CFA Institute®
CFA Program

ETHICAL AND PROFESSIONAL STANDARDS, QUANTITATIVE METHODS, AND ECONOMICS

CFA® Program Curriculum
2020 • LEVEL II • VOLUME 1

ISBN 978-1-946442-82-6 (paper)
ISBN 978-1-950157-06-8 (ebk)

10 9 8 7 6 5 4 3 2 1

CONTENTS

indicates an optional segment

O indicates an optional segment

Quantitative Methods

Economics

How to Use the CFA Program Curriculum

Congratulations on reaching Level II of the Chartered Financial Analyst® (CFA®) Program. This exciting and rewarding program of study reflects your desire to become a serious investment professional. You have embarked on a program noted for its high ethical standards and the breadth of knowledge, skills, and abilities (competencies) it develops. Your commitment to the CFA Program should be educationally and professionally rewarding.

The credential you seek is respected around the world as a mark of accomplishment and dedication. Each level of the program represents a distinct achievement in professional development. Successful completion of the program is rewarded with membership in a prestigious global community of investment professionals. CFA charterholders are dedicated to life-long learning and maintaining currency with the ever-changing dynamics of a challenging profession. The CFA Program represents the first step toward a career-long commitment to professional education.

The CFA examination measures your mastery of the core knowledge, skills, and abilities required to succeed as an investment professional. These core competencies are the basis for the Candidate Body of Knowledge (CBOK™). The CBOK consists of four components:

- A broad outline that lists the major topic areas covered in the CFA Program (https://www.cfainstitute.org/programs/cfa/curriculum/cbok);
- Topic area weights that indicate the relative exam weightings of the top-level topic areas (https://www.cfainstitute.org/programs/cfa/curriculum/overview);
- Learning outcome statements (LOS) that advise candidates about the specific knowledge, skills, and abilities they should acquire from readings covering a topic area (LOS are provided in candidate study sessions and at the beginning of each reading); and
- The CFA Program curriculum that candidates receive upon examination registration.

Therefore, the key to your success on the CFA examinations is studying and understanding the CBOK. The following sections provide background on the CBOK, the organization of the curriculum, features of the curriculum, and tips for designing an effective personal study program.

BACKGROUND ON THE CBOK

The CFA Program is grounded in the practice of the investment profession. Beginning with the Global Body of Investment Knowledge (GBIK), CFA Institute performs a continuous practice analysis with investment professionals around the world to determine the competencies that are relevant to the profession. Regional expert panels and targeted surveys are conducted annually to verify and reinforce the continuous feedback about the GBIK. The practice analysis process ultimately defines the CBOK. The

CBOK reflects the competencies that are generally accepted and applied by investment professionals. These competencies are used in practice in a generalist context and are expected to be demonstrated by a recently qualified CFA charterholder.

The CFA Institute staff, in conjunction with the Education Advisory Committee and Curriculum Level Advisors that consist of practicing CFA charterholders, designs the CFA Program curriculum in order to deliver the CBOK to candidates. The examinations, also written by CFA charterholders, are designed to allow you to demonstrate your mastery of the CBOK as set forth in the CFA Program curriculum. As you structure your personal study program, you should emphasize mastery of the CBOK and the practical application of that knowledge. For more information on the practice analysis, CBOK, and development of the CFA Program curriculum, please visit www.cfainstitute.org.

ORGANIZATION OF THE CURRICULUM

The Level II CFA Program curriculum is organized into 10 topic areas. Each topic area begins with a brief statement of the material and the depth of knowledge expected. It is then divided into one or more study sessions. These study sessions—17 sessions in the Level II curriculum—should form the basic structure of your reading and preparation. Each study session includes a statement of its structure and objective and is further divided into assigned readings. An outline illustrating the organization of these 17 study sessions can be found at the front of each volume of the curriculum.

The readings are commissioned by CFA Institute and written by content experts, including investment professionals and university professors. Each reading includes LOS and the core material to be studied, often a combination of text, exhibits, and in-text examples and questions. A reading typically ends with practice problems followed by solutions to these problems to help you understand and master the material. The LOS indicate what you should be able to accomplish after studying the material. The LOS, the core material, and the practice problems are dependent on each other, with the core material and the practice problems providing context for understanding the scope of the LOS and enabling you to apply a principle or concept in a variety of scenarios.

The entire readings, including the practice problems at the end of the readings, are the basis for all examination questions and are selected or developed specifically to teach the knowledge, skills, and abilities reflected in the CBOK.

You should use the LOS to guide and focus your study because each examination question is based on one or more LOS and the core material and practice problems associated with the LOS. As a candidate, you are responsible for the entirety of the required material in a study session.

We encourage you to review the information about the LOS on our website (www.cfainstitute.org/programs/cfa/curriculum/study-sessions), including the descriptions of LOS "command words" on the candidate resources page at www.cfainstitute.org.

FEATURES OF THE CURRICULUM

OPTIONAL SEGMENT

Required vs. Optional Segments You should read all of an assigned reading. In some cases, though, we have reprinted an entire publication and marked certain parts of the reading as "optional." The CFA examination is based only on the required segments, and the optional segments are included only when it is determined that they might help you to better understand the required segments (by seeing the required material in its full context). When an optional segment begins, you will see an icon and a dashed vertical bar in the outside margin that will continue until the optional segment ends, accompanied by another icon. *Unless the material is specifically marked as optional, you should assume it is required.* You should rely on the required segments and the reading-specific LOS in preparing for the examination.

END OPTIONAL SEGMENT

Practice Problems/Solutions *All practice problems at the end of the readings as well as their solutions are part of the curriculum and are required material for the examination.* In addition to the in-text examples and questions, these practice problems should help demonstrate practical applications and reinforce your understanding of the concepts presented. Some of these practice problems are adapted from past CFA examinations and/or may serve as a basis for examination questions.

Glossary For your convenience, each volume includes a comprehensive glossary. Throughout the curriculum, a **bolded** word in a reading denotes a term defined in the glossary.

Note that the digital curriculum that is included in your examination registration fee is searchable for key words, including glossary terms.

LOS Self-Check We have inserted checkboxes next to each LOS that you can use to track your progress in mastering the concepts in each reading.

Source Material The CFA Institute curriculum cites textbooks, journal articles, and other publications that provide additional context and information about topics covered in the readings. As a candidate, you are not responsible for familiarity with the original source materials cited in the curriculum.

Note that some readings may contain a web address or URL. The referenced sites were live at the time the reading was written or updated but may have been deactivated since then.

Some readings in the curriculum cite articles published in the *Financial Analysts Journal®*, which is the flagship publication of CFA Institute. Since its launch in 1945, the *Financial Analysts Journal* has established itself as the leading practitioner-oriented journal in the investment management community. Over the years, it has advanced the knowledge and understanding of the practice of investment management through the publication of peer-reviewed practitioner-relevant research from leading academics and practitioners. It has also featured thought-provoking opinion pieces that advance the common level of discourse within the investment management profession. Some of the most influential research in the area of investment management has appeared in the pages of the *Financial Analysts Journal*, and several Nobel laureates have contributed articles.

Candidates are not responsible for familiarity with *Financial Analysts Journal* articles that are cited in the curriculum. But, as your time and studies allow, we strongly encourage you to begin supplementing your understanding of key investment management issues by reading this practice-oriented publication. Candidates have full online access to the *Financial Analysts Journal* and associated resources. All you need is to log in on www.cfapubs.org using your candidate credentials.

Errata The curriculum development process is rigorous and includes multiple rounds of reviews by content experts. Despite our efforts to produce a curriculum that is free of errors, there are times when we must make corrections. Curriculum errata are periodically updated and posted on the candidate resources page at www.cfainstitute.org.

DESIGNING YOUR PERSONAL STUDY PROGRAM

Create a Schedule An orderly, systematic approach to examination preparation is critical. You should dedicate a consistent block of time every week to reading and studying. Complete all assigned readings and the associated problems and solutions in each study session. Review the LOS both before and after you study each reading to ensure that you have mastered the applicable content and can demonstrate the knowledge, skills, and abilities described by the LOS and the assigned reading. Use the LOS self-check to track your progress and highlight areas of weakness for later review.

Successful candidates report an average of more than 300 hours preparing for each examination. Your preparation time will vary based on your prior education and experience, and you will probably spend more time on some study sessions than on others. As the Level II curriculum includes 17 study sessions, a good plan is to devote 15–20 hours per week for 17 weeks to studying the material and use the final four to six weeks before the examination to review what you have learned and practice with practice questions and mock examinations. This recommendation, however, may underestimate the hours needed for appropriate examination preparation depending on your individual circumstances, relevant experience, and academic background. You will undoubtedly adjust your study time to conform to your own strengths and weaknesses and to your educational and professional background.

You should allow ample time for both in-depth study of all topic areas and additional concentration on those topic areas for which you feel the least prepared.

As part of the supplemental study tools that are included in your examination registration fee, you have access to a study planner to help you plan your study time. The study planner calculates your study progress and pace based on the time remaining until examination. For more information on the study planner and other supplemental study tools, please visit www.cfainstitute.org.

As you prepare for your examination, we will e-mail you important examination updates, testing policies, and study tips. Be sure to read these carefully.

CFA Institute Practice Questions Your examination registration fee includes digital access to hundreds of practice questions that are additional to the practice problems at the end of the readings. These practice questions are intended to help you assess your mastery of individual topic areas as you progress through your studies. After each practice question, you will be able to receive immediate feedback noting the correct responses and indicating the relevant assigned reading so you can identify areas of weakness for further study. For more information on the practice questions, please visit www.cfainstitute.org.

CFA Institute Mock Examinations Your examination registration fee also includes digital access to three-hour mock examinations that simulate the morning and afternoon sessions of the actual CFA examination. These mock examinations are intended to be taken after you complete your study of the full curriculum and take practice questions so you can test your understanding of the curriculum and your readiness for the examination. You will receive feedback at the end of the mock examination, noting the correct responses and indicating the relevant assigned readings so you can assess areas of weakness for further study during your review period. We recommend that you take mock examinations during the final stages of your preparation for the actual CFA examination. For more information on the mock examinations, please visit www.cfainstitute.org.

Preparatory Providers After you enroll in the CFA Program, you may receive numerous solicitations for preparatory courses and review materials. When considering a preparatory course, make sure the provider belongs to the CFA Institute Approved Prep Provider Program. Approved Prep Providers have committed to follow CFA Institute guidelines and high standards in their offerings and communications with candidates. For more information on the Approved Prep Providers, please visit www.cfainstitute.org/programs/cfa/exam/prep-providers.

Remember, however, that there are no shortcuts to success on the CFA examinations; reading and studying the CFA curriculum *is* the key to success on the examination. The CFA examinations reference only the CFA Institute assigned curriculum—no preparatory course or review course materials are consulted or referenced.

SUMMARY

Every question on the CFA examination is based on the content contained in the required readings and on one or more LOS. Frequently, an examination question is based on a specific example highlighted within a reading or on a specific practice problem and its solution. To make effective use of the CFA Program curriculum, please remember these key points:

1. All pages of the curriculum are required reading for the examination except for occasional sections marked as optional. You may read optional pages as background, but you will not be tested on them.
2. All questions, problems, and their solutions—found at the end of readings—are part of the curriculum and are required study material for the examination.
3. You should make appropriate use of the practice questions and mock examinations as well as other supplemental study tools and candidate resources available at www.cfainstitute.org.
4. Create a schedule and commit sufficient study time to cover the 17 study sessions using the study planner. You should also plan to review the materials and take topic tests and mock examinations.
5. Some of the concepts in the study sessions may be superseded by updated rulings and/or pronouncements issued after a reading was published. Candidates are expected to be familiar with the overall analytical framework contained in the assigned readings. Candidates are not responsible for changes that occur after the material was written.

FEEDBACK

At CFA Institute, we are committed to delivering a comprehensive and rigorous curriculum for the development of competent, ethically grounded investment professionals. We rely on candidate and investment professional comments and feedback as we work to improve the curriculum, supplemental study tools, and candidate resources.

Please send any comments or feedback to info@cfainstitute.org. You can be assured that we will review your suggestions carefully. Ongoing improvements in the curriculum will help you prepare for success on the upcoming examinations and for a lifetime of learning as a serious investment professional.

Ethical and Professional Standards

STUDY SESSIONS

Study Session 1	Ethical and Professional Standards

TOPIC LEVEL LEARNING OUTCOME

The candidate should be able to demonstrate a thorough knowledge of the CFA Institute Code of Ethics and Standards of Professional Conduct (Code and Standards), identify violations of the Code and Standards, and recommend appropriate corrective measures.

The topic of ethics is one of fundamental importance to the investment profession. Acting responsibly with high levels of integrity builds trust, upon which the investment profession is built. Behaving ethically, and in the best interest of clients, is critical to the long-term success of the investment profession and those choosing a career within it.

The Code and Standards form the ethical foundation for the CFA Institute self-regulatory program. The *Standards of Practice Handbook* provides practical guidance in the interpretation and implementation of the Code and Standards. Together these seek to hold CFA Institute members[1] and CFA Program candidates to the highest levels of professional behavior.

1 Eligibility and requirements for becoming a member of CFA Institute vary by jurisdiction. Please consult www.cfainstitute.org for further details.

ETHICAL AND PROFESSIONAL STANDARDS

STUDY SESSION

1

Ethical and Professional Standards

The readings in this study session address the CFA Institute Code of Ethics and Standards of Professional Conduct (Code and Standards. The Code and Standards provide guidance to help identify and resolve ethical conflicts present in everyday activities in the investment profession. “Guidance” in the *Standards of Practice Handbook* addresses the practical application of the Code and Standards by reviewing the purpose and scope of each Standard, presenting recommended procedures for compliance, and providing examples of the Standard in practice.

READING ASSIGNMENTS

Reading 1	Code of Ethics and Standards of Professional Conduct *Standards of Practice Handbook*, Eleventh Edition
Reading 2	Guidance for Standards I–VII *Standards of Practice Handbook*, Eleventh Edition
Reading 3	Application of the Code and Standards: Level II

READING

1

Code of Ethics and Standards of Professional Conduct

LEARNING OUTCOMES

Mastery	The candidate should be able to:
☐	**a.** describe the six components of the Code of Ethics and the seven Standards of Professional Conduct;
☐	**b.** explain the ethical responsibilities required of CFA Institute members and candidates in the CFA Program by the Code and Standards.

PREFACE

The *Standards of Practice Handbook* (*Handbook*) provides guidance to the people who grapple with real ethical dilemmas in the investment profession on a daily basis; the *Handbook* addresses the professional intersection where theory meets practice and where the concept of ethical behavior crosses from the abstract to the concrete. The *Handbook* is intended for a diverse and global audience: CFA Institute members navigating ambiguous ethical situations; supervisors and direct/indirect reports determining the nature of their responsibilities to each other, to existing and potential clients, and to the broader financial markets; and candidates preparing for the Chartered Financial Analyst (CFA) examinations.

Recent events in the global financial markets have tested the ethical mettle of financial market participants, including CFA Institute members. The standards taught in the CFA Program and by which CFA Institute members and candidates must abide represent timeless ethical principles and professional conduct for all market conditions. Through adherence to these standards, which continue to serve as the model for ethical behavior in the investment professional globally, each market participant does his or her part to improve the integrity and efficient operations of the financial markets.

The *Handbook* provides guidance in understanding the interconnectedness of the aspirational and practical principles and provisions of the Code of Ethics and Standards of Professional Conduct (Code and Standards). The Code contains high-level aspirational ethical principles that drive members and candidates to create a positive and reputable investment profession. The Standards contain practical ethical principles of conduct that members and candidates must follow to achieve the broader industry expectations. However, applying the principles individually may not capture the complexity of ethical requirements related to the investment industry. The Code and Standards should be viewed and interpreted as an interwoven tapestry of ethical requirements. Through members' and candidates' adherence to these principles as a whole, the integrity of and trust in the capital markets are improved.

Evolution of the CFA Institute Code of Ethics and Standards of Professional Conduct

Generally, changes to the Code and Standards over the years have been minor. CFA Institute has revised the language of the Code and Standards and occasionally added a new standard to address a prominent issue of the day. For instance, in 1992, CFA Institute added the standard addressing performance presentation to the existing list of standards.

Major changes came in 2005 with the ninth edition of the *Handbook*. CFA Institute adopted new standards, revised some existing standards, and reorganized the standards. The revisions were intended to clarify the requirements of the Code and Standards and effectively convey to its global membership what constitutes "best practice" in a number of areas relating to the investment profession.

The Code and Standards must be regularly reviewed and updated if they are to remain effective and continue to represent the highest ethical standards in the global investment industry. CFA Institute strongly believes that revisions of the Code and Standards are not undertaken for cosmetic purposes but to add value by addressing legitimate concerns and improving comprehension.

Changes to the Code and Standards have far-reaching implications for the CFA Institute membership, the CFA Program, and the investment industry as a whole. CFA Institute members and candidates are *required* to adhere to the Code and Standards. In addition, the Code and Standards are increasingly being adopted, in whole or in part, by firms and regulatory authorities. Their relevance goes well beyond CFA Institute members and candidates.

Standards of Practice Handbook

The periodic revisions of the Code and Standards have come in conjunction with updates of the *Standards of Practice Handbook*. The *Handbook* is the fundamental element of the ethics education effort of CFA Institute and the primary resource for guidance in interpreting and implementing the Code and Standards. The *Handbook* seeks to educate members and candidates on how to apply the Code and Standards to their professional lives and thereby benefit their clients, employers, and the investing public in general. The *Handbook* explains the purpose of the Code and Standards and how they apply in a variety of situations. The sections discuss and amplify each standard and suggest procedures to prevent violations.

Examples in the "Application of the Standard" sections are meant to illustrate how the standard applies to hypothetical but factual situations. The names contained in the examples are fictional and are not meant to refer to any actual person or entity. Unless otherwise stated (e.g., one or more people specifically identified), individuals in each example are CFA Institute members and holders of the CFA designation. Because factual circumstances vary so widely and often involve gray areas, the explanatory material and examples are not intended to be all inclusive. Many examples set forth in the application sections involve standards that have legal counterparts; ***members are strongly urged to discuss with their supervisors and legal and compliance departments the content of the Code and Standards and the members' general obligations under the Code and Standards***.

CFA Institute recognizes that the presence of any set of ethical standards may create a false sense of security unless the documents are fully understood, enforced, and made a meaningful part of everyday professional activities. The *Handbook* is intended to provide a useful frame of reference that suggests ethical professional behavior in the investment decision-making process. This book cannot cover every contingency or circumstance, however, and it does not attempt to do so. The development and interpretation of the Code and Standards are evolving processes; the Code and Standards will be subject to continuing refinement.

Summary of Changes in the Eleventh Edition

The comprehensive review of the Code and Standards in 2005 resulted in principle requirements that remain applicable today. The review carried out for the eleventh edition focused on market practices that have evolved since the tenth edition. Along with updates to the guidance and examples within the *Handbook*, the eleventh edition includes an update to the Code of Ethics that embraces the members' role of maintaining the social contract between the industry and investors. Additionally, there are three changes to the Standards of Professional Conduct, which recognize the importance of proper supervision, clear communications with clients, and the expanding educational programs of CFA Institute.

Inclusion of Updated CFA Institute Mission

The CFA Institute Board of Governors approved an updated mission for the organization that is included in the Preamble to the Code and Standards. The new mission conveys the organization's conviction in the investment industry's role in the betterment of society at large.

> ***Mission:***
>
> To lead the investment profession globally by promoting the highest standards of ethics, education, and professional excellence for the ultimate benefit of society.

Updated Code of Ethics Principle

One of the bullets in the Code of Ethics was updated to reflect the role that the capital markets have in the greater society. As members work to promote and maintain the integrity of the markets, their actions should also help maintain the social contract with investors.

> ***Old:***
>
> Promote the integrity of and uphold the rules governing capital markets.

> ***New:***
>
> Promote the integrity and viability of the global capital markets for the ultimate benefit of society.

New Standard Regarding Responsibilities of Supervisors [IV(C)]

The standard for members and candidates with supervision or authority over others within their firms was updated to bring about improvements in preventing illegal and unethical actions from occurring. The prior version of Standard IV(C) focused on the detection and prevention of violations. The updated version stresses broader compliance expectations, which include the detection and prevention aspects of the original version.

> ***Old:***
>
> Members and Candidates must make reasonable efforts to detect and prevent violations of applicable laws, rules, regulations, and the Code and Standards by anyone subject to their supervision or authority.
>
> ***New:***
>
> Members and Candidates must make reasonable efforts to ensure that anyone subject to their supervision or authority complies with applicable laws, rules, regulations, and the Code and Standards.

Additional Requirement under the Standard for Communication with Clients and Prospective Clients [V(B)]

Given the constant development of new and exotic financial instruments and strategies, the standard regarding communicating with clients now includes an implicit requirement to discuss the risks and limitations of recommendations being made to clients. The new principle and related guidance take into account the fact that levels of disclosure will differ between products and services. Members and candidates, along with their firms, must determine the specific disclosures their clients should receive while ensuring appropriate transparency of the individual firms' investment processes.

> ***Addition:***
>
> Disclose to clients and prospective clients significant limitations and risks associated with the investment process.

Modification to Standard VII(A)

Since this standard was developed, CFA Institute has launched additional educational programs. The updated standard not only maintains the integrity of the CFA Program but also expands the same ethical considerations when members or candidates participate in such programs as the CIPM Program and the CFA Institute Investment Foundations certificate program. Whether participating as a member assisting with the curriculum or an examination or as a sitting candidate within a program, we expect them to engage in these programs as they would participate in the CFA Program.

> ***Old:***
>
> Conduct as Members and Candidates in the CFA Program
>
> Members and Candidates must not engage in any conduct that compromises the reputation or integrity of CFA Institute or the CFA designation or the integrity, validity, or security of the CFA examinations.
>
> ***New:***
>
> Conduct as Participants in CFA Institute Programs
>
> Members and Candidates must not engage in any conduct that compromises the reputation or integrity of CFA Institute or the CFA designation or the integrity, validity, or security of CFA Institute programs.

General Guidance and Example Revision

The guidance and examples were updated to reflect practices and scenarios applicable to today's investment industry. Two concepts that appear frequently in the updates in this edition relate to the increased use of social media for business communications and the use of and reliance on the output of quantitative models. The use of social media platforms has increased significantly since the publication of the tenth edition. And although financial modeling is not new to the industry, this update reflects upon actions that are viewed as possible contributing factors to the financial crises of the past decade.

CFA Institute Professional Conduct Program

All CFA Institute members and candidates enrolled in the CFA Program are required to comply with the Code and Standards. The CFA Institute Board of Governors maintains oversight and responsibility for the Professional Conduct Program (PCP), which, in conjunction with the Disciplinary Review Committee (DRC), is responsible for enforcement of the Code and Standards. The DRC is a volunteer committee of CFA charterholders who serve on panels to review conduct and partner with Professional Conduct staff to establish and review professional conduct policies. The CFA Institute Bylaws and Rules of Procedure for Professional Conduct (Rules of Procedure) form the basic structure for enforcing the Code and Standards. The Professional Conduct division is also responsible for enforcing testing policies of other CFA Institute education programs as well as the professional conduct of Certificate in Investment Performance Measurement (CIPM) certificants.

Professional Conduct inquiries come from a number of sources. First, members and candidates must self-disclose on the annual Professional Conduct Statement all matters that question their professional conduct, such as involvement in civil litigation or a criminal investigation or being the subject of a written complaint. Second, written complaints received by Professional Conduct staff can bring about an investigation. Third, CFA Institute staff may become aware of questionable conduct by a member or candidate through the media, regulatory notices, or another public source. Fourth, candidate conduct is monitored by proctors who complete reports on candidates suspected to have violated testing rules on exam day. Lastly, CFA Institute may also conduct analyses of scores and exam materials after the exam, as well as monitor online and social media to detect disclosure of confidential exam information.

When an inquiry is initiated, the Professional Conduct staff conducts an investigation that may include requesting a written explanation from the member or candidate; interviewing the member or candidate, complaining parties, and third parties; and collecting documents and records relevant to the investigation. Upon reviewing the material obtained during the investigation, the Professional Conduct staff may conclude the inquiry with no disciplinary sanction, issue a cautionary letter, or continue proceedings to discipline the member or candidate. If the Professional Conduct staff believes a violation of the Code and Standards or testing policies has occurred, the member or candidate has the opportunity to reject or accept any charges and the proposed sanctions.

If the member or candidate does not accept the charges and proposed sanction, the matter is referred to a panel composed of DRC members. Panels review materials and presentations from Professional Conduct staff and from the member or candidate. The panel's task is to determine whether a violation of the Code and Standards or testing policies occurred and, if so, what sanction should be imposed.

Sanctions imposed by CFA Institute may have significant consequences; they include public censure, suspension of membership and use of the CFA designation, and revocation of the CFA charter. Candidates enrolled in the CFA Program who have violated the Code and Standards or testing policies may be suspended or prohibited from further participation in the CFA Program.

Adoption of the Code and Standards

The Code and Standards apply to individual members of CFA Institute and candidates in the CFA Program. CFA Institute does encourage firms to adopt the Code and Standards, however, as part of their code of ethics. Those who claim compliance should fully understand the requirements of each of the principles of the Code and Standards.

Once a party—nonmember or firm—ensures its code of ethics meets the principles of the Code and Standards, that party should make the following statement whenever claiming compliance:

> "[Insert name of party] claims compliance with the CFA Institute Code of Ethics and Standards of Professional Conduct. This claim has not been verified by CFA Institute."

CFA Institute welcomes public acknowledgement, when appropriate, that firms are complying with the CFA Institute Code of Ethics and Standards of Professional Conduct and encourages firms to notify us of the adoption plans. For firms that would like to distribute the Code and Standards to clients and potential clients, attractive one-page copies of the Code and Standards, including translations, are available on the CFA Institute website (www.cfainstitute.org).

CFA Institute has also published the Asset Manager Code of Professional Conduct, which is designed, in part, to help asset managers comply with the regulations mandating codes of ethics for investment advisers. Whereas the Code and Standards are aimed at individual investment professionals who are members of CFA Institute or candidates in the CFA Program, the Asset Manager Code was drafted specifically for firms. The Asset Manager Code provides specific, practical guidelines for asset managers in six areas: loyalty to clients, the investment process, trading, compliance, performance evaluation, and disclosure. The Asset Manager Code and the appropriate steps to acknowledge adoption or compliance can be found on the CFA Institute website (www.cfainstitute.org).

Acknowledgments

CFA Institute is a not-for-profit organization that is heavily dependent on the expertise and intellectual contributions of member volunteers. Members devote their time because they share a mutual interest in the organization's mission to promote and achieve ethical practice in the investment profession. CFA Institute owes much to the volunteers' abundant generosity and energy in extending ethical integrity.

The CFA Institute Standards of Practice Council (SPC), a group consisting of CFA charterholder volunteers from many different countries, is charged with maintaining and interpreting the Code and Standards and ensuring that they are effective. The SPC draws its membership from a broad spectrum of organizations in the securities field, including brokers, investment advisers, banks, and insurance companies. In most instances, the SPC members have important supervisory responsibilities within their firms.

The SPC continually evaluates the Code and Standards, as well as the guidance in the *Handbook*, to ensure that they are

- representative of high standards of professional conduct,
- relevant to the changing nature of the investment profession,
- globally applicable,
- sufficiently comprehensive, practical, and specific,
- enforceable, and
- testable for the CFA Program.

The SPC has spent countless hours reviewing and discussing revisions to the Code and Standards and updates to the guidance that make up the eleventh edition of the *Handbook*. Following is a list of the current and former members of the SPC who generously donated their time and energy to this effort.

James E. Hollis III, CFA, Chair
Rik Albrecht, CFA
Terence E. Burns, CFA
Laura Dagan, CFA
Samuel B. Jones, Jr., CFA
Ulrike Kaiser-Boeing, CFA
Jinliang (Jack) Li, CFA
Christopher C. Loop, CFA,
James M. Meeth, CFA
Guy G. Rutherfurd, Jr., CFA
Edouard Senechal, CFA
Wenliang (Richard) Wang, CFA
Peng Lian Wee, CFA

ETHICS AND THE INVESTMENT INDUSTRY

Society ultimately benefits from efficient markets where capital can freely flow to the most productive or innovative destination. Well-functioning capital markets efficiently match those needing capital with those seeking to invest their assets in revenue-generating ventures. In order for capital markets to be efficient, investors must be able to trust that the markets are fair and transparent and offer them the opportunity to be rewarded for the risk they choose to take. Laws, regulations, and enforcement play a vital role but are insufficient alone to guarantee fair and transparent markets. The markets depend on an ethical foundation to guide participants' judgment and behavior. CFA Institute maintains and promotes the Code of Ethics and Standards of Professional Conduct in order to create a culture of ethics for the ultimate benefit of society.

Why Ethics Matters

Ethics can be defined as a set of moral principles or rules of conduct that provide guidance for our behavior when it affects others. Widely acknowledged fundamental ethical principles include honesty, fairness, diligence, and care and respect for others. Ethical conduct follows those principles and balances self-interest with both the direct and the indirect consequences of that behavior for other people.

Not only does unethical behavior by individuals have serious personal consequences—ranging from job loss and reputational damage to fines and even jail—but unethical conduct from market participants, investment professionals, and those who service investors can damage investor trust and thereby impair the sustainability of the global capital markets as a whole. Unfortunately, there seems to be an unending parade of stories bringing to light accounting frauds and manipulations, Ponzi schemes, insider-trading scandals, and other misdeeds. Not surprisingly, this has led to erosion in public confidence in investment professionals. Empirical evidence from numerous surveys documents the low standing in the eyes of the investing public of banks and financial services firms—the very institutions that are entrusted with the economic well-being and retirement security of society.

Governments and regulators have historically tried to combat misconduct in the industry through regulatory reform, with various levels of success. Global capital markets are highly regulated to protect investors and other market participants. However, compliance with regulation alone is insufficient to fully earn investor trust. Individuals and firms must develop a "culture of integrity" that permeates all levels of operations and promotes the ethical principles of stewardship of investor assets and working in the best interests of clients, above and beyond strict compliance with the law. A strong ethical culture that helps honest, ethical people engage in ethical behavior will foster the trust of investors, lead to robust global capital markets, and ultimately benefit society. That is why ethics matters.

Ethics, Society, and the Capital Markets

CFA Institute recently added the concept "for the ultimate benefit of society" to its mission. The premise is that we want to live in a socially, politically, and financially stable society that fosters individual well-being and welfare of the public. A key ingredient for this goal is global capital markets that facilitate the efficient allocation of resources so that the available capital finds its way to places where it most benefits that society. These investments are then used to produce goods and services, to fund innovation and jobs, and to promote improvements in standards of living. Indeed, such a function serves the interests of the society. Efficient capital markets, in turn, provide a host of benefits to those providing the investment capital. Investors are provided the opportunity to transfer and transform risk because the capital markets serve as an information exchange, create investment products, provide liquidity, and limit transaction costs.

However, a well-functioning and efficient capital market system is dependent on trust of the participants. If investors believe that capital market participants—investment professionals and firms—cannot be trusted with their financial assets or that the capital markets are unfair such that only insiders can be successful, they will be unlikely to invest or, at the very least, will require a higher risk premium. Decreased investment capital can reduce innovation and job creation and hurt the economy and society as a whole. Reduced trust in capital markets can also result in a less vibrant, if not smaller, investment industry.

Ethics for a global investment industry should be universal and ultimately support trust and integrity above acceptable local or regional customs and culture. Universal ethics for a global industry strongly supports the efficiency, values, and mission of the industry as a whole. Different countries may be at different stages of development in establishing standards of practice, but the end goal must be to achieve rules, regulations, and standards that support and promote fundamental ethical principles on a global basis.

Capital Market Sustainability and the Actions of One

Individuals and firms also have to look at the indirect impacts of their actions on the broader investment community. The increasingly interconnected nature of global finance brings to the fore an added consideration of market sustainability that was, perhaps, less appreciated in years past. In addition to committing to the highest levels of ethical behavior, today's investment professionals and their employers should consider the long-term health of the market as a whole.

As recent events have demonstrated, apparently isolated and unrelated decisions, however innocuous when considered on an individual basis, in aggregate can precipitate a market crisis. In an interconnected global economy and marketplace, each participant should strive to be aware of how his or her actions or the products he or she distributes may have an impact on capital market participants in other regions or countries.

Investment professionals should consider how their investment decision-making processes affect the global financial markets in the broader context of how they apply their ethical and professional obligations. Those in positions of authority have a special responsibility to consider the broader context of market sustainability in their development and approval of corporate policies, particularly those involving risk management and product development. In addition, corporate compensation strategies should not encourage otherwise ethically sound individuals to engage in unethical or questionable conduct for financial gain. Ethics, sustainability, and properly functioning capital markets are components of the same concept of protecting the best interests of all. To always place the interests of clients ahead of both investment professionals' own interests and those of their employer remains a key ethos.

The Relationship between Ethics and Regulations

Some equate ethical behavior with legal behavior: If you are following the law, you must be acting appropriately. Ethical principles, like laws and regulations, prescribe appropriate constraints on our natural tendency to pursue self-interest that could harm the interests of others. Laws and regulations often attempt to guide people toward ethical behavior, but they do not cover all unethical behavior. Ethical behavior is often distinguished from legal conduct by describing legal behavior as what is required and ethical behavior as conduct that is morally correct. Ethical principles go beyond that which is legally sufficient and encompass what is the right thing to do.

Given many regulators' lack of sufficient resources to enforce well-conceived rules and regulations, relying on a regulatory framework to lead the charge in establishing ethical behavior has its challenges. Therefore, reliance on compliance with laws and regulation alone is insufficient to ensure ethical behavior of investment professionals or to create a truly ethical culture in the industry.

The recent past has shown us that some individuals will succeed at circumventing the regulatory rules for their personal gain. Only the application of strong ethical principles, at both the individual level and the firm level, will limit abuses. Knowing the rules or regulations to apply in a particular situation, although important, may not be sufficient to ensure ethical conduct. Individuals must be able both to recognize areas that are prone to ethical pitfalls and to identify and process those circumstances and influences that can impair ethical judgment.

Applying an Ethical Framework

Laws, regulations, professional standards, and codes of ethics can guide ethical behavior, but individual judgment is a critical ingredient in making principled choices and engaging in appropriate conduct. When faced with an ethical dilemma, individuals must have a well-developed set of principles; otherwise, their thought processes can lead to, at best, equivocation and indecision and, at worst, fraudulent conduct and destruction of the public trust. Establishing an ethical framework for an internal thought process prior to deciding to act is a crucial step in engaging in ethical conduct.

Most investment professionals are used to making decisions from a business (profit/loss) outlook. But given the importance of ethical behavior in carrying out professional responsibilities, it is critical to also analyze decisions and potential conduct from an ethical perspective. Utilizing a framework for ethical decision making will help investment professionals effectively examine their conduct in the context of conflicting interests common to their professional obligations (e.g., researching and gathering information, developing investment recommendations, and managing money for others). Such a framework will allow investment professionals to analyze their conduct in a way that meets high standards of ethical behavior.

An ethical decision-making framework can come in many forms but should provide investment professionals with a tool for following the principles of the firm's code of ethics. Through analyzing the particular circumstances of each decision, investment professionals are able to determine the best course of action to fulfill their responsibilities in an ethical manner.

Commitment to Ethics by Firms

A firm's code of ethics risks becoming a largely ignored, dusty compilation if it is not truly integrated into the fabric of the business. The ability to relate an ethical decision-making framework to a firm's code of ethics allows investment professionals to bring the aspirations and principles of the code of ethics to life—transforming it from a compliance exercise to something that is at the heart of a firm's culture.

An investment professional's natural desire to "do the right thing" must be reinforced by building a culture of integrity in the workplace. Development, maintenance, and demonstration of a strong culture of integrity within the firm by senior management may be the single most important factor in promoting ethical behavior among the firm's employees. Adopting a code that clearly lays out the ethical principles that guide the thought processes and conduct the firm expects from its employees is a critical first step. But a code of ethics, while necessary, is insufficient.

Simply nurturing an inclination to do right is no match for the multitude of daily decisions that investment managers make. We need to exercise ethical decision-making skills to develop the muscle memory necessary for fundamentally ethical people to make good decisions despite the reality of agent conflicts. Just as coaching and practice transform our natural ability to run across a field into the technique and endurance required to run a race, teaching, reinforcing, and practicing ethical decision-making skills prepare us to confront the hard issues effectively. It is good for business, individuals, firms, the industry, and the markets, as well as society as a whole, to engage in the investment management profession in a highly ethical manner.

Ethical Commitment of CFA Institute

An important goal of CFA Institute is to ensure that the organization and its members and candidates develop, promote, and follow the highest ethical standards in the investment industry. The CFA Institute Code of Ethics (Code) and Standards of Professional Conduct (Standards) are the foundation supporting the organization's quest to uphold the industry's highest standards of individual and corporate practice and to help serve the greater good. The Code is a set of principles that define the overarching conduct CFA Institute expects from its members and CFA Program candidates. The Code works in tandem with the Standards, which outline professional conduct that constitutes fair and ethical business practices.

For more than 50 years, CFA Institute members and candidates have been required to abide by the organization's Code and Standards. Periodically, CFA Institute has revised and updated its Code and Standards to ensure that they remain relevant to the changing nature of the investment profession and representative of the highest standard of professional conduct. Within this *Handbook*, CFA Institute addresses ethical principles for the profession, including individual professionalism; responsibilities to capital markets, clients, and employers; ethics involved in investment analysis, recommendations, and actions; and possible conflicts of interest. Although the investment world has become a far more complex place since the first publication of the *Standard of Practice Handbook*, distinguishing right from wrong remains the paramount principle of the Code and Standards.

New challenges will continually arise for members and candidates in applying the Code and Standards because many decisions are not unambiguously right or wrong. The dilemma exists because the choice between right and wrong is not always clear.

Even well-intentioned investment professionals can find themselves in circumstances that may tempt them to cut corners. Situational influences can overpower the best of intentions.

CFA Institute has made a significant commitment to providing members and candidates with the resources to extend and deepen their understanding of how to appropriately apply the principles of the Code and Standards. The product offerings from CFA Institute offer a wealth of material. Through publications, conferences, webcasts, and podcasts, the ethical challenges of investment professionals are brought to light. Archived issues of these items are available on the CFA Institute website (www.cfainstitute.org).

By reviewing these resources and discussing with their peers, market participants can further enhance their abilities to apply an effective ethical decision-making framework. In time, this should help restore some of the trust recently lost by investors.

Markets function to an important extent on trust. Recent events have shown the fragility of this foundation and the devastating consequences that can ensue when it is fundamentally questioned. Investment professionals should remain mindful of the long-term health of financial markets and incorporate this concern for the market's sustainability in their investment decision making. CFA Institute and the Standards of Practice Council hope this edition of the *Handbook* will assist and guide investment professionals in meeting the ethical demands of the highly interconnected global capital markets for the ultimate benefit of society.

CFA INSTITUTE CODE OF ETHICS AND STANDARDS OF PROFESSIONAL CONDUCT

Preamble

The CFA Institute Code of Ethics and Standards of Professional Conduct are fundamental to the values of CFA Institute and essential to achieving its mission to lead the investment profession globally by promoting the highest standards of ethics, education, and professional excellence for the ultimate benefit of society. High ethical standards are critical to maintaining the public's trust in financial markets and in the investment profession. Since their creation in the 1960s, the Code and Standards have promoted the integrity of CFA Institute members and served as a model for measuring the ethics of investment professionals globally, regardless of job function, cultural differences, or local laws and regulations. All CFA Institute members (including holders of the Chartered Financial Analyst [CFA] designation) and CFA candidates have the personal responsibility to embrace and uphold the provisions of the Code and Standards and are encouraged to notify their employer of this responsibility. Violations may result in disciplinary sanctions by CFA Institute. Sanctions can include revocation of membership, revocation of candidacy in the CFA Program, and revocation of the right to use the CFA designation.

The Code of Ethics

Members of CFA Institute (including CFA charterholders) and candidates for the CFA designation ("Members and Candidates") must:

- Act with integrity, competence, diligence, and respect and in an ethical manner with the public, clients, prospective clients, employers, employees, colleagues in the investment profession, and other participants in the global capital markets.

- Place the integrity of the investment profession and the interests of clients above their own personal interests.
- Use reasonable care and exercise independent professional judgment when conducting investment analysis, making investment recommendations, taking investment actions, and engaging in other professional activities.
- Practice and encourage others to practice in a professional and ethical manner that will reflect credit on themselves and the profession.
- Promote the integrity and viability of the global capital markets for the ultimate benefit of society.
- Maintain and improve their professional competence and strive to maintain and improve the competence of other investment professionals.

Standards of Professional Conduct

I. PROFESSIONALISM

A Knowledge of the Law

Members and Candidates must understand and comply with all applicable laws, rules, and regulations (including the CFA Institute Code of Ethics and Standards of Professional Conduct) of any government, regulatory organization, licensing agency, or professional association governing their professional activities. In the event of conflict, Members and Candidates must comply with the more strict law, rule, or regulation. Members and Candidates must not knowingly participate or assist in and must dissociate from any violation of such laws, rules, or regulations.

B Independence and Objectivity

Members and Candidates must use reasonable care and judgment to achieve and maintain independence and objectivity in their professional activities. Members and Candidates must not offer, solicit, or accept any gift, benefit, compensation, or consideration that reasonably could be expected to compromise their own or another's independence and objectivity.

C Misrepresentation

Members and Candidates must not knowingly make any misrepresentations relating to investment analysis, recommendations, actions, or other professional activities.

D Misconduct

Members and Candidates must not engage in any professional conduct involving dishonesty, fraud, or deceit or commit any act that reflects adversely on their professional reputation, integrity, or competence.

II. INTEGRITY OF CAPITAL MARKETS

A Material Nonpublic Information

Members and Candidates who possess material nonpublic information that could affect the value of an investment must not act or cause others to act on the information.

B Market Manipulation

Members and Candidates must not engage in practices that distort prices or artificially inflate trading volume with the intent to mislead market participants.

III. DUTIES TO CLIENTS

A Loyalty, Prudence, and Care

Members and Candidates have a duty of loyalty to their clients and must act with reasonable care and exercise prudent judgment. Members and Candidates must act for the benefit of their clients and place their clients' interests before their employer's or their own interests.

B Fair Dealing

Members and Candidates must deal fairly and objectively with all clients when providing investment analysis, making investment recommendations, taking investment action, or engaging in other professional activities.

C Suitability

1 When Members and Candidates are in an advisory relationship with a client, they must:

a Make a reasonable inquiry into a client's or prospective client's investment experience, risk and return objectives, and financial constraints prior to making any investment recommendation or taking investment action and must reassess and update this information regularly.

b Determine that an investment is suitable to the client's financial situation and consistent with the client's written objectives, mandates, and constraints before making an investment recommendation or taking investment action.

c Judge the suitability of investments in the context of the client's total portfolio.

2 When Members and Candidates are responsible for managing a portfolio to a specific mandate, strategy, or style, they must make only investment recommendations or take only investment actions that are consistent with the stated objectives and constraints of the portfolio.

D Performance Presentation

When communicating investment performance information, Members and Candidates must make reasonable efforts to ensure that it is fair, accurate, and complete.

E Preservation of Confidentiality

Members and Candidates must keep information about current, former, and prospective clients confidential unless:

1 The information concerns illegal activities on the part of the client or prospective client,

2 Disclosure is required by law, or

3 The client or prospective client permits disclosure of the information.

IV. DUTIES TO EMPLOYERS

A Loyalty

In matters related to their employment, Members and Candidates must act for the benefit of their employer and not deprive their employer of the advantage of their skills and abilities, divulge confidential information, or otherwise cause harm to their employer.

B Additional Compensation Arrangements

Members and Candidates must not accept gifts, benefits, compensation, or consideration that competes with or might reasonably be expected to create a conflict of interest with their employer's interest unless they obtain written consent from all parties involved.

C Responsibilities of Supervisors

Members and Candidates must make reasonable efforts to ensure that anyone subject to their supervision or authority complies with applicable laws, rules, regulations, and the Code and Standards.

V. INVESTMENT ANALYSIS, RECOMMENDATIONS, AND ACTIONS

A Diligence and Reasonable Basis

Members and Candidates must:

1 Exercise diligence, independence, and thoroughness in analyzing investments, making investment recommendations, and taking investment actions.

2 Have a reasonable and adequate basis, supported by appropriate research and investigation, for any investment analysis, recommendation, or action.

B Communication with Clients and Prospective Clients

Members and Candidates must:

1 Disclose to clients and prospective clients the basic format and general principles of the investment processes they use to analyze investments, select securities, and construct portfolios and must promptly disclose any changes that might materially affect those processes.

2 Disclose to clients and prospective clients significant limitations and risks associated with the investment process.

3 Use reasonable judgment in identifying which factors are important to their investment analyses, recommendations, or actions and include those factors in communications with clients and prospective clients.

4 Distinguish between fact and opinion in the presentation of investment analysis and recommendations.

C Record Retention

Members and Candidates must develop and maintain appropriate records to support their investment analyses, recommendations, actions, and other investment-related communications with clients and prospective clients.

VI. CONFLICTS OF INTEREST

A Disclosure of Conflicts

Members and Candidates must make full and fair disclosure of all matters that could reasonably be expected to impair their independence and objectivity or interfere with respective duties to their clients, prospective clients, and employer. Members and Candidates must ensure that such disclosures are prominent, are delivered in plain language, and communicate the relevant information effectively.

B Priority of Transactions

Investment transactions for clients and employers must have priority over investment transactions in which a Member or Candidate is the beneficial owner.

C Referral Fees

Members and Candidates must disclose to their employer, clients, and prospective clients, as appropriate, any compensation, consideration, or benefit received from or paid to others for the recommendation of products or services.

VII. RESPONSIBILITIES AS A CFA INSTITUTE MEMBER OR CFA CANDIDATE

A Conduct as Participants in CFA Institute Programs

Members and Candidates must not engage in any conduct that compromises the reputation or integrity of CFA Institute or the CFA designation or the integrity, validity, or security of CFA Institute programs.

B Reference to CFA Institute, the CFA Designation, and the CFA Program

When referring to CFA Institute, CFA Institute membership, the CFA designation, or candidacy in the CFA Program, Members and Candidates must not misrepresent or exaggerate the meaning or implications of membership in CFA Institute, holding the CFA designation, or candidacy in the CFA Program.

READING

2

Guidance for Standards I–VII

LEARNING OUTCOMES

Mastery	The candidate should be able to:
☐	**a.** demonstrate a thorough knowledge of the CFA Institute Code of Ethics and Standards of Professional Conduct by applying the Code and Standards to specific situations;
☐	**b.** recommend practices and procedures designed to prevent violations of the Code of Ethics and Standards of Professional Conduct.

STANDARD I: PROFESSIONALISM

Standard I(A) Knowledge of the Law

Members and Candidates must understand and comply with all applicable laws, rules, and regulations (including the CFA Institute Code of Ethics and Standards of Professional Conduct) of any government, regulatory organization, licensing agency, or professional association governing their professional activities. In the event of conflict, Members and Candidates must comply with the more strict law, rule, or regulation. Members and Candidates must not knowingly participate or assist in and must dissociate from any violation of such laws, rules, or regulations.

Guidance

Highlights:

- *Relationship between the Code and Standards and Applicable Law*
- *Participation in or Association with Violations by Others*
- *Investment Products and Applicable Laws*

Members and candidates must understand the applicable laws and regulations of the countries and jurisdictions where they engage in professional activities. These activities may include, but are not limited to, trading of securities or other financial instruments, providing investment advice, conducting research, or performing other investment services. On the basis of their reasonable and good faith understanding, members and candidates must comply with the laws and regulations that directly govern their professional activities and resulting outcomes and that protect the interests of the clients.

When questions arise, members and candidates should know their firm's policies and procedures for accessing compliance guidance. This standard does not require members and candidates to become experts, however, in compliance. Additionally, members and candidates are not required to have detailed knowledge of or be experts on all the laws that could potentially govern their activities.

During times of changing regulations, members and candidates must remain vigilant in maintaining their knowledge of the requirements for their professional activities. New financial products and processes, along with uncovered ethical missteps, create an environment for recurring and potentially wide-ranging regulatory changes. Members and candidates are also continually provided improved and enhanced methods of communicating with both clients and potential clients, such as mobile applications and web-based social networking platforms. As new local, regional, and global requirements are updated to address these and other changes, members, candidates, and their firms must adjust their procedures and practices to remain in compliance.

Relationship between the Code and Standards and Applicable Law

Some members or candidates may live, work, or provide investment services to clients living in a country that has no law or regulation governing a particular action or that has laws or regulations that differ from the requirements of the Code and Standards. When applicable law and the Code and Standards require different conduct, members and candidates must follow the more strict of the applicable law or the Code and Standards.

"Applicable law" is the law that governs the member's or candidate's conduct. Which law applies will depend on the particular facts and circumstances of each case. The "more strict" law or regulation is the law or regulation that imposes greater restrictions on the action of the member or candidate or calls for the member or candidate to exert a greater degree of action that protects the interests of investors. For example, applicable law or regulation may not require members and candidates to disclose referral fees received from or paid to others for the recommendation of investment products or services. Because the Code and Standards impose this obligation, however, members and candidates must disclose the existence of such fees.

Members and candidates must adhere to the following principles:

- Members and candidates must comply with applicable laws or regulations related to their professional activities.
- Members and candidates must not engage in conduct that constitutes a violation of the Code and Standards, even though it may otherwise be legal.
- In the absence of any applicable law or regulation or when the Code and Standards impose a higher degree of responsibility than applicable laws and regulations, members and candidates must adhere to the Code and Standards. Applications of these principles are outlined in Exhibit 1.

The applicable laws governing the responsibilities of a member or candidate should be viewed as the minimal threshold of acceptable actions. When members and candidates take actions that exceed the minimal requirements, they further support the conduct required of Standard I(A).

CFA Institute members are obligated to abide by the CFA Institute Articles of Incorporation, Bylaws, Code of Ethics, Standards of Professional Conduct, Rules of Procedure, Membership Agreement, and other applicable rules promulgated by CFA Institute, all as amended periodically. CFA candidates who are not members must also abide by these documents (except for the Membership Agreement) as well as rules and regulations related to the administration of the CFA examination, the Candidate Responsibility Statement, and the Candidate Pledge.

Participation in or Association with Violations by Others

Members and candidates are responsible for violations in which they *knowingly* participate or assist. Although members and candidates are presumed to have knowledge of all applicable laws, rules, and regulations, CFA Institute acknowledges that members may not recognize violations if they are not aware of all the facts giving rise to the violations. Standard I(A) applies when members and candidates know or should know that their conduct may contribute to a violation of applicable laws, rules, or regulations or the Code and Standards.

If a member or candidate has reasonable grounds to believe that imminent or ongoing client or employer activities are illegal or unethical, the member or candidate must dissociate, or separate, from the activity. In extreme cases, dissociation may require a member or candidate to leave his or her employment. Members and candidates may take the following intermediate steps to dissociate from ethical violations of others when direct discussions with the person or persons committing the violation are unsuccessful. The first step should be to attempt to stop the behavior by bringing it to the attention of the employer through a supervisor or the firm's compliance department. If this attempt is unsuccessful, then members and candidates have a responsibility to step away and dissociate from the activity. Dissociation practices will differ on the basis of the member's or candidate's role in the investment industry. It may include removing one's name from written reports or recommendations, asking for a different assignment, or refusing to accept a new client or continue to advise a current client. Inaction combined with continuing association with those involved in illegal or unethical conduct may be construed as participation or assistance in the illegal or unethical conduct.

CFA Institute strongly encourages members and candidates to report potential violations of the Code and Standards committed by fellow members and candidates. Although a failure to report is less likely to be construed as a violation than a failure to dissociate from unethical conduct, the impact of inactivity on the integrity of capital markets can be significant. Although the Code and Standards do not compel members and candidates to report violations to their governmental or regulatory organizations unless such disclosure is mandatory under applicable law (voluntary reporting is often referred to as whistleblowing), such disclosure may be prudent under certain circumstances. Members and candidates should consult their legal and compliance advisers for guidance.

Additionally, CFA Institute encourages members, nonmembers, clients, and the investing public to report violations of the Code and Standards by CFA Institute members or CFA candidates by submitting a complaint in writing to the CFA Institute Professional Conduct Program via e-mail (pcprogram@cfainstitute.org) or the CFA Institute website (www.cfainstitute.org).

Investment Products and Applicable Laws

Members and candidates involved in creating or maintaining investment services or investment products or packages of securities and/or derivatives should be mindful of where these products or packages will be sold as well as their places of origination. The applicable laws and regulations of the countries or regions of origination and expected sale should be understood by those responsible for the supervision of the services or creation and maintenance of the products or packages. Members or candidates should make reasonable efforts to review whether associated firms that are distributing products or services developed by their employing firm also abide by the laws and regulations of the countries and regions of distribution. Members and candidates should undertake the necessary due diligence when transacting cross-border business to understand the multiple applicable laws and regulations in order to protect the reputation of their firm and themselves.

Given the complexity that can arise with business transactions in today's market, there may be some uncertainty surrounding which laws or regulations are considered applicable when activities are being conducted in multiple jurisdictions. Members and candidates should seek the appropriate guidance, potentially including the firm's compliance or legal departments and legal counsel outside the organization, to gain a reasonable understanding of their responsibilities and how to implement them appropriately.

Exhibit 1 Global Application of the Code and Standards

Members and candidates who practice in multiple jurisdictions may be subject to varied securities laws and regulations. If applicable law is stricter than the requirements of the Code and Standards, members and candidates must adhere to applicable law; otherwise, they must adhere to the Code and Standards. The following chart provides illustrations involving a member who may be subject to the securities laws and regulations of three different types of countries:

NS: country with no securities laws or regulations

LS: country with ***less*** strict securities laws and regulations than the Code and Standards

MS: country with ***more*** strict securities laws and regulations than the Code and Standards

Applicable Law	Duties	Explanation
Member resides in NS country, does business in LS country; LS law applies.	Member must adhere to the Code and Standards.	Because applicable law is less strict than the Code and Standards, the member must adhere to the Code and Standards.
Member resides in NS country, does business in MS country; MS law applies.	Member must adhere to the law of MS country.	Because applicable law is stricter than the Code and Standards, member must adhere to the more strict applicable law.
Member resides in LS country, does business in NS country; LS law applies.	Member must adhere to the Code and Standards.	Because applicable law is less strict than the Code and Standards, member must adhere to the Code and Standards.
Member resides in LS country, does business in MS country; MS law applies.	Member must adhere to the law of MS country.	Because applicable law is stricter than the Code and Standards, member must adhere to the more strict applicable law.

Exhibit 1 (Continued)

Applicable Law	Duties	Explanation
Member resides in LS country, does business in NS country; LS law applies, but it states that law of locality where business is conducted governs.	Member must adhere to the Code and Standards.	Because applicable law states that the law of the locality where the business is conducted governs and there is no local law, the member must adhere to the Code and Standards.
Member resides in LS country, does business in MS country; LS law applies, but it states that law of locality where business is conducted governs.	Member must adhere to the law of MS country.	Because applicable law of the locality where the business is conducted governs and local law is stricter than the Code and Standards, member must adhere to the more strict applicable law.
Member resides in MS country, does business in LS country; MS law applies.	Member must adhere to the law of MS country.	Because applicable law is stricter than the Code and Standards, member must adhere to the more strict applicable law.
Member resides in MS country, does business in LS country; MS law applies, but it states that law of locality where business is conducted governs.	Member must adhere to the Code and Standards.	Because applicable law states that the law of the locality where the business is conducted governs and local law is less strict than the Code and Standards, member must adhere to the Code and Standards.
Member resides in MS country, does business in LS country with a client who is a citizen of LS country; MS law applies, but it states that the law of the client's home country governs.	Member must adhere to the Code and Standards.	Because applicable law states that the law of the client's home country governs (which is less strict than the Code and Standards), member must adhere to the Code and Standards.
Member resides in MS country, does business in LS country with a client who is a citizen of MS country; MS law applies, but it states that the law of the client's home country governs.	Member must adhere to the law of MS country.	Because applicable law states that the law of the client's home country governs and the law of the client's home country is stricter than the Code and Standards, the member must adhere to the more strict applicable law.

Recommended Procedures for Compliance

Members and Candidates

Suggested methods by which members and candidates can acquire and maintain understanding of applicable laws, rules, and regulations include the following:

- *Stay informed*: Members and candidates should establish or encourage their employers to establish a procedure by which employees are regularly informed about changes in applicable laws, rules, regulations, and case law. In many instances, the employer's compliance department or legal counsel can provide such information in the form of memorandums distributed to employees in the organization. Also, participation in an internal or external continuing education program is a practical method of staying current.
- *Review procedures*: Members and candidates should review, or encourage their employers to review, the firm's written compliance procedures on a regular basis to ensure that the procedures reflect current law and provide adequate guidance to employees about what is permissible conduct under the law and/or the Code and Standards. Recommended compliance procedures for specific items of the Code and Standards are discussed in this *Handbook* in the "Guidance" sections associated with each standard.
- *Maintain current files*: Members and candidates should maintain or encourage their employers to maintain readily accessible current reference copies of applicable statutes, rules, regulations, and important cases.

Distribution Area Laws

Members and candidates should make reasonable efforts to understand the applicable laws—both country and regional—for the countries and regions where their investment products are developed and are most likely to be distributed to clients.

Legal Counsel

When in doubt about the appropriate action to undertake, it is recommended that a member or candidate seek the advice of compliance personnel or legal counsel concerning legal requirements. If a potential violation is being committed by a fellow employee, it may also be prudent for the member or candidate to seek the advice of the firm's compliance department or legal counsel.

Dissociation

When dissociating from an activity that violates the Code and Standards, members and candidates should document the violation and urge their firms to attempt to persuade the perpetrator(s) to cease such conduct. To dissociate from the conduct, a member or candidate may have to resign his or her employment.

Firms

The formality and complexity of compliance procedures for firms depend on the nature and size of the organization and the nature of its investment operations. Members and candidates should encourage their firms to consider the following policies and procedures to support the principles of Standard I(A):

- *Develop and/or adopt a code of ethics*: The ethical culture of an organization starts at the top. Members and candidates should encourage their supervisors or managers to adopt a code of ethics. Adhering to a code of ethics facilitates solutions when people face ethical dilemmas and can prevent the need for employees to resort to a "whistleblowing" solution publicly alleging concealed misconduct. CFA Institute has published the *Asset Manager Code of Professional Conduct*, which firms may adopt or use as the basis for their codes (visit www.cfainstitute.org).

- *Provide information on applicable laws*: Pertinent information that highlights applicable laws and regulations might be distributed to employees or made available in a central location. Information sources might include primary information developed by the relevant government, governmental agencies, regulatory organizations, licensing agencies, and professional associations (e.g., from their websites); law firm memorandums or newsletters; and association memorandums or publications (e.g., *CFA Institute Magazine*).
- *Establish procedures for reporting violations*: Firms might provide written protocols for reporting suspected violations of laws, regulations, or company policies.

Application of the Standard

Example 1 (Notification of Known Violations):

Michael Allen works for a brokerage firm and is responsible for an underwriting of securities. A company official gives Allen information indicating that the financial statements Allen filed with the regulator overstate the issuer's earnings. Allen seeks the advice of the brokerage firm's general counsel, who states that it would be difficult for the regulator to prove that Allen has been involved in any wrongdoing.

> *Comment*: Although it is recommended that members and candidates seek the advice of legal counsel, the reliance on such advice does not absolve a member or candidate from the requirement to comply with the law or regulation. Allen should report this situation to his supervisor, seek an independent legal opinion, and determine whether the regulator should be notified of the error.

Example 2 (Dissociating from a Violation):

Lawrence Brown's employer, an investment banking firm, is the principal underwriter for an issue of convertible debentures by the Courtney Company. Brown discovers that the Courtney Company has concealed severe third-quarter losses in its foreign operations. The preliminary prospectus has already been distributed.

> *Comment*: Knowing that the preliminary prospectus is misleading, Brown should report his findings to the appropriate supervisory persons in his firm. If the matter is not remedied and Brown's employer does not dissociate from the underwriting, Brown should sever all his connections with the underwriting. Brown should also seek legal advice to determine whether additional reporting or other action should be taken.

Example 3 (Dissociating from a Violation):

Kamisha Washington's firm advertises its past performance record by showing the 10-year return of a composite of its client accounts. Washington discovers, however, that the composite omits the performance of accounts that have left the firm during the 10-year period, whereas the description of the composite indicates the inclusion of all firm accounts. This omission has led to an inflated performance figure. Washington is asked to use promotional material that includes the erroneous performance number when soliciting business for the firm.

Comment: Misrepresenting performance is a violation of the Code and Standards. Although she did not calculate the performance herself, Washington would be assisting in violating Standard I(A) if she were to use the inflated performance number when soliciting clients. She must dissociate herself from the activity. If discussing the misleading number with the person responsible is not an option for correcting the problem, she can bring the situation to the attention of her supervisor or the compliance department at her firm. If her firm is unwilling to recalculate performance, she must refrain from using the misleading promotional material and should notify the firm of her reasons. If the firm insists that she use the material, she should consider whether her obligation to dissociate from the activity requires her to seek other employment.

Example 4 (Following the Highest Requirements):

James Collins is an investment analyst for a major Wall Street brokerage firm. He works in a developing country with a rapidly modernizing economy and a growing capital market. Local securities laws are minimal—in form and content—and include no punitive prohibitions against insider trading.

Comment: Collins must abide by the requirements of the Code and Standards, which might be more strict than the rules of the developing country. He should be aware of the risks that a small market and the absence of a fairly regulated flow of information to the market represent to his ability to obtain information and make timely judgments. He should include this factor in formulating his advice to clients. In handling material nonpublic information that accidentally comes into his possession, he must follow Standard II(A)–Material Nonpublic Information.

Example 5 (Following the Highest Requirements):

Laura Jameson works for a multinational investment adviser based in the United States. Jameson lives and works as a registered investment adviser in the tiny, but wealthy, island nation of Karramba. Karramba's securities laws state that no investment adviser registered and working in that country can participate in initial public offerings (IPOs) for the adviser's personal account. Jameson, believing that, as a US citizen working for a US-based company, she should comply only with US law, has ignored this Karrambian law. In addition, Jameson believes that as a charterholder, as long as she adheres to the Code and Standards requirement that she disclose her participation in any IPO to her employer and clients when such ownership creates a conflict of interest, she is meeting the highest ethical requirements.

Comment: Jameson is in violation of Standard I(A). As a registered investment adviser in Karramba, Jameson is prevented by Karrambian securities law from participating in IPOs regardless of the law of her home country. In addition, because the law of the country where she is working is stricter than the Code and Standards, she must follow the stricter requirements of the local law rather than the requirements of the Code and Standards.

Example 6 (Laws and Regulations Based on Religious Tenets):

Amanda Janney is employed as a fixed-income portfolio manager for a large international firm. She is on a team within her firm that is responsible for creating and managing a fixed-income hedge fund to be sold throughout the firm's distribution centers to high-net-worth clients. Her firm receives expressions of interest from potential clients from the Middle East who are seeking investments that comply with Islamic law. The marketing and promotional materials for the fixed-income hedge fund do not specify whether or not the fund is a suitable investment for an investor seeking compliance with Islamic law. Because the fund is being distributed globally, Janney is concerned about the reputation of the fund and the firm and believes disclosure of whether or not the fund complies with Islamic law could help minimize potential mistakes with placing this investment.

> *Comment*: As the financial market continues to become globalized, members and candidates will need to be aware of the differences between cultural and religious laws and requirements as well as the different governmental laws and regulations. Janney and the firm could be proactive in their efforts to acknowledge areas where the new fund may not be suitable for clients.

Example 7 (Reporting Potential Unethical Actions):

Krista Blume is a junior portfolio manager for high-net-worth portfolios at a large global investment manager. She observes a number of new portfolios and relationships coming from a country in Europe where the firm did not have previous business and is told that a broker in that country is responsible for this new business. At a meeting on allocation of research resources to third-party research firms, Blume notes that this broker has been added to the list and is allocated payments for research. However, she knows the portfolios do not invest in securities in the broker's country, and she has not seen any research come from this broker. Blume asks her supervisor about the name being on the list and is told that someone in marketing is receiving the research and that the name being on the list is OK. She believes that what may be going on is that the broker is being paid for new business through the inappropriate research payments, and she wishes to dissociate from the misconduct.

> *Comment*: Blume should follow the firm's policies and procedures for reporting potential unethical activity, which may include discussions with her supervisor or someone in a designated compliance department. She should communicate her concerns appropriately while advocating for disclosure between the new broker relationship and the research payments.

Example 8 (Failure to Maintain Knowledge of the Law):

Colleen White is excited to use new technology to communicate with clients and potential clients. She recently began posting investment information, including performance reports and investment opinions and recommendations, to her Facebook page. In addition, she sends out brief announcements, opinions, and thoughts via her Twitter account (for example, "Prospects for future growth of XYZ company look good! #makingmoney4U"). Prior to White's use of these social media platforms, the local regulator had issued new requirements and guidance governing online electronic communication. White's communications appear to conflict with the recent regulatory announcements.

> *Comment*: White is in violation of Standard I(A) because her communications do not comply with the existing guidance and regulation governing use of social media. White must be aware of the evolving legal requirements pertaining to new and dynamic areas of the financial services industry that are applicable to her. She should seek guidance from appropriate, knowledgeable, and reliable sources, such as her firm's compliance department, external service providers, or outside counsel, unless she diligently follows legal and regulatory trends affecting her professional responsibilities.

Standard I(B) Independence and Objectivity

Members and Candidates must use reasonable care and judgment to achieve and maintain independence and objectivity in their professional activities. Members and Candidates must not offer, solicit, or accept any gift, benefit, compensation, or consideration that reasonably could be expected to compromise their own or another's independence and objectivity.

Guidance

Highlights:

- *Buy-Side Clients*
- *Fund Manager and Custodial Relationships*
- *Investment Banking Relationships*
- *Performance Measurement and Attribution*
- *Public Companies*
- *Credit Rating Agency Opinions*
- *Influence during the Manager Selection/Procurement Process*
- *Issuer-Paid Research*
- *Travel Funding*

Standard I(B) states the responsibility of CFA Institute members and candidates in the CFA Program to maintain independence and objectivity so that their clients will have the benefit of their work and opinions unaffected by any potential conflict of interest or other circumstance adversely affecting their judgment. Every member and candidate should endeavor to avoid situations that could cause or be perceived to cause a loss of independence or objectivity in recommending investments or taking investment action.

External sources may try to influence the investment process by offering analysts and portfolio managers a variety of benefits. Corporations may seek expanded research coverage, issuers and underwriters may wish to promote new securities offerings, brokers may want to increase commission business, and independent rating agencies may be influenced by the company requesting the rating. Benefits may include gifts, invitations to lavish functions, tickets, favors, or job referrals. One type of benefit is the allocation of shares in oversubscribed IPOs to investment managers for their personal accounts. This practice affords managers the opportunity to make quick profits that may not be available to their clients. Such a practice is prohibited under Standard I(B). Modest gifts and entertainment are acceptable, but special care must be taken by members and candidates to resist subtle and not-so-subtle pressures to act in conflict with the interests of their clients. Best practice dictates that members and candidates reject any offer of gift or entertainment that could be expected to threaten their independence and objectivity.

Receiving a gift, benefit, or consideration from a *client* can be distinguished from gifts given by entities seeking to influence a member or candidate to the detriment of other clients. In a client relationship, the client has already entered some type of compensation arrangement with the member, candidate, or his or her firm. A gift from a client could be considered supplementary compensation. The potential for obtaining influence to the detriment of other clients, although present, is not as great as in situations where no compensation arrangement exists. When possible, prior to accepting "bonuses" or gifts from clients, members and candidates should disclose to their employers such benefits offered by clients. If notification is not possible prior to acceptance, members and candidates must disclose to their employer benefits previously accepted from clients. Disclosure allows the employer of a member or candidate to make an independent determination about the extent to which the gift may affect the member's or candidate's independence and objectivity.

Members and candidates may also come under pressure from their own firms to, for example, issue favorable research reports or recommendations for certain companies with potential or continuing business relationships with the firm. The situation may be aggravated if an executive of the company sits on the bank or investment firm's board and attempts to interfere in investment decision making. Members and candidates acting in a sales or marketing capacity must be especially mindful of their objectivity in promoting appropriate investments for their clients.

Left unmanaged, pressures that threaten independence place research analysts in a difficult position and may jeopardize their ability to act independently and objectively. One of the ways that research analysts have coped with these pressures in the past is to use subtle and ambiguous language in their recommendations or to temper the tone of their research reports. Such subtleties are lost on some investors, however, who reasonably expect research reports and recommendations to be straightforward and transparent and to communicate clearly an analyst's views based on unbiased analysis and independent judgment.

Members and candidates are personally responsible for maintaining independence and objectivity when preparing research reports, making investment recommendations, and taking investment action on behalf of clients. Recommendations must convey the member's or candidate's true opinions, free of bias from internal or external pressures, and be stated in clear and unambiguous language.

Members and candidates also should be aware that some of their professional or social activities within CFA Institute or its member societies may subtly threaten their independence or objectivity. When seeking corporate financial support for conventions, seminars, or even weekly society luncheons, the members or candidates responsible for the activities should evaluate both the actual effect of such solicitations on their independence and whether their objectivity might be perceived to be compromised in the eyes of their clients.

Buy-Side Clients

One source of pressure on sell-side analysts is buy-side clients. Institutional clients are traditionally the primary users of sell-side research, either directly or with soft dollar brokerage. Portfolio managers may have significant positions in the security of a company under review. A rating downgrade may adversely affect the portfolio's performance, particularly in the short term, because the sensitivity of stock prices to ratings changes has increased in recent years. A downgrade may also affect the manager's compensation, which is usually tied to portfolio performance. Moreover, portfolio performance is subject to media and public scrutiny, which may affect the manager's professional reputation. Consequently, some portfolio managers implicitly or explicitly support sell-side ratings inflation.

Portfolio managers have a responsibility to respect and foster the intellectual honesty of sell-side research. Therefore, it is improper for portfolio managers to threaten or engage in retaliatory practices, such as reporting sell-side analysts to the covered company in order to instigate negative corporate reactions. Although most portfolio managers do not engage in such practices, the perception by the research analyst that a reprisal is possible may cause concern and make it difficult for the analyst to maintain independence and objectivity.

Fund Manager and Custodial Relationships

Research analysts are not the only people who must be concerned with maintaining their independence. Members and candidates who are responsible for hiring and retaining outside managers and third-party custodians should not accepts gifts, entertainment, or travel funding that may be perceived as impairing their decisions. The use of secondary fund managers has evolved into a common practice to manage specific asset allocations. The use of third-party custodians is common practice for independent investment advisory firms and helps them with trading capabilities and reporting requirements. Primary and secondary fund managers, as well as third-party custodians, often arrange educational and marketing events to inform others about their business strategies, investment process, or custodial services. Members and candidates must review the merits of each offer individually in determining whether they may attend yet maintain their independence.

Investment Banking Relationships

Some sell-side firms may exert pressure on their analysts to issue favorable research reports on current or prospective investment banking clients. For many of these firms, income from investment banking has become increasingly important to overall firm profitability because brokerage income has declined as a result of price competition. Consequently, firms offering investment banking services work hard to develop and maintain relationships with investment banking clients and prospects. These companies are often covered by the firm's research analysts because companies often select their investment banks on the basis of the reputation of their research analysts, the quality of their work, and their standing in the industry.

In some countries, research analysts frequently work closely with their investment banking colleagues to help evaluate prospective investment banking clients. In other countries, because of past abuses in managing the obvious conflicts of interest, regulators have established clear rules prohibiting the interaction of these groups. Although collaboration between research analysts and investment banking colleagues may benefit the firm and enhance market efficiency (e.g., by allowing firms to assess risks more accurately and make better pricing assumptions), it requires firms to carefully balance the conflicts of interest inherent in the collaboration. Having analysts work with investment bankers is appropriate only when the conflicts are adequately and effectively managed and disclosed. Firm managers have a responsibility to provide an environment in which analysts are neither coerced nor enticed into issuing research that does not reflect their true opinions. Firms should require public disclosure of actual conflicts of interest to investors.

Members, candidates, and their firms must adopt and follow perceived best practices in maintaining independence and objectivity in the corporate culture and protecting analysts from undue pressure by their investment banking colleagues. The "firewalls" traditionally built between these two functions must be managed to minimize conflicts of interest; indeed, enhanced firewall policies may go as far as prohibiting all communications between these groups. A key element of an enhanced firewall is separate reporting structures for personnel on the research side and personnel on the investment banking side. For example, investment banking personnel should not have any authority to approve, disapprove, or make changes to research reports or recommendations. Another element should be a compensation arrangement that minimizes the pressures on research analysts and rewards objectivity and accuracy. Compensation arrangements should not link analyst remuneration directly to investment banking assignments in which the analyst may participate as a team member. Firms should also regularly review their policies and procedures to determine whether analysts are adequately safeguarded and to improve the transparency of disclosures relating to conflicts of interest. The highest level of transparency is achieved when disclosures are prominent and specific rather than marginalized and generic.

Performance Measurement and Attribution

Members and candidates working within a firm's investment performance measurement department may also be presented with situations that challenge their independence and objectivity. As performance analysts, their analyses may reveal instances where managers may appear to have strayed from their mandate. Additionally, the performance analyst may receive requests to alter the construction of composite indexes owing to negative results for a selected account or fund. The member or candidate must not allow internal or external influences to affect their independence and objectivity as they faithfully complete their performance calculation and analysis-related responsibilities.

Public Companies

Analysts may be pressured to issue favorable reports and recommendations by the companies they follow. Not every stock is a "buy," and not every research report is favorable—for many reasons, including the cyclical nature of many business activities and market fluctuations. For instance, a "good company" does not always translate into a "good stock" rating if the current stock price is fully valued. In making an investment recommendation, the analyst is responsible for anticipating, interpreting, and assessing a company's prospects and stock price performance in a factual manner. Many company managers, however, believe that their company's stock is undervalued, and these managers may find it difficult to accept critical research reports or ratings downgrades. Company managers' compensation may also be dependent on stock performance.

Due diligence in financial research and analysis involves gathering information from a wide variety of sources, including public disclosure documents (such as proxy statements, annual reports, and other regulatory filings) and also company management and investor-relations personnel, suppliers, customers, competitors, and other relevant sources. Research analysts may justifiably fear that companies will limit their ability to conduct thorough research by denying analysts who have "negative" views direct access to company managers and/or barring them from conference calls and other communication venues. Retaliatory practices include companies bringing legal action against analysts personally and/or their firms to seek monetary damages for the economic effects of negative reports and recommendations. Although few companies engage in such behavior, the perception that a reprisal is possible is a reasonable concern for analysts. This concern may make it difficult for them to conduct the comprehensive research needed to make objective recommendations. For further information and guidance, members and candidates should refer to the CFA Institute publication *Best Practice Guidelines Governing Analyst/Corporate Issuer Relations* (www.cfainstitute.org).

Credit Rating Agency Opinions

Credit rating agencies provide a service by grading the fixed-income products offered by companies. Analysts face challenges related to incentives and compensation schemes that may be tied to the final rating and successful placement of the product. Members and candidates employed at rating agencies should ensure that procedures and processes at the agencies prevent undue influences from a sponsoring company during the analysis. Members and candidates should abide by their agencies' and the industry's standards of conduct regarding the analytical process and the distribution of their reports.

The work of credit rating agencies also raises concerns similar to those inherent in investment banking relationships. Analysts may face pressure to issue ratings at a specific level because of other services the agency offers companies—namely, advising on the development of structured products. The rating agencies need to develop the necessary firewalls and protections to allow the independent operations of their different business lines.

When using information provided by credit rating agencies, members and candidates should be mindful of the potential conflicts of interest. And because of the potential conflicts, members and candidates may need to independently validate the rating granted.

Influence during the Manager Selection/Procurement Process

Members and candidates may find themselves on either side of the manager selection process. An individual may be on the hiring side as a representative of a pension organization or an investment committee member of an endowment or a charitable organization. Additionally, other members may be representing their organizations in attempts to earn new investment allocation mandates. The responsibility of members and candidates to maintain their independence and objectivity extends to the hiring or firing of those who provide business services beyond investment management.

When serving in a hiring capacity, members and candidates should not solicit gifts, contributions, or other compensation that may affect their independence and objectivity. Solicitations do not have to benefit members and candidates personally to conflict with Standard I(B). Requesting contributions to a favorite charity or political organization may also be perceived as an attempt to influence the decision-making process. Additionally, members and candidates serving in a hiring capacity should refuse gifts, donations, and other offered compensation that may be perceived to influence their decision-making process.

When working to earn a new investment allocation, members and candidates should not offer gifts, contributions, or other compensation to influence the decision of the hiring representative. The offering of these items with the intent to impair the independence and objectivity of another person would not comply with Standard I(B). Such prohibited actions may include offering donations to a charitable organization or political candidate referred by the hiring representative.

A clear example of improperly influencing hiring representatives was displayed in the "pay-to-play" scandal involving government-sponsored pension funds in the United States. Managers looking to gain lucrative allocations from the large funds made requested donations to the political campaigns of individuals directly responsible for the hiring decisions. This scandal and other similar events have led to new laws requiring additional reporting concerning political contributions and bans on hiring—or hiring delays for—managers that made campaign contributions to representatives associated with the decision-making process.

Issuer-Paid Research

In light of the recent reduction of sell-side research coverage, many companies, seeking to increase visibility both in the financial markets and with potential investors, have hired analysts to produce research reports analyzing their companies. These reports bridge the gap created by the lack of coverage and can be an effective method of communicating with investors.

Issuer-paid research conducted by independent analysts, however, is fraught with potential conflicts. Depending on how the research is written and distributed, investors may be misled into believing that the research is from an independent source when, in reality, it has been paid for by the subject company.

Members and candidates must adhere to strict standards of conduct that govern how the research is to be conducted and what disclosures must be made in the report. Analysts must engage in thorough, independent, and unbiased analysis and must fully disclose potential conflicts of interest, including the nature of their compensation. Otherwise, analysts risk misleading investors.

Investors need clear, credible, and thorough information about companies, and they need research based on independent thought. At a minimum, issuer-paid research should include a thorough analysis of the company's financial statements based on publicly disclosed information, benchmarking within a peer group, and industry analysis. Analysts must exercise diligence, independence, and thoroughness in conducting their research in an objective manner. Analysts must distinguish between fact and opinion in their reports. Conclusions must have a reasonable and adequate basis and must be supported by appropriate research.

Independent analysts must also strictly limit the type of compensation that they accept for conducting issuer-paid research. Otherwise, the content and conclusions of the reports could reasonably be expected to be determined or affected by compensation from the sponsoring companies. Compensation that might influence the research report could be direct, such as payment based on the conclusions of the report, or indirect, such as stock warrants or other equity instruments that could increase in value on the basis of positive coverage in the report. In such instances, the independent analyst has an incentive to avoid including negative information or making negative conclusions. Best practice is for independent analysts, prior to writing their reports, to negotiate only a flat fee for their work that is not linked to their conclusions or recommendations.

Travel Funding

The benefits related to accepting paid travel extend beyond the cost savings to the member or candidate and his firm, such as the chance to talk exclusively with the executives of a company or learning more about the investment options provided by an investment organization. Acceptance also comes with potential concerns; for example, members and candidates may be influenced by these discussions when flying on a corporate or chartered jet or attending sponsored conferences where many expenses, including airfare and lodging, are covered. To avoid the appearance of compromising their independence and objectivity, best practice dictates that members and candidates always use commercial transportation at their expense or at the expense of their firm rather than accept paid travel arrangements from an outside company. Should commercial transportation be unavailable, members and candidates may accept modestly arranged travel to participate in appropriate information-gathering events, such as a property tour.

Recommended Procedures for Compliance

Members and candidates should adhere to the following practices and should encourage their firms to establish procedures to avoid violations of Standard I(B):

- *Protect the integrity of opinions*: Members, candidates, and their firms should establish policies stating that every research report concerning the securities of a corporate client should reflect the unbiased opinion of the analyst. Firms should also design compensation systems that protect the integrity of the investment decision process by maintaining the independence and objectivity of analysts.

- *Create a restricted list*: If the firm is unwilling to permit dissemination of adverse opinions about a corporate client, members and candidates should encourage the firm to remove the controversial company from the research universe and put it on a restricted list so that the firm disseminates only factual information about the company.
- *Restrict special cost arrangements*: When attending meetings at an issuer's headquarters, members and candidates should pay for commercial transportation and hotel charges. No corporate issuer should reimburse members or candidates for air transportation. Members and candidates should encourage issuers to limit the use of corporate aircraft to situations in which commercial transportation is not available or in which efficient movement could not otherwise be arranged. Members and candidates should take particular care that when frequent meetings are held between an individual issuer and an individual member or candidate, the issuer should not always host the member or candidate.
- *Limit gifts*: Members and candidates must limit the acceptance of gratuities and/or gifts to token items. Standard I(B) does not preclude customary, ordinary business-related entertainment as long as its purpose is not to influence or reward members or candidates. Firms should consider a strict value limit for acceptable gifts that is based on the local or regional customs and should address whether the limit is per gift or an aggregate annual value.
- *Restrict investments*: Members and candidates should encourage their investment firms to develop formal policies related to employee purchases of equity or equity-related IPOs. Firms should require prior approval for employee participation in IPOs, with prompt disclosure of investment actions taken following the offering. Strict limits should be imposed on investment personnel acquiring securities in private placements.
- *Review procedures*: Members and candidates should encourage their firms to implement effective supervisory and review procedures to ensure that analysts and portfolio managers comply with policies relating to their personal investment activities.
- *Independence policy*: Members, candidates, and their firms should establish a formal written policy on the independence and objectivity of research and implement reporting structures and review procedures to ensure that research analysts do not report to and are not supervised or controlled by any department of the firm that could compromise the independence of the analyst. More detailed recommendations related to a firm's policies regarding research objectivity are set forth in the CFA Institute statement *Research Objectivity Standards* (www.cfainstitute.org).
- *Appointed officer*: Firms should appoint a senior officer with oversight responsibilities for compliance with the firm's code of ethics and all regulations concerning its business. Firms should provide every employee with the procedures and policies for reporting potentially unethical behavior, violations of regulations, or other activities that may harm the firm's reputation.

Application of the Standard

Example 1 (Travel Expenses):

Steven Taylor, a mining analyst with Bronson Brokers, is invited by Precision Metals to join a group of his peers in a tour of mining facilities in several western US states. The company arranges for chartered group flights from site to site and for accommodations in Spartan Motels, the only chain with accommodations near the mines, for three nights. Taylor allows Precision Metals to pick up his tab, as do the other analysts, with one exception—John Adams, an employee of a large trust company who insists on following his company's policy and paying for his hotel room himself.

> *Comment*: The policy of the company where Adams works complies closely with Standard I(B) by avoiding even the appearance of a conflict of interest, but Taylor and the other analysts were not necessarily violating Standard I(B). In general, when allowing companies to pay for travel and/or accommodations in these circumstances, members and candidates must use their judgment. They must be on guard that such arrangements not impinge on a member's or candidate's independence and objectivity. In this example, the trip was strictly for business and Taylor was not accepting irrelevant or lavish hospitality. The itinerary required chartered flights, for which analysts were not expected to pay. The accommodations were modest. These arrangements are not unusual and did not violate Standard I(B) as long as Taylor's independence and objectivity were not compromised. In the final analysis, members and candidates should consider both whether they can remain objective and whether their integrity might be perceived by their clients to have been compromised.

Example 2 (Research Independence):

Susan Dillon, an analyst in the corporate finance department of an investment services firm, is making a presentation to a potential new business client that includes the promise that her firm will provide full research coverage of the potential client.

> *Comment*: Dillon may agree to provide research coverage, but she must not commit her firm's research department to providing a favorable recommendation. The firm's recommendation (favorable, neutral, or unfavorable) must be based on an independent and objective investigation and analysis of the company and its securities.

Example 3 (Research Independence and Intrafirm Pressure):

Walter Fritz is an equity analyst with Hilton Brokerage who covers the mining industry. He has concluded that the stock of Metals & Mining is overpriced at its current level, but he is concerned that a negative research report will hurt the good relationship between Metals & Mining and the investment banking division of his firm. In fact, a senior manager of Hilton Brokerage has just sent him a copy of a proposal his firm has made to Metals & Mining to underwrite a debt offering. Fritz needs to produce a report right away and is concerned about issuing a less-than-favorable rating.

> *Comment*: Fritz's analysis of Metals & Mining must be objective and based solely on consideration of company fundamentals. Any pressure from other divisions of his firm is inappropriate. This conflict could have been eliminated if, in anticipation of the offering, Hilton Brokerage had placed Metals & Mining on a restricted list for its sales force.

Example 4 (Research Independence and Issuer Relationship Pressure):

As in Example 3, Walter Fritz has concluded that Metals & Mining stock is overvalued at its current level, but he is concerned that a negative research report might jeopardize a close rapport that he has nurtured over the years with Metals & Mining's CEO, chief finance officer, and investment relations officer. Fritz is concerned that a negative report might result also in management retaliation—for instance, cutting him off from participating in conference calls when a quarterly earnings release is made, denying him the ability to ask questions on such calls, and/or denying him access to top management for arranging group meetings between Hilton Brokerage clients and top Metals & Mining managers.

> *Comment*: As in Example 3, Fritz's analysis must be objective and based solely on consideration of company fundamentals. Any pressure from Metals & Mining is inappropriate. Fritz should reinforce the integrity of his conclusions by stressing that his investment recommendation is based on relative valuation, which may include qualitative issues with respect to Metals & Mining's management.

Example 5 (Research Independence and Sales Pressure):

As support for the sales effort of her corporate bond department, Lindsey Warner offers credit guidance to purchasers of fixed-income securities. Her compensation is closely linked to the performance of the corporate bond department. Near the quarter's end, Warner's firm has a large inventory position in the bonds of Milton, Ltd., and has been unable to sell the bonds because of Milton's recent announcement of an operating problem. Salespeople have asked her to contact large clients to push the bonds.

> *Comment*: Unethical sales practices create significant potential violations of the Code and Standards. Warner's opinion of the Milton bonds must not be affected by internal pressure or compensation. In this case, Warner must refuse to push the Milton bonds unless she is able to justify that the market price has already adjusted for the operating problem.

Example 6 (Research Independence and Prior Coverage):

Jill Jorund is a securities analyst following airline stocks and a rising star at her firm. Her boss has been carrying a "buy" recommendation on International Airlines and asks Jorund to take over coverage of that airline. He tells Jorund that under no circumstances should the prevailing buy recommendation be changed.

> *Comment*: Jorund must be independent and objective in her analysis of International Airlines. If she believes that her boss's instructions have compromised her, she has two options: She can tell her boss that she cannot cover the company under these constraints, or she can take over coverage of the company, reach her own independent conclusions, and if they conflict with her boss's opinion, share the conclusions with her boss or other supervisors in the firm so that they can make appropriate recommendations. Jorund must issue only recommendations that reflect her independent and objective opinion.

Example 7 (Gifts and Entertainment from Related Party):

Edward Grant directs a large amount of his commission business to a New York–based brokerage house. In appreciation for all the business, the brokerage house gives Grant two tickets to the World Cup in South Africa, two nights at a nearby resort, several meals, and transportation via limousine to the game. Grant fails to disclose receiving this package to his supervisor.

> *Comment*: Grant has violated Standard I(B) because accepting these substantial gifts may impede his independence and objectivity. Every member and candidate should endeavor to avoid situations that might cause or be perceived to cause a loss of independence or objectivity in recommending

investments or taking investment action. By accepting the trip, Grant has opened himself up to the accusation that he may give the broker favored treatment in return.

Example 8 (Gifts and Entertainment from Client):

Theresa Green manages the portfolio of Ian Knowlden, a client of Tisbury Investments. Green achieves an annual return for Knowlden that is consistently better than that of the benchmark she and the client previously agreed to. As a reward, Knowlden offers Green two tickets to Wimbledon and the use of Knowlden's flat in London for a week. Green discloses this gift to her supervisor at Tisbury.

Comment: Green is in compliance with Standard I(B) because she disclosed the gift from one of her clients in accordance with the firm's policies. Members and candidates may accept bonuses or gifts from clients as long as they disclose them to their employer because gifts in a client relationship are deemed less likely to affect a member's or candidate's objectivity and independence than gifts in other situations. Disclosure is required, however, so that supervisors can monitor such situations to guard against employees favoring a gift-giving client to the detriment of other fee-paying clients (such as by allocating a greater proportion of IPO stock to the gift-giving client's portfolio).

Best practices for monitoring include comparing the transaction costs of the Knowlden account with the costs of other accounts managed by Green and other similar accounts within Tisbury. The supervisor could also compare the performance returns with the returns of other clients with the same mandate. This comparison will assist in determining whether a pattern of favoritism by Green is disadvantaging other Tisbury clients or the possibility that this favoritism could affect her future behavior.

Example 9 (Travel Expenses from External Manager):

Tom Wayne is the investment manager of the Franklin City Employees Pension Plan. He recently completed a successful search for a firm to manage the foreign equity allocation of the plan's diversified portfolio. He followed the plan's standard procedure of seeking presentations from a number of qualified firms and recommended that his board select Penguin Advisors because of its experience, well-defined investment strategy, and performance record. The firm claims compliance with the Global Investment Performance Standards (GIPS) and has been verified. Following the selection of Penguin, a reporter from the *Franklin City Record* calls to ask if there was any connection between this action and the fact that Penguin was one of the sponsors of an "investment fact-finding trip to Asia" that Wayne made earlier in the year. The trip was one of several conducted by the Pension Investment Academy, which had arranged the itinerary of meetings with economic, government, and corporate officials in major cities in several Asian countries. The Pension Investment Academy obtains support for the cost of these trips from a number of investment managers, including Penguin Advisors; the Academy then pays the travel expenses of the various pension plan managers on the trip and provides all meals and accommodations. The president of Penguin Advisors was also one of the travelers on the trip.

Comment: Although Wayne can probably put to good use the knowledge he gained from the trip in selecting portfolio managers and in other areas of managing the pension plan, his recommendation of Penguin Advisors may be tainted by the possible conflict incurred when he participated in a trip partly paid for by Penguin Advisors and when he was in the daily company of the president of Penguin Advisors. To avoid violating Standard I(B), Wayne's basic expenses for travel and accommodations should have been paid by his employer or the pension plan; contact with the president of Penguin Advisors should have been limited to informational or educational events only; and the trip, the organizer, and the sponsor should have been made a matter of public record. Even if his actions were not in violation of Standard I(B), Wayne should have been sensitive to the public perception of the trip when reported in the newspaper and the extent to which the subjective elements of his decision might have been affected by the familiarity that the daily contact of such a trip would encourage. This advantage would probably not be shared by firms competing with Penguin Advisors.

Example 10 (Research Independence and Compensation Arrangements):

Javier Herrero recently left his job as a research analyst for a large investment adviser. While looking for a new position, he was hired by an investor-relations firm to write a research report on one of its clients, a small educational software company. The investor-relations firm hopes to generate investor interest in the technology company. The firm will pay Herrero a flat fee plus a bonus if any new investors buy stock in the company as a result of Herrero's report.

Comment: If Herrero accepts this payment arrangement, he will be in violation of Standard I(B) because the compensation arrangement can reasonably be expected to compromise his independence and objectivity. Herrero will receive a bonus for attracting investors, which provides an incentive to draft a positive report regardless of the facts and to ignore or play down any negative information about the company. Herrero should accept only a flat fee that is not tied to the conclusions or recommendations of the report. Issuer-paid research that is objective and unbiased can be done under the right circumstances as long as the analyst takes steps to maintain his or her objectivity and includes in the report proper disclosures regarding potential conflicts of interest.

Example 11 (Recommendation Objectivity and Service Fees):

Two years ago, Bob Wade, trust manager for Central Midas Bank, was approached by Western Funds about promoting its family of funds, with special interest in the service-fee class of funds. To entice Central to promote this class, Western Funds offered to pay the bank a service fee of 0.25%. Without disclosing the fee being offered to the bank, Wade asked one of the investment managers to review Western's funds to determine whether they were suitable for clients of Central Midas Bank. The manager completed the normal due diligence review and determined that the new funds were fairly valued in the market with fee structures on a par with competitors. Wade decided to accept Western's offer and instructed the team of portfolio managers to exclusively promote these funds and the service-fee class to clients seeking to invest new funds or transfer from their current investments.

Now, two years later, the funds managed by Western begin to underperform their peers. Wade is counting on the fees to reach his profitability targets and continues to push these funds as acceptable investments for Central's clients.

Comment: Wade is violating Standard I(B) because the fee arrangement has affected the objectivity of his recommendations. Wade is relying on the fee as a component of the department's profitability and is unwilling to offer other products that may affect the fees received.

See also Standard VI(A)–Disclosure of Conflicts.

Example 12 (Recommendation Objectivity):

Bob Thompson has been doing research for the portfolio manager of the fixed-income department. His assignment is to do sensitivity analysis on securitized subprime mortgages. He has discussed with the manager possible scenarios to use to calculate expected returns. A key assumption in such calculations is housing price appreciation (HPA) because it drives "prepays" (prepayments of mortgages) and losses. Thompson is concerned with the significant appreciation experienced over the previous five years as a result of the increased availability of funds from subprime mortgages. Thompson insists that the analysis should include a scenario run with –10% for Year 1, –5% for Year 2, and then (to project a worst-case scenario) 0% for Years 3 through 5. The manager replies that these assumptions are too dire because there has never been a time in their available database when HPA was negative.

Thompson conducts his research to better understand the risks inherent in these securities and evaluates these securities in the worst-case scenario, an unlikely but possible environment. Based on the results of the enhanced scenarios, Thompson does not recommend the purchase of the securitization. Against the general market trends, the manager follows Thompson's recommendation and does not invest. The following year, the housing market collapses. In avoiding the subprime investments, the manager's portfolio outperforms its peer group that year.

> *Comment*: Thompson's actions in running the worst-case scenario against the protests of the portfolio manager are in alignment with the principles of Standard I(B). Thompson did not allow his research to be pressured by the general trends of the market or the manager's desire to limit the research to historical norms.
>
> See also Standard V(A)–Diligence and Reasonable Basis.

Example 13 (Influencing Manager Selection Decisions):

Adrian Mandel, CFA, is a senior portfolio manager for ZZYY Capital Management who oversees a team of investment professionals who manage labor union pension funds. A few years ago, ZZYY sought to win a competitive asset manager search to manage a significant allocation of the pension fund of the United Doughnut and Pretzel Bakers Union (UDPBU). UDPBU's investment board is chaired by a recognized key decision maker and long-time leader of the union, Ernesto Gomez. To improve ZZYY's chances of winning the competition, Mandel made significant monetary contributions to Gomez's union reelection campaign fund. Even after ZZYY was hired as a primary manager of the pension, Mandel believed that his firm's position was not secure. Mandel continued to contribute to Gomez's reelection campaign chest as well as to entertain lavishly the union leader and his family at top restaurants on a regular basis. All of Mandel's outlays were routinely handled as marketing expenses reimbursed by ZZYY's expense accounts and were disclosed to his senior management as being instrumental in maintaining a strong close relationship with an important client.

> *Comment*: Mandel not only offered but actually gave monetary gifts, benefits, and other considerations that reasonably could be expected to compromise Gomez's objectivity. Therefore, Mandel was in violation of Standard I(B).

Example 14 (Influencing Manager Selection Decisions):

Adrian Mandel, CFA, had heard about the manager search competition for the UDPBU Pension Fund through a broker/dealer contact. The contact told him that a well-known retired professional golfer, Bobby "The Bear" Finlay, who had become a licensed broker/dealer serving as a pension consultant, was orchestrating the UDPBU manager search. Finlay had gained celebrity status with several labor union pension fund boards by entertaining their respective board members and regaling them with colorful stories of fellow pro golfers' antics in clubhouses around the world. Mandel decided to improve ZZYY's chances of being invited to participate in the search competition by befriending Finlay to curry his favor. Knowing Finlay's love of entertainment, Mandel wined and dined Finlay at high-profile bistros where Finlay could glow in the fan recognition lavished on him by all the other patrons. Mandel's endeavors paid off handsomely when Finlay recommended to the UDPBU board that ZZYY be entered as one of three finalist asset management firms in its search.

> *Comment*: Similar to Example 13, Mandel lavished gifts, benefits, and other considerations in the form of expensive entertainment that could reasonably be expected to influence the consultant to recommend the hiring of his firm. Therefore, Mandel was in violation of Standard I(B).

Example 15 (Fund Manager Relationships):

Amie Scott is a performance analyst within her firm with responsibilities for analyzing the performance of external managers. While completing her quarterly analysis, Scott notices a change in one manager's reported composite construction. The change concealed the bad performance of a particularly large account by placing that account into a new residual composite. This change allowed the manager to remain at the top of the list of manager performance. Scott knows her firm has a large allocation to this manager, and the fund's manager is a close personal friend of the CEO. She needs to deliver her final report but is concerned with pointing out the composite change.

> *Comment*: Scott would be in violation of Standard I(B) if she did not disclose the change in her final report. The analysis of managers' performance should not be influenced by personal relationships or the size of the allocation to the outside managers. By not including the change, Scott would not be providing an independent analysis of the performance metrics for her firm.

Example 16 (Intrafirm Pressure):

Jill Stein is head of performance measurement for her firm. During the last quarter, many members of the organization's research department were removed because of the poor quality of their recommendations. The subpar research caused one larger account holder to experience significant underperformance, which resulted in the client withdrawing his money after the end of the quarter. The head of sales requests that Stein remove this account from the firm's performance composite because the performance decline can be attributed to the departed research team and not the client's adviser.

> *Comment*: Pressure from other internal departments can create situations that cause a member or candidate to violate the Code and Standards. Stein must maintain her independence and objectivity and refuse to exclude specific accounts from the firm's performance composites to which they belong. As long as the client invested under a strategy similar to that of the defined composite, it cannot be excluded because of the poor stock selections that led to the underperformance and asset withdrawal.

Standard I(C) Misrepresentation

Members and Candidates must not knowingly make any misrepresentations relating to investment analysis, recommendations, actions, or other professional activities.

Guidance

Highlights:

- *Impact on Investment Practice*
- *Performance Reporting*
- *Social Media*
- *Omissions*
- *Plagiarism*
- *Work Completed for Employer*

Trust is the foundation of the investment profession. Investors must be able to rely on the statements and information provided to them by those with whom the investors have trusted their financial well-being. Investment professionals who make false or misleading statements not only harm investors but also reduce the level of investor confidence in the investment profession and threaten the integrity of capital markets as a whole.

A misrepresentation is any untrue statement or omission of a fact or any statement that is otherwise false or misleading. A member or candidate must not knowingly omit or misrepresent information or give a false impression of a firm, organization, or security in the member's or candidate's oral representations, advertising (whether in the press or through brochures), electronic communications, or written materials (whether publicly disseminated or not). In this context, "knowingly" means that the member or candidate either knows or should have known that the misrepresentation was being made or that omitted information could alter the investment decision-making process.

Written materials include, but are not limited to, research reports, underwriting documents, company financial reports, market letters, newspaper columns, and books. Electronic communications include, but are not limited to, internet communications, webpages, mobile applications, and e-mails. Members and candidates who use webpages should regularly monitor materials posted on these sites to ensure that they contain current information. Members and candidates should also ensure that all reasonable precautions have been taken to protect the site's integrity and security and that the site does not misrepresent any information and does provide full disclosure.

Standard I(C) prohibits members and candidates from guaranteeing clients any specific return on volatile investments. Most investments contain some element of risk that makes their return inherently unpredictable. For such investments, guaranteeing either a particular rate of return or a guaranteed preservation of investment capital (e.g., "I can guarantee that you will earn 8% on equities this year" or "I can guarantee that you will not lose money on this investment") is misleading to investors. Standard I(C) does not prohibit members and candidates from providing clients with information on investment products that have guarantees built into the structure of the products themselves or for which an institution has agreed to cover any losses.

Impact on Investment Practice

Members and candidates must not misrepresent any aspect of their practice, including (but not limited to) their qualifications or credentials, the qualifications or services provided by their firm, their performance record and the record of their firm, and the characteristics of an investment. Any misrepresentation made by a member or candidate relating to the member's or candidate's professional activities is a breach of this standard.

Members and candidates should exercise care and diligence when incorporating third-party information. Misrepresentations resulting from the use of the credit ratings, research, testimonials, or marketing materials of outside parties become the responsibility of the investment professional when it affects that professional's business practices.

Investing through outside managers continues to expand as an acceptable method of investing in areas outside a firm's core competencies. Members and candidates must disclose their intended use of external managers and must not represent those managers' investment practices as their own. Although the level of involvement of outside managers may change over time, appropriate disclosures by members and candidates are important in avoiding misrepresentations, especially if the primary activity is to invest directly with a single external manager. Standard V(B)–Communication with Clients and Prospective Clients discusses in further detail communicating the firm's investment practices.

Performance Reporting

The performance benchmark selection process is another area where misrepresentations may occur. Members and candidates may misrepresent the success of their performance record through presenting benchmarks that are not comparable to their strategies. Further, clients can be misled if the benchmark's results are not reported on a basis comparable to that of the fund's or client's results. Best practice is selecting the most appropriate available benchmark from a universe of available options. The transparent presentation of appropriate performance benchmarks is an important aspect in providing clients with information that is useful in making investment decisions.

However, Standard I(C) does not require that a benchmark always be provided in order to comply. Some investment strategies may not lend themselves to displaying an appropriate benchmark because of the complexity or diversity of the investments included. Furthermore, some investment strategies may use reference indexes that do not reflect the opportunity set of the invested assets—for example, a hedge fund comparing its performance with a "cash plus" basis. When such a benchmark is used, members and candidates should make reasonable efforts to ensure that they disclose the reasons behind the use of this reference index to avoid misrepresentations of their performance. Members and candidates should discuss with clients on a continuous basis the appropriate benchmark to be used for performance evaluations and related fee calculations.

Reporting misrepresentations may also occur when valuations for illiquid or non-traded securities are available from more than one source. When different options are available, members and candidates may be tempted to switch providers to obtain higher security valuations. The process of shopping for values may misrepresent a security's worth, lead to misinformed decisions to sell or hold an investment, and result in overcharging clients advisory fees.

Members and candidates should take reasonable steps to provide accurate and reliable security pricing information to clients on a consistent basis. Changing pricing providers should not be based solely on the justification that the new provider reports a higher current value of a security. Consistency in the reported information will improve the perception of the valuation process for illiquid securities. Clients will likely have additional confidence that they were able to make an informed decision about continuing to hold these securities in their portfolios.

Social Media

The advancement of online discussion forums and communication platforms, commonly referred to as "social media," is placing additional responsibilities on members and candidates. When communicating through social media channels, members and candidates should provide only the same information they are allowed to distribute to clients and potential clients through other traditional forms of communication. The online or interactive aspects of social media do not remove the need to be open and honest about the information being distributed.

Along with understanding and following existing and newly developing rules and regulations regarding the allowed use of social media, members and candidates should also ensure that all communications in this format adhere to the requirements of the Code and Standards. The perceived anonymity granted through these platforms may entice individuals to misrepresent their qualifications or abilities or those of their employer. Actions undertaken through social media that knowingly misrepresent investment recommendations or professional activities are considered a violation of Standard I(C).

Omissions

The omission of a fact or outcome can be misleading, especially given the growing use of models and technical analysis processes. Many members and candidates rely on such models and processes to scan for new investment opportunities, to develop investment vehicles, and to produce investment recommendations and ratings. When inputs are knowingly omitted, the resulting outcomes may provide misleading information to those who rely on it for making investment decisions. Additionally, the outcomes from models shall not be presented as fact because they represent the expected results based on the inputs and analysis process incorporated.

Omissions in the performance measurement and attribution process can also misrepresent a manager's performance and skill. Members and candidates should encourage their firms to develop strict policies for composite development to prevent cherry picking—situations in which selected accounts are presented as representative of the firm's abilities. The omission of any accounts appropriate for the defined composite may misrepresent to clients the success of the manager's implementation of its strategy.

Plagiarism

Standard I(C) also prohibits plagiarism in the preparation of material for distribution to employers, associates, clients, prospects, or the general public. Plagiarism is defined as copying or using in substantially the same form materials prepared by others without acknowledging the source of the material or identifying the author and publisher of such material. Members and candidates must not copy (or represent as their own) original ideas or material without permission and must acknowledge and identify the source of ideas or material that is not their own.

The investment profession uses a myriad of financial, economic, and statistical data in the investment decision-making process. Through various publications and presentations, the investment professional is constantly exposed to the work of others and to the temptation to use that work without proper acknowledgment.

Misrepresentation through plagiarism in investment management can take various forms. The simplest and most flagrant example is to take a research report or study done by another firm or person, change the names, and release the material as one's own original analysis. This action is a clear violation of Standard I(C). Other practices include (1) using excerpts from articles or reports prepared by others either verbatim or with only slight changes in wording without acknowledgment, (2) citing specific quotations as attributable to "leading analysts" and "investment experts" without naming the specific references, (3) presenting statistical estimates of forecasts prepared by others and identifying the sources but without including the qualifying statements or caveats that may have been used, (4) using charts and graphs without stating their sources, and (5) copying proprietary computerized spreadsheets or algorithms without seeking the cooperation or authorization of their creators.

In the case of distributing third-party, outsourced research, members and candidates may use and distribute such reports as long as they do not represent themselves as the report's authors. Indeed, the member or candidate may add value for the client by sifting through research and repackaging it for clients. In such cases, clients should be fully informed that they are paying for the ability of the member or candidate to find the best research from a wide variety of sources. Members and candidates must not misrepresent their abilities, the extent of their expertise, or the extent of their work in a way that would mislead their clients or prospective clients. Members and candidates should disclose whether the research being presented to clients comes from another source—from either within or outside the member's or candidate's firm. This allows clients to understand who has the expertise behind the report or whether the work is being done by the analyst, other members of the firm, or an outside party.

Standard I(C) also applies to plagiarism in oral communications, such as through group meetings; visits with associates, clients, and customers; use of audio/video media (which is rapidly increasing); and telecommunications, including electronic data transfer and the outright copying of electronic media.

One of the most egregious practices in violation of this standard is the preparation of research reports based on multiple sources of information without acknowledging the sources. Examples of information from such sources include ideas, statistical compilations, and forecasts combined to give the appearance of original work. Although there is no monopoly on ideas, members and candidates must give credit where it is clearly due. Analysts should not use undocumented forecasts, earnings projections, asset values, and so on. Sources must be revealed to bring the responsibility directly back to the author of the report or the firm involved.

Work Completed for Employer

The preceding paragraphs address actions that would constitute a violation of Standard I(C). In some situations, however, members or candidates may use research conducted or models developed by others within the same firm without committing a violation. The most common example relates to the situation in which one (or more) of the original analysts is no longer with the firm. Research and models developed while employed by a firm are the property of the firm. The firm retains the right to continue using the work completed after a member or candidate has left the organization. The firm may issue future reports without providing attribution to the prior analysts. A member or candidate cannot, however, reissue a previously released report solely under his or her name.

Recommended Procedures for Compliance

Factual Presentations

Members and candidates can prevent unintentional misrepresentations of their qualifications or the services they or their firms provide if each member and candidate understands the limit of the firm's or individual's capabilities and the need to be accurate and complete in presentations. Firms can provide guidance for employees who make written or oral presentations to clients or potential clients by providing a written list of the firm's available services and a description of the firm's qualifications. This list should suggest ways of describing the firm's services, qualifications, and compensation that are both accurate and suitable for client or customer presentations. Firms can also help prevent misrepresentation by specifically designating which employees are authorized to speak on behalf of the firm. Regardless of whether the firm provides guidance, members and candidates should make certain that they understand the services the firm can perform and its qualifications.

Qualification Summary

In addition, to ensure accurate presentations to clients, each member and candidate should prepare a summary of his or her own qualifications and experience and a list of the services the member or candidate is capable of performing. Firms can assist member and candidate compliance by periodically reviewing employee correspondence and documents that contain representations of individual or firm qualifications.

Verify Outside Information

When providing information to clients from a third party, members and candidates share a responsibility for the accuracy of the marketing and distribution materials that pertain to the third party's capabilities, services, and products. Misrepresentation by third parties can damage the member's or candidate's reputation, the reputation of the firm, and the integrity of the capital markets. Members and candidates should encourage their employers to develop procedures for verifying information of third-party firms.

Maintain Webpages

Members and candidates who publish a webpage should regularly monitor materials posted on the site to ensure that the site contains current information. Members and candidates should also ensure that all reasonable precautions have been taken to protect the site's integrity, confidentiality, and security and that the site does not misrepresent any information and provides full disclosure.

Plagiarism Policy

To avoid plagiarism in preparing research reports or conclusions of analysis, members and candidates should take the following steps:

- *Maintain copies*: Keep copies of all research reports, articles containing research ideas, material with new statistical methodologies, and other materials that were relied on in preparing the research report.
- *Attribute quotations*: Attribute to their sources any direct quotations, including projections, tables, statistics, model/product ideas, and new methodologies prepared by persons other than recognized financial and statistical reporting services or similar sources.
- *Attribute summaries*: Attribute to their sources any paraphrases or summaries of material prepared by others. For example, to support his analysis of Brown Company's competitive position, the author of a research report on Brown might summarize another analyst's report on Brown's chief competitor, but the author of the Brown report must acknowledge in his own report the reliance on the other analyst's report.

Application of the Standard

Example 1 (Disclosure of Issuer-Paid Research):

Anthony McGuire is an issuer-paid analyst hired by publicly traded companies to electronically promote their stocks. McGuire creates a website that promotes his research efforts as a seemingly independent analyst. McGuire posts a profile and a strong buy recommendation for each company on the website indicating that the stock is expected to increase in value. He does not disclose the contractual relationships with the companies he covers on his website, in the research reports he issues, or in the statements he makes about the companies in internet chat rooms.

> *Comment*: McGuire has violated Standard I(C) because the website is misleading to potential investors. Even if the recommendations are valid and supported with thorough research, his omissions regarding the true relationship between himself and the companies he covers constitute a misrepresentation. McGuire has also violated Standard VI(A)–Disclosure of Conflicts by not disclosing the existence of an arrangement with the companies through which he receives compensation in exchange for his services.

Example 2 (Correction of Unintentional Errors):

Hijan Yao is responsible for the creation and distribution of the marketing materials for his firm, which claims compliance with the GIPS standards. Yao creates and distributes a presentation of performance by the firm's Asian equity composite that states the composite has ¥350 billion in assets. In fact, the composite has only ¥35 billion in assets, and the higher figure on the presentation is a result of a typographical error. Nevertheless, the erroneous material is distributed to a number of clients before Yao catches the mistake.

> *Comment*: Once the error is discovered, Yao must take steps to cease distribution of the incorrect material and correct the error by informing those who have received the erroneous information. Because Yao did not knowingly make the misrepresentation, however, he did not violate Standard I(C). Because his firm claims compliance with the GIPS standards, it must also comply with the GIPS Guidance Statement on Error Correction in relation to the error.

Example 3 (Noncorrection of Known Errors):

Syed Muhammad is the president of an investment management firm. The promotional material for the firm, created by the firm's marketing department, incorrectly claims that Muhammad has an advanced degree in finance from a prestigious business school in addition to the CFA designation. Although Muhammad attended the school for a short period of time, he did not receive a degree. Over the years, Muhammad and others in the firm have distributed this material to numerous prospective clients and consultants.

> *Comment*: Even though Muhammad may not have been directly responsible for the misrepresentation of his credentials in the firm's promotional material, he used this material numerous times over an extended period and should have known of the misrepresentation. Thus, Muhammad has violated Standard I(C).

Example 4 (Plagiarism):

Cindy Grant, a research analyst for a Canadian brokerage firm, has specialized in the Canadian mining industry for the past 10 years. She recently read an extensive research report on Jefferson Mining, Ltd., by Jeremy Barton, another analyst. Barton provided extensive statistics on the mineral reserves, production capacity, selling rates, and marketing factors affecting Jefferson's operations. He also noted that initial drilling results on a new ore body, which had not been made public, might show the existence of mineral zones that could increase the life of Jefferson's main mines, but Barton cited no specific data as to the initial drilling results. Grant called an officer of Jefferson, who gave her the initial drilling results over the telephone. The data indicated that the expected life of the main mines would be tripled. Grant added these statistics to Barton's report and circulated it within her firm as her own report.

> *Comment*: Grant plagiarized Barton's report by reproducing large parts of it in her own report without acknowledgment.

Example 5 (Misrepresentation of Information):

When Ricki Marks sells mortgage-backed derivatives called "interest-only strips" (IOs) to public pension plan clients, she describes them as "guaranteed by the US government." Purchasers of the IOs are entitled only to the interest stream generated by the mortgages, however, not the notional principal itself. One particular municipality's investment policies and local law require that securities purchased by its public pension plans be guaranteed by the US government. Although the underlying mortgages are guaranteed, neither the investor's investment nor the interest stream on the IOs is guaranteed. When interest rates decline, causing an increase in prepayment of mortgages, interest payments to the IOs' investors decline, and these investors lose a portion of their investment.

> *Comment*: Marks violated Standard I(C) by misrepresenting the terms and character of the investment.

Example 6 (Potential Information Misrepresentation):

Khalouck Abdrabbo manages the investments of several high-net-worth individuals in the United States who are approaching retirement. Abdrabbo advises these individuals that a portion of their investments be moved from equity to bank-sponsored certificates of deposit and money market accounts so that the principal will be "guaranteed" up to a certain amount. The interest is not guaranteed.

> *Comment*: Although there is risk that the institution offering the certificates of deposits and money market accounts could go bankrupt, in the United States, these accounts are insured by the US government through the Federal Deposit Insurance Corporation. Therefore, using the term "guaranteed" in this context is not inappropriate as long as the amount is within the government-insured limit. Abdrabbo should explain these facts to the clients.

Example 7 (Plagiarism):

Steve Swanson is a senior analyst in the investment research department of Ballard and Company. Apex Corporation has asked Ballard to assist in acquiring the majority ownership of stock in the Campbell Company, a financial consulting firm, and to prepare a report recommending that stockholders of Campbell agree to the acquisition. Another investment firm, Davis and Company, had already prepared a report for Apex analyzing both Apex and Campbell and recommending an exchange ratio. Apex has given the Davis report to Ballard officers, who have passed it on to Swanson. Swanson reviews the Davis report and other available material on Apex and Campbell. From his analysis, he concludes that the common stocks of Campbell and Apex represent good value at their current prices; he believes, however, that the Davis report does not consider all the factors a Campbell stockholder would need to know to make a decision. Swanson reports his conclusions to the partner in charge, who tells him to "use the Davis report, change a few words, sign your name, and get it out."

> *Comment*: If Swanson does as requested, he will violate Standard I(C). He could refer to those portions of the Davis report that he agrees with if he identifies Davis as the source; he could then add his own analysis and conclusions to the report before signing and distributing it.

Example 8 (Plagiarism):

Claude Browning, a quantitative analyst for Double Alpha, Inc., returns from a seminar in great excitement. At that seminar, Jack Jorrely, a well-known quantitative analyst at a national brokerage firm, discussed one of his new models in great detail, and Browning is intrigued by the new concepts. He proceeds to test the model, making some minor mechanical changes but retaining the concepts, until he produces some very positive results. Browning quickly announces to his supervisors at Double Alpha that he has discovered a new model and that clients and prospective clients should be informed of this positive finding as ongoing proof of Double Alpha's continuing innovation and ability to add value.

> *Comment*: Although Browning tested Jorrely's model on his own and even slightly modified it, he must still acknowledge the original source of the idea. Browning can certainly take credit for the final, practical results; he can also support his conclusions with his own test. The credit for the innovative thinking, however, must be awarded to Jorrely.

Example 9 (Plagiarism):

Fernando Zubia would like to include in his firm's marketing materials some "plain-language" descriptions of various concepts, such as the price-to-earnings (P/E) multiple and why standard deviation is used as a measure of risk. The descriptions come from other sources, but Zubia wishes to use them without reference to the original authors. Would this use of material be a violation of Standard I(C)?

> *Comment*: Copying verbatim any material without acknowledgement, including plain-language descriptions of the P/E multiple and standard deviation, violates Standard I(C). Even though these concepts are general, best practice would be for Zubia to describe them in his own words or cite the sources from which the descriptions are quoted. Members and candidates would be violating Standard I(C) if they either were responsible for creating marketing materials without attribution or knowingly use plagiarized materials.

Example 10 (Plagiarism):

Through a mainstream media outlet, Erika Schneider learns about a study that she would like to cite in her research. Should she cite both the mainstream intermediary source as well as the author of the study itself when using that information?

> *Comment*: In all instances, a member or candidate must cite the actual source of the information. Best practice for Schneider would be to obtain the information directly from the author and review it before citing it in

a report. In that case, Schneider would not need to report how she found out about the information. For example, suppose Schneider read in the *Financial Times* about a study issued by CFA Institute; best practice for Schneider would be to obtain a copy of the study from CFA Institute, review it, and then cite it in her report. If she does not use any interpretation of the report from the *Financial Times* and the newspaper does not add value to the report itself, the newspaper is merely a conduit of the original information and does not need to be cited. If she does not obtain the report and review the information, Schneider runs the risk of relying on second-hand information that may misstate facts. If, for example, the *Financial Times* erroneously reported some information from the original CFA Institute study and Schneider copied that erroneous information without acknowledging CFA Institute, she could be the object of complaints. Best practice would be either to obtain the complete study from its original author and cite only that author or to use the information provided by the intermediary and cite both sources.

Example 11 (Misrepresentation of Information):

Paul Ostrowski runs a two-person investment management firm. Ostrowski's firm subscribes to a service from a large investment research firm that provides research reports that can be repackaged by smaller firms for those firms' clients. Ostrowski's firm distributes these reports to clients as its own work.

Comment: Ostrowski can rely on third-party research that has a reasonable and adequate basis, but he cannot imply that he is the author of such research. If he does, Ostrowski is misrepresenting the extent of his work in a way that misleads the firm's clients or prospective clients.

Example 12 (Misrepresentation of Information):

Tom Stafford is part of a team within Appleton Investment Management responsible for managing a pool of assets for Open Air Bank, which distributes structured securities to offshore clients. He becomes aware that Open Air is promoting the structured securities as a much less risky investment than the investment management policy followed by him and the team to manage the original pool of assets. Also, Open Air has procured an independent rating for the pool that significantly overstates the quality of the investments. Stafford communicates his concerns to his supervisor, who responds that Open Air owns the product and is responsible for all marketing and distribution. Stafford's supervisor goes on to say that the product is outside of the US regulatory regime that Appleton follows and that all risks of the product are disclosed at the bottom of page 184 of the prospectus.

Comment: As a member of the investment team, Stafford is qualified to recognize the degree of accuracy of the materials that characterize the portfolio, and he is correct to be worried about Appleton's responsibility for a misrepresentation of the risks. Thus, he should continue to pursue the issue of Open Air's inaccurate promotion of the portfolio according to the firm's policies and procedures.

The Code and Standards stress protecting the reputation of the firm and the sustainability and integrity of the capital markets. Misrepresenting the quality and risks associated with the investment pool may lead to negative consequences for others well beyond the direct investors.

Example 13 (Avoiding a Misrepresentation):

Trina Smith is a fixed-income portfolio manager at a pension fund. She has observed that the market for highly structured mortgages is the focus of salespeople she meets and that these products represent a significant number of trading opportunities. In discussions about this topic with her team, Smith learns that calculating yields on changing cash flows within the deal structure requires very specialized vendor software. After more research, they find out that each deal is unique and that deals can have more than a dozen layers and changing cash flow priorities. Smith comes to the conclusion that, because of the complexity of these securities, the team cannot effectively distinguish between potentially good and bad investment options. To avoid misrepresenting their understanding, the team decides that the highly structured mortgage segment of the securitized market should not become part of the core of the fund's portfolio; they will allow some of the less complex securities to be part of the core.

> *Comment*: Smith is in compliance with Standard I(C) by not investing in securities that she and her team cannot effectively understand. Because she is not able to describe the risk and return profile of the securities to the pension fund beneficiaries and trustees, she appropriately limits the fund's exposure to this sector.

Example 14 (Misrepresenting Composite Construction):

Robert Palmer is head of performance for a fund manager. When asked to provide performance numbers to fund rating agencies, he avoids mentioning that the fund manager is quite liberal in composite construction. The reason accounts are included/excluded is not fully explained. The performance values reported to the rating agencies for the composites, although accurate for the accounts shown each period, may not present a true representation of the fund manager's ability.

> *Comment*: "Cherry picking" accounts to include in either published reports or information provided to rating agencies conflicts with Standard I(C). Moving accounts into or out of a composite to influence the overall performance results materially misrepresents the reported values over time. Palmer should work with his firm to strengthen its reporting practices concerning composite construction to avoid misrepresenting the firm's track record or the quality of the information being provided.

Example 15 (Presenting Out-of-Date Information):

David Finch is a sales director at a commercial bank, where he directs the bank's client advisers in the sale of third-party mutual funds. Each quarter, he holds a division-wide training session where he provides fact sheets on investment funds the bank is allowed to offer to clients. These fact sheets, which can be redistributed to potential clients, are created by the fund firms and contain information about the funds, including investment strategy and target distribution rates.

Finch knows that some of the fact sheets are out of date; for example, one long-only fund approved the use of significant leverage last quarter as a method to enhance returns. He continues to provide the sheets to the sales team without updates because the bank has no control over the marketing material released by the mutual fund firms.

> *Comment*: Finch is violating Standard I(C) by providing information that misrepresents aspects of the funds. By not providing the sales team and, ultimately, the clients with the updated information, he is misrepresenting the potential risks associated with the funds with outdated fact sheets. Finch

can instruct the sales team to clarify the deficiencies in the fact sheets with clients and ensure they have the most recent fund prospectus document before accepting orders for investing in any fund.

Example 16 (Overemphasis of Firm Results):

Bob Anderson is chief compliance officer for Optima Asset Management Company, a firm currently offering eight funds to clients. Seven of the eight had 10-year returns below the median for their respective sectors. Anderson approves a recent advertisement, which includes this statement: "Optima Asset Management is achieving excellent returns for its investors. The Optima Emerging Markets Equity fund, for example, has 10-year returns that exceed the sector median by more than 10%."

Comment: From the information provided it is difficult to determine whether a violation has occurred as long as the sector outperformance is correct. Anderson may be attempting to mislead potential clients by citing the performance of the sole fund that achieved such results. Past performance is often used to demonstrate a firm's skill and abilities in comparison to funds in the same sectors.

However, if all the funds outperformed their respective benchmarks, then Anderson's assertion that the company "is achieving excellent returns" may be factual. Funds may exhibit positive returns for investors, exceed benchmarks, and yet have returns below the median in their sectors.

Members and candidates need to ensure that their marketing efforts do not include statements that misrepresent their skills and abilities to remain compliant with Standard I(C). Unless the returns of a single fund reflect the performance of a firm as a whole, the use of a singular fund for performance comparisons should be avoided.

Standard I(D) Misconduct

Members and Candidates must not engage in any professional conduct involving dishonesty, fraud, or deceit or commit any act that reflects adversely on their professional reputation, integrity, or competence.

Guidance

Whereas Standard I(A) addresses the obligation of members and candidates to comply with applicable law that governs their professional activities, Standard I(D) addresses *all* conduct that reflects poorly on the professional integrity, good reputation, or competence of members and candidates. Any act that involves lying, cheating, stealing, or other dishonest conduct is a violation of this standard if the offense reflects adversely on a member's or candidate's professional activities. Although CFA Institute discourages any sort of unethical behavior by members and candidates, the Code and Standards are primarily aimed at conduct and actions related to a member's or candidate's professional life.

Conduct that damages trustworthiness or competence may include behavior that, although not illegal, nevertheless negatively affects a member's or candidate's ability to perform his or her responsibilities. For example, abusing alcohol during business hours might constitute a violation of this standard because it could have a detrimental effect on the member's or candidate's ability to fulfill his or her professional responsibilities. Personal bankruptcy may not reflect on the integrity or trustworthiness of the person declaring bankruptcy, but if the circumstances of the bankruptcy involve fraudulent or deceitful business conduct, the bankruptcy may be a violation of this standard.

In some cases, the absence of appropriate conduct or the lack of sufficient effort may be a violation of Standard I(D). The integrity of the investment profession is built on trust. A member or candidate—whether an investment banker, rating or research analyst, or portfolio manager—is expected to conduct the necessary due diligence to properly understand the nature and risks of an investment before making an investment recommendation. By not taking these steps and, instead, relying on someone else in the process to perform them, members or candidates may violate the trust their clients have placed in them. This loss of trust may have a significant impact on the reputation of the member or candidate and the operations of the financial market as a whole.

Individuals may attempt to abuse the CFA Institute Professional Conduct Program by actively seeking CFA Institute enforcement of the Code and Standards, and Standard I(D) in particular, as a method of settling personal, political, or other disputes unrelated to professional ethics. CFA Institute is aware of this issue, and appropriate disciplinary policies, procedures, and enforcement mechanisms are in place to address misuse of the Code and Standards and the Professional Conduct Program in this way.

Recommended Procedures for Compliance

In addition to ensuring that their own behavior is consistent with Standard I(D), to prevent general misconduct, members and candidates should encourage their firms to adopt the following policies and procedures to support the principles of Standard I(D):

- *Code of ethics*: Develop and/or adopt a code of ethics to which every employee must subscribe, and make clear that any personal behavior that reflects poorly on the individual involved, the institution as a whole, or the investment industry will not be tolerated.
- *List of violations*: Disseminate to all employees a list of potential violations and associated disciplinary sanctions, up to and including dismissal from the firm.
- *Employee references*: Check references of potential employees to ensure that they are of good character and not ineligible to work in the investment industry because of past infractions of the law.

Application of the Standard

Example 1 (Professionalism and Competence):

Simon Sasserman is a trust investment officer at a bank in a small affluent town. He enjoys lunching every day with friends at the country club, where his clients have observed him having numerous drinks. Back at work after lunch, he clearly is intoxicated while making investment decisions. His colleagues make a point of handling any business with Sasserman in the morning because they distrust his judgment after lunch.

> *Comment*: Sasserman's excessive drinking at lunch and subsequent intoxication at work constitute a violation of Standard I(D) because this conduct has raised questions about his professionalism and competence. His behavior reflects poorly on him, his employer, and the investment industry.

Example 2 (Fraud and Deceit):

Howard Hoffman, a security analyst at ATZ Brothers, Inc., a large brokerage house, submits reimbursement forms over a two-year period to ATZ's self-funded health insurance program for more than two dozen bills, most of which have been altered to increase the amount due. An investigation by the firm's director of employee benefits uncovers the inappropriate conduct. ATZ subsequently terminates Hoffman's employment and notifies CFA Institute.

> *Comment*: Hoffman violated Standard I(D) because he engaged in intentional conduct involving fraud and deceit in the workplace that adversely reflected on his integrity.

Example 3 (Fraud and Deceit):

Jody Brink, an analyst covering the automotive industry, volunteers much of her spare time to local charities. The board of one of the charitable institutions decides to buy five new vans to deliver hot lunches to low-income elderly people. Brink offers to donate her time to handle purchasing agreements. To pay a long-standing debt to a friend who operates an automobile dealership—and to compensate herself for her trouble—she agrees to a price 20% higher than normal and splits the surcharge with her friend. The director of the charity ultimately discovers the scheme and tells Brink that her services, donated or otherwise, are no longer required.

> *Comment*: Brink engaged in conduct involving dishonesty, fraud, and misrepresentation and has violated Standard I(D).

Example 4 (Personal Actions and Integrity):

Carmen Garcia manages a mutual fund dedicated to socially responsible investing. She is also an environmental activist. As the result of her participation in nonviolent protests, Garcia has been arrested on numerous occasions for trespassing on the property of a large petrochemical plant that is accused of damaging the environment.

> *Comment*: Generally, Standard I(D) is not meant to cover legal transgressions resulting from acts of civil disobedience in support of personal beliefs because such conduct does not reflect poorly on the member's or candidate's professional reputation, integrity, or competence.

Example 5 (Professional Misconduct):

Meredith Rasmussen works on a buy-side trading desk of an investment management firm and concentrates on in-house trades for a hedge fund subsidiary managed by a team at the investment management firm. The hedge fund has been very successful and is marketed globally by the firm. From her experience as the trader for much of the activity of the fund, Rasmussen has become quite knowledgeable about the hedge fund's strategy, tactics, and performance. When a distinct break in the market occurs and many of the securities involved in the hedge fund's strategy decline markedly in value, Rasmussen observes that the reported performance of the hedge fund does not reflect this decline. In her experience, the lack of effect is a very unlikely occurrence. She approaches the head of trading about her concern and is told that she should not ask any questions and that the fund is big and successful and is not her concern. She is fairly sure something is not right, so she contacts the compliance officer, who also tells her to stay away from the issue of the hedge fund's reporting.

> *Comment*: Rasmussen has clearly come across an error in policies, procedures, and compliance practices within the firm's operations. According to the firm's procedures for reporting potentially unethical activity, she

should pursue the issue by gathering some proof of her reason for doubt. Should all internal communications within the firm not satisfy her concerns, Rasmussen should consider reporting the potential unethical activity to the appropriate regulator.

See also Standard IV(A) for guidance on whistleblowing and Standard IV(C) for the duties of a supervisor.

STANDARD II: INTEGRITY OF CAPITAL MARKETS

Standard II(A) Material Nonpublic Information

Members and Candidates who possess material nonpublic information that could affect the value of an investment must not act or cause others to act on the information.

Guidance

Highlights:

- *What Is "Material" Information?*
- *What Constitutes "Nonpublic" Information?*
- *Mosaic Theory*
- *Social Media*
- *Using Industry Experts*
- *Investment Research Reports*

Trading or inducing others to trade on material nonpublic information erodes confidence in capital markets, institutions, and investment professionals by supporting the idea that those with inside information and special access can take unfair advantage of the general investing public. Although trading on inside information may lead to short-term profits, in the long run, individuals and the profession as a whole suffer from such trading. These actions have caused and will continue to cause investors to avoid capital markets because the markets are perceived to be "rigged" in favor of the knowledgeable insider. When the investing public avoids capital markets, the markets and capital allocation become less efficient and less supportive of strong and vibrant economies. Standard II(A) promotes and maintains a high level of confidence in market integrity, which is one of the foundations of the investment profession.

The prohibition on using this information goes beyond the direct buying and selling of individual securities or bonds. Members and candidates must not use material nonpublic information to influence their investment actions related to derivatives (e.g., swaps or option contracts), mutual funds, or other alternative investments. *Any* trading based on material nonpublic information constitutes a violation of Standard II(A). The expansion of financial products and the increasing interconnectivity of financial markets globally have resulted in new potential opportunities for trading on material nonpublic information.

What Is "Material" Information?

Information is "material" if its disclosure would probably have an impact on the price of a security or if reasonable investors would want to know the information before making an investment decision. In other words, information is material if it would significantly alter the total mix of information currently available about a security in such a way that the price of the security would be affected.

The specificity of the information, the extent of its difference from public information, its nature, and its reliability are key factors in determining whether a particular piece of information fits the definition of material. For example, material information may include, but is not limited to, information on the following:

- earnings;
- mergers, acquisitions, tender offers, or joint ventures;
- changes in assets or asset quality;
- innovative products, processes, or discoveries (e.g., new product trials or research efforts);
- new licenses, patents, registered trademarks, or regulatory approval/rejection of a product;
- developments regarding customers or suppliers (e.g., the acquisition or loss of a contract);
- changes in management;
- change in auditor notification or the fact that the issuer may no longer rely on an auditor's report or qualified opinion;
- events regarding the issuer's securities (e.g., defaults on senior securities, calls of securities for redemption, repurchase plans, stock splits, changes in dividends, changes to the rights of security holders, and public or private sales of additional securities);
- bankruptcies;
- significant legal disputes;
- government reports of economic trends (employment, housing starts, currency information, etc.);
- orders for large trades before they are executed; and
- new or changing equity or debt ratings issued by a third party (e.g., sell-side recommendations and credit ratings).

In addition to the substance and specificity of the information, the source or relative reliability of the information also determines materiality. The less reliable a source, the less likely the information provided would be considered material. For example, factual information from a corporate insider regarding a significant new contract for a company is likely to be material, whereas an assumption based on speculation by a competitor about the same contract is likely to be less reliable and, therefore, not material. Additionally, information about trials of a new drug, product, or service under development from qualified personnel involved in the trials is likely to be material, whereas educated conjecture by subject experts not connected to the trials is unlikely to be material.

Also, the more ambiguous the effect of the information on price, the less material that information is considered. If it is unclear whether and to what extent the information will affect the price of a security, the information may not be considered material. The passage of time may also render information that was once important immaterial.

What Constitutes "Nonpublic" Information?

Information is "nonpublic" until it has been disseminated or is available to the marketplace in general (as opposed to a select group of investors). "Disseminated" can be defined as "made known." For example, a company report of profits that is posted on the internet and distributed widely through a press release or accompanied by a filing has been effectively disseminated to the marketplace. Members and candidates must have a reasonable expectation that people have received the information before it can be considered public. It is not necessary, however, to wait for the slowest method of delivery. Once the information is disseminated to the market, it is public information that is no longer covered by this standard.

Members and candidates must be particularly aware of information that is selectively disclosed by corporations to a small group of investors, analysts, or other market participants. Information that is made available to analysts remains nonpublic until it is made available to investors in general. Corporations that disclose information on a limited basis create the potential for insider-trading violations.

Issues of selective disclosure often arise when a corporate insider provides material information to analysts in a briefing or conference call before that information is released to the public. Analysts must be aware that a disclosure made to a room full of analysts does not necessarily make the disclosed information "public." Analysts should also be alert to the possibility that they are selectively receiving material nonpublic information when a company provides them with guidance or interpretation of such publicly available information as financial statements or regulatory filings.

A member or candidate may use insider information provided legitimately by the source company for the specific purpose of conducting due diligence according to the business agreement between the parties for such activities as mergers, loan underwriting, credit ratings, and offering engagements. In such instances, the investment professional would not be considered in violation of Standard II(A) by using the material information. However, the use of insider information provided by the source company for other purposes, especially to trade or entice others to trade the securities of the firm, conflicts with this standard.

Mosaic Theory

A financial analyst gathers and interprets large quantities of information from many sources. The analyst may use significant conclusions derived from the analysis of public and nonmaterial nonpublic information as the basis for investment recommendations and decisions even if those conclusions would have been material inside information had they been communicated directly to the analyst by a company. Under the "mosaic theory," financial analysts are free to act on this collection, or mosaic, of information without risking violation.

The practice of financial analysis depends on the free flow of information. For the fair and efficient operation of the capital markets, analysts and investors must have the greatest amount of information possible to facilitate making well-informed investment decisions about how and where to invest capital. Accurate, timely, and intelligible communication is essential if analysts and investors are to obtain the data needed to make informed decisions about how and where to invest capital. These disclosures must go beyond the information mandated by the reporting requirements of the securities laws and should include specific business information about items used to guide a company's future growth, such as new products, capital projects, and the competitive environment. Analysts seek and use such information to compare and contrast investment alternatives.

Much of the information used by analysts comes directly from companies. Analysts often receive such information through contacts with corporate insiders, especially investor-relations staff and financial officers. Information may be disseminated in the form of press releases, through oral presentations by company executives in analysts' meetings or conference calls, or during analysts' visits to company premises. In seeking to develop the most accurate and complete picture of a company, analysts should also reach beyond contacts with companies themselves and collect information from other sources, such as customers, contractors, suppliers, and the companies' competitors.

Analysts are in the business of formulating opinions and insights that are not obvious to the general investing public about the attractiveness of particular securities. In the course of their work, analysts actively seek out corporate information not generally known to the market for the express purpose of analyzing that information, forming an opinion on its significance, and informing their clients, who can be expected to trade on the basis of the recommendation. Analysts' initiatives to discover and analyze information and communicate their findings to their clients significantly enhance market efficiency, thus benefiting all investors (see *Dirks v. Securities and Exchange Commission*). Accordingly, violations of Standard II(A) will *not* result when a perceptive analyst reaches a conclusion about a corporate action or event through an analysis of public information and items of nonmaterial nonpublic information.

Investment professionals should note, however, that although analysts are free to use mosaic information in their research reports, they should save and document all their research [see Standard V(C)–Record Retention]. Evidence of the analyst's knowledge of public and nonmaterial nonpublic information about a corporation strengthens the assertion that the analyst reached his or her conclusions solely through appropriate methods rather than through the use of material nonpublic information.

Social Media

The continuing advancement in technology allows members, candidates, and the industry at large to exchange information at rates not previously available. It is important for investment professionals to understand the implications of using information from the internet and social media platforms because all such information may not actually be considered public.

Some social media platforms require membership in specific groups in order to access the published content. Members and candidates participating in groups with membership limitations should verify that material information obtained from these sources can also be accessed from a source that would be considered available to the public (e.g., company filings, webpages, and press releases).

Members and candidates may use social media platforms to communicate with clients or investors without conflicting with this standard. As long as the information reaches all clients or is open to the investing public, the use of these platforms would be comparable with other traditional forms of communications, such as e-mails and press releases. Members and candidates, as required by Standard I(A), should also complete all appropriate regulatory filings related to information distributed through social media platforms.

Using Industry Experts

The increased demand for insights for understanding the complexities of some industries has led to an expansion of engagement with outside experts. As the level of engagement increased, new businesses formed to connect analysts and investors with individuals who have specialized knowledge of their industry (e.g., technology or pharmaceuticals). These networks offer investors the opportunity to reach beyond their usual business circles to speak with experts regarding economic conditions, industry trends, and technical issues relating to specific products and services.

Members and candidates may provide compensation to individuals for their insights without violating this standard. However, members and candidates are ultimately responsible for ensuring that they are not requesting or acting on confidential information received from external experts, which is in violation of security regulations and laws or duties to others. As the recent string of insider-trading cases displayed, some experts are willing to provide confidential and protected information for the right incentive.

Firms connecting experts with members or candidates often require both parties to sign agreements concerning the disclosure of material nonpublic information. Even with the protections from such compliance practices, if an expert provides material nonpublic information, members and candidates would be prohibited from taking investment actions on the associated firm until the information became publicly known to the market.

Investment Research Reports

When a particularly well-known or respected analyst issues a report or makes changes to his or her recommendation, that information alone may have an effect on the market and thus may be considered material. Theoretically, under Standard II(A), such a report would have to be made public at the time it was distributed to clients. The analyst is not a company insider, however, and does not have access to inside information. Presumably, the analyst created the report from information available to the public (mosaic theory) and by using his or her expertise to interpret the information. The analyst's hard work, paid for by the client, generated the conclusions.

Simply because the public in general would find the conclusions material does not require that the analyst make his or her work public. Investors who are not clients of the analyst can either do the work themselves or become clients of the analyst to gain access to the analyst's expertise.

Recommended Procedures for Compliance

Achieve Public Dissemination

If a member or candidate determines that information is material, the member or candidate should make reasonable efforts to achieve public dissemination of the information. These efforts usually entail encouraging the issuing company to make the information public. If public dissemination is not possible, the member or candidate must communicate the information only to the designated supervisory and compliance personnel within the member's or candidate's firm and must not take investment action or alter current investment recommendations on the basis of the information. Moreover, members and candidates must not knowingly engage in any conduct that may induce company insiders to privately disclose material nonpublic information.

Adopt Compliance Procedures

Members and candidates should encourage their firms to adopt compliance procedures to prevent the misuse of material nonpublic information. Particularly important is improving compliance in such areas as the review of employee and proprietary trading, the review of investment recommendations, documentation of firm procedures, and the supervision of interdepartmental communications in multiservice firms. Compliance procedures should suit the particular characteristics of a firm, including its size and the nature of its business.

Members and candidates are encouraged to inform their supervisor and compliance personnel of suspected inappropriate use of material nonpublic information as the basis for security trading activities or recommendations being made within their firm.

Adopt Disclosure Procedures

Members and candidates should encourage their firms to develop and follow disclosure policies designed to ensure that information is disseminated to the marketplace in an equitable manner. For example, analysts from small firms should receive the same information and attention from a company as analysts from large firms receive. Similarly, companies should not provide certain information to buy-side analysts but not to sell-side analysts, or vice versa. Furthermore, a company should not discriminate among analysts in the provision of information or "blackball" particular analysts who have given negative reports on the company in the past.

Within investment and research firms, members and candidates should encourage the development of and compliance with procedures for distributing new and updated investment opinions to clients. Recommendations of this nature may represent material market-moving information that needs to be communicated to all clients fairly.

Issue Press Releases

Companies should consider issuing press releases prior to analyst meetings and conference calls and scripting those meetings and calls to decrease the chance that further information will be disclosed. If material nonpublic information is disclosed for the first time in an analyst meeting or call, the company should promptly issue a press release or otherwise make the information publicly available.

Firewall Elements

An information barrier commonly referred to as a "firewall" is the most widely used approach for preventing the communication of material nonpublic information within firms. It restricts the flow of confidential information to those who need to know the information to perform their jobs effectively. The minimum elements of such a system include, but are not limited to, the following:

- substantial control of relevant interdepartmental communications, preferably through a clearance area within the firm in either the compliance or legal department;
- review of employee trading through the maintenance of "watch," "restricted," and "rumor" lists;
- documentation of the procedures designed to limit the flow of information between departments and of the actions taken to enforce those procedures; and
- heightened review or restriction of proprietary trading while a firm is in possession of material nonpublic information.

Appropriate Interdepartmental Communications

Although documentation requirements must, for practical reasons, take into account the differences between the activities of small firms and those of large, multiservice firms, firms of all sizes and types benefit by improving the documentation of their internal enforcement of firewall procedures. Therefore, even at small firms, procedures concerning interdepartmental communication, the review of trading activity, and the investigation of possible violations should be compiled and formalized.

Physical Separation of Departments

As a practical matter, to the greatest extent possible, firms should consider the physical separation of departments and files to prevent the communication of sensitive information that should not be shared. For example, the investment banking and corporate finance areas of a brokerage firm should be separated from the sales and research departments, and a bank's commercial lending department should be segregated from its trust and research departments.

Prevention of Personnel Overlap

There should be no overlap of personnel between the investment banking and corporate finance areas of a brokerage firm and the sales and research departments or between a bank's commercial lending department and its trust and research departments. For a firewall to be effective in a multiservice firm, an employee should be on only one side of the firewall at any time. Inside knowledge may not be limited to information about a specific offering or the current financial condition of a company. Analysts may be exposed to much information about the company, including new product developments or future budget projections that clearly constitute inside knowledge and thus preclude the analyst from returning to his or her research function. For example, an analyst who follows a particular company may provide limited assistance to the investment bankers under carefully controlled circumstances when the firm's investment banking department is involved in a deal with the company. That analyst must then be treated as though he or she were an investment banker; the analyst must remain on the investment banking side of the wall until any information he or she learns is publicly disclosed. In short, the analyst cannot use any information learned in the course of the project for research purposes and cannot share that information with colleagues in the research department.

A Reporting System

A primary objective of an effective firewall procedure is to establish a reporting system in which authorized people review and approve communications between departments. If an employee behind a firewall believes that he or she needs to share confidential information with someone on the other side of the wall, the employee should consult a designated compliance officer to determine whether sharing the information is necessary and how much information should be shared. If the sharing is necessary, the compliance officer should coordinate the process of "looking over the wall" so that the necessary information will be shared and the integrity of the procedure will be maintained.

A single supervisor or compliance officer should have the specific authority and responsibility of deciding whether information is material and whether it is sufficiently public to be used as the basis for investment decisions. Ideally, the supervisor or compliance officer responsible for communicating information to a firm's research or brokerage area would not be a member of that area.

Personal Trading Limitations

Firms should consider restrictions or prohibitions on personal trading by employees and should carefully monitor both proprietary trading and personal trading by employees. Firms should require employees to make periodic reports (to the extent that such reporting is not already required by securities laws) of their own transactions and transactions made for the benefit of family members. Securities should be placed on a restricted list when a firm has or may have material nonpublic information. The broad distribution of a restricted list often triggers the sort of trading the list was developed to avoid. Therefore, a watch list shown to only the few people responsible for compliance should be used to monitor transactions in specified securities. The use of a watch list in combination with a restricted list is an increasingly common means of ensuring effective control of personal trading.

Record Maintenance

Multiservice firms should maintain written records of the communications between various departments. Firms should place a high priority on training and should consider instituting comprehensive training programs, particularly for employees in sensitive areas.

Proprietary Trading Procedures

Procedures concerning the restriction or review of a firm's proprietary trading while the firm possesses material nonpublic information will necessarily depend on the types of proprietary trading in which the firm may engage. A prohibition on all types of proprietary activity when a firm comes into possession of material nonpublic information is *not* appropriate. For example, when a firm acts as a market maker, a prohibition on proprietary trading may be counterproductive to the goals of maintaining the confidentiality of information and market liquidity. This concern is particularly important in the relationships between small, regional broker/dealers and small issuers. In many situations, a firm will take a small issuer public with the understanding that the firm will continue to be a market maker in the stock. In such instances, a withdrawal by the firm from market-making activities would be a clear tip to outsiders. Firms that continue market-making activity while in the possession of material nonpublic information should, however, instruct their market makers to remain passive with respect to the market—that is, to take only the contra side of unsolicited customer trades.

In risk-arbitrage trading, the case for a trading prohibition is more compelling than it is in the case of market making. The impetus for arbitrage trading is neither passive nor reactive, and the potential for illegal profits is greater than in market making. The most prudent course for firms is to suspend arbitrage activity when a security is placed on the watch list. Those firms that continue arbitrage activity face a high hurdle in proving the adequacy of their internal procedures for preventing trading on material nonpublic information and must demonstrate a stringent review and documentation of firm trades.

Communication to All Employees

Members and candidates should encourage their employers to circulate written compliance policies and guidelines to all employees. Policies and guidelines should be used in conjunction with training programs aimed at enabling employees to recognize material nonpublic information. Such information is not always clearly identifiable.

Employees must be given sufficient training to either make an informed decision or to realize they need to consult a supervisor or compliance officer before engaging in questionable transactions. Appropriate policies reinforce that using material nonpublic information is illegal in many countries. Such trading activities based on material nonpublic information undermine the integrity of the individual, the firm, and the capital markets.

Application of the Standard

Example 1 (Acting on Nonpublic Information):

Frank Barnes, the president and controlling shareholder of the SmartTown clothing chain, decides to accept a tender offer and sell the family business at a price almost double the market price of its shares. He describes this decision to his sister (SmartTown's treasurer), who conveys it to her daughter (who owns no stock in the family company at present), who tells her husband, Staple. Staple, however, tells his stockbroker, Alex Halsey, who immediately buys SmartTown stock for himself.

> *Comment*: The information regarding the pending sale is both material and nonpublic. Staple has violated Standard II(A) by communicating the inside information to his broker. Halsey also has violated the standard by buying the shares on the basis of material nonpublic information.

Example 2 (Controlling Nonpublic Information):

Samuel Peter, an analyst with Scotland and Pierce Incorporated, is assisting his firm with a secondary offering for Bright Ideas Lamp Company. Peter participates, via telephone conference call, in a meeting with Scotland and Pierce investment banking employees and Bright Ideas' CEO. Peter is advised that the company's earnings projections for the next year have significantly dropped. Throughout the telephone conference call, several Scotland and Pierce salespeople and portfolio managers walk in and out of Peter's office, where the telephone call is taking place. As a result, they are aware of the drop in projected earnings for Bright Ideas. Before the conference call is concluded, the salespeople trade the stock of the company on behalf of the firm's clients and other firm personnel trade the stock in a firm proprietary account and in employees' personal accounts.

> *Comment*: Peter has violated Standard II(A) because he failed to prevent the transfer and misuse of material nonpublic information to others in his firm. Peter's firm should have adopted information barriers to prevent the communication of nonpublic information between departments of the firm. The salespeople and portfolio managers who traded on the information have also violated Standard II(A) by trading on inside information.

Example 3 (Selective Disclosure of Material Information):

Elizabeth Levenson is based in Hanoi and covers the Vietnamese market for her firm, which is based in Singapore. She is invited, together with the other 10 largest shareholders of a manufacturing company, to meet the finance director of that company. During the meeting, the finance director states that the company expects its workforce to strike next Friday, which will cripple productivity and distribution. Can Levenson use this information as a basis to change her rating on the company from "buy" to "sell"?

> *Comment*: Levenson must first determine whether the material information is public. According to Standard II(A), if the company has not made this information public (a small group forum does not qualify as a method of public dissemination), she cannot use the information.

Example 4 (Determining Materiality):

Leah Fechtman is trying to decide whether to hold or sell shares of an oil-and-gas exploration company that she owns in several of the funds she manages. Although the company has underperformed the index for some time already, the trends in the industry sector signal that companies of this type might become takeover targets. While she is considering her decision, her doctor, who casually follows the markets, mentions that she thinks that the company in question will soon be bought out by a large multinational conglomerate and that it would be a good idea to buy the stock right now. After talking to various investment professionals and checking their opinions on the company as well as checking industry trends, Fechtman decides the next day to accumulate more stock in the oil-and-gas exploration company.

> *Comment*: Although information on an expected takeover bid may be of the type that is generally material and nonpublic, in this case, the source of information is unreliable, so the information cannot be considered material. Therefore, Fechtman is not prohibited from trading the stock on the basis of this information.

Example 5 (Applying the Mosaic Theory):

Jagdish Teja is a buy-side analyst covering the furniture industry. Looking for an attractive company to recommend as a buy, he analyzes several furniture makers by studying their financial reports and visiting their operations. He also talks to some designers and retailers to find out which furniture styles are trendy and popular. Although none of the companies that he analyzes are a clear buy, he discovers that one of them, Swan Furniture Company (SFC), may be in financial trouble. SFC's extravagant new designs have been introduced at substantial cost. Even though these designs initially attracted attention, the public is now buying more conservative furniture from other makers. Based on this information and on a profit-and-loss analysis, Teja believes that SFC's next quarter earnings will drop substantially. He issues a sell recommendation for SFC. Immediately after receiving that recommendation, investment managers start reducing the SFC stock in their portfolios.

> *Comment*: Information on quarterly earnings data is material and nonpublic. Teja arrived at his conclusion about the earnings drop on the basis of public information and on pieces of nonmaterial nonpublic information (such as opinions of designers and retailers). Therefore, trading based on Teja's correct conclusion is not prohibited by Standard II(A).

Example 6 (Applying the Mosaic Theory):

Roger Clement is a senior financial analyst who specializes in the European automobile sector at Rivoli Capital. Because he has been repeatedly nominated by many leading industry magazines and newsletters as a "best analyst" for the automobile industry, he is widely regarded as an authority on the sector. After speaking with representatives of Turgot Chariots—a European auto manufacturer with sales primarily in South Korea—and after conducting interviews with salespeople, labor leaders, his firm's Korean currency analysts, and banking officials, Clement analyzed Turgot Chariots and concluded that (1) its newly introduced model will probably not meet sales expectations, (2) its corporate restructuring strategy may well face serious opposition from unions, (3) the depreciation of the Korean won should lead to pressure on margins for the industry in general and Turgot's market segment in particular, and (4) banks could take a tougher-than-expected stance in the upcoming round of credit renegotiations with the company. For these reasons, he changes his conclusion about the company from "market outperform" to "market underperform." Clement retains the support material used to reach his conclusion in case questions later arise.

> *Comment*: To reach a conclusion about the value of the company, Clement has pieced together a number of nonmaterial or public bits of information that affect Turgot Chariots. Therefore, under the mosaic theory, Clement has not violated Standard II(A) in drafting the report.

Example 7 (Analyst Recommendations as Material Nonpublic Information):

The next day, Clement is preparing to be interviewed on a global financial news television program where he will discuss his changed recommendation on Turgot Chariots for the first time in public. While preparing for the program, he mentions to the show's producers and Mary Zito, the journalist who will be interviewing him, the information he will be discussing. Just prior to going on the air, Zito sells her holdings in Turgot Chariots. She also phones her father with the information because she knows that he and other family members have investments in Turgot Chariots.

> *Comment*: When Zito receives advance notice of Clement's change of opinion, she knows it will have a material impact on the stock price, even if she is not totally aware of Clement's underlying reasoning. She is not a client

of Clement but obtains early access to the material nonpublic information prior to publication. Her trades are thus based on material nonpublic information and violate Standard II(A).

Zito further violates the Standard by relaying the information to her father. It would not matter if he or any other family member traded; the act of providing the information violates Standard II(A). The fact that the information is provided to a family member does not absolve someone of the prohibition of using or communicating material nonpublic information.

Example 8 (Acting on Nonpublic Information):

Ashton Kellogg is a retired investment professional who manages his own portfolio. He owns shares in National Savings, a large local bank. A close friend and golfing buddy, John Mayfield, is a senior executive at National. National has seen its stock price drop considerably, and the news and outlook are not good. In a conversation about the economy and the banking industry on the golf course, Mayfield relays the information that National will surprise the investment community in a few days when it announces excellent earnings for the quarter. Kellogg is pleasantly surprised by this information, and thinking that Mayfield, as a senior executive, knows the law and would not disclose inside information, he doubles his position in the bank. Subsequently, National announces that it had good operating earnings but had to set aside reserves for anticipated significant losses on its loan portfolio. The combined news causes the stock to go down 60%.

Comment: Even though Kellogg believes that Mayfield would not break the law by disclosing inside information and money was lost on the purchase, Kellogg should not have purchased additional shares of National. It is the member's or candidate's responsibility to make sure, before executing investment actions, that comments about earnings are not material nonpublic information. Kellogg has violated Standard II(A).

Example 9 (Mosaic Theory):

John Doll is a research analyst for a hedge fund that also sells its research to a select group of paying client investment firms. Doll's focus is medical technology companies and products, and he has been in the business long enough and has been successful enough to build up a very credible network of friends and experts in the business. Doll has been working on a major research report recommending Boyce Health, a medical device manufacturer. He recently ran into an old acquaintance at a wedding who is a senior executive at Boyce, and Doll asked about the business. Doll was drawn to a statement that the executive, who has responsibilities in the new products area, made about a product: "I would not get too excited about the medium-term prospects; we have a lot of work to do first." Doll incorporated this and other information about the new Boyce product in his long-term recommendation of Boyce.

Comment: Doll's conversation with the senior executive is part of the mosaic of information used in recommending Boyce. When holding discussions with a firm executive, Doll would need to guard against soliciting or obtaining material nonpublic information. Before issuing the report, the executive's statement about the continuing development of the product would need to be weighed against the other known public facts to determine whether it would be considered material.

Example 10 (Materiality Determination):

Larry Nadler, a trader for a mutual fund, gets a text message from another firm's trader, whom he has known for years. The message indicates a software company is going to report strong earnings when the firm publicly announces in two days. Nadler has a buy order from a portfolio manager within his firm to purchase several hundred thousand shares of the stock. Nadler is aggressive in placing the portfolio manager's order and completes the purchases by the following morning, a day ahead of the firm's planned earnings announcement.

> *Comment*: There are often rumors and whisper numbers before a release of any kind. The text message from the other trader would most likely be considered market noise. Unless Nadler knew that the trader had an ongoing business relationship with the public firm, he had no reason to suspect he was receiving material nonpublic information that would prevent him from completing the trading request of the portfolio manager.

Example 11 (Using an Expert Network):

Mary McCoy is the senior drug analyst at a mutual fund. Her firm hires a service that connects her to experts in the treatment of cancer. Through various phone conversations, McCoy enhances her understanding of the latest therapies for successful treatment. This information is critical to Mary making informed recommendations of the companies producing these drugs.

> *Comment*: McCoy is appropriately using the expert networks to enhance her evaluation process. She has neither asked for nor received information that may be considered material and nonpublic, such as preliminary trial results. McCoy is allowed to seek advice from professionals within the industry that she follows.

Example 12 (Using an Expert Network):

Tom Watson is a research analyst working for a hedge fund. To stay informed, Watson relies on outside experts for information on such industries as technology and pharmaceuticals, where new advancements occur frequently. The meetings with the industry experts often are arranged through networks or placement agents that have specific policies and procedures in place to deter the exchange of material nonpublic information.

Watson arranges a call to discuss future prospects for one of the fund's existing technology company holdings, a company that was testing a new semiconductor product. The scientist leading the tests indicates his disappointment with the performance of the new semiconductor. Following the call, Watson relays the insights he received to others at the fund. The fund sells its current position in the company and buys many put options because the market is anticipating the success of the new semiconductor and the share price reflects the market's optimism.

> *Comment*: Watson has violated Standard II(A) by passing along material nonpublic information concerning the ongoing product tests, which the fund used to trade in the securities and options of the related company. Watson cannot simply rely on the agreements signed by individuals who participate in expert networks that state that he has not received information that would prohibit his trading activity. He must make his own determination whether information he received through these arrangements reaches a materiality threshold that would affect his trading abilities.

Standard II(B) Market Manipulation

Members and Candidates must not engage in practices that distort prices or artificially inflate trading volume with the intent to mislead market participants.

Guidance

Highlights:

- *Information-Based Manipulation*
- *Transaction-Based Manipulation*

Standard II(B) requires that members and candidates uphold market integrity by prohibiting market manipulation. Market manipulation includes practices that distort security prices or trading volume with the intent to deceive people or entities that rely on information in the market. Market manipulation damages the interests of all investors by disrupting the smooth functioning of financial markets and lowering investor confidence.

Market manipulation may lead to a lack of trust in the fairness of the capital markets, resulting in higher risk premiums and reduced investor participation. A reduction in the efficiency of a local capital market may negatively affect the growth and economic health of the country and may also influence the operations of the globally interconnected capital markets. Although market manipulation may be less likely to occur in mature financial markets than in emerging markets, cross-border investing increasingly exposes all global investors to the potential for such practices.

Market manipulation includes (1) the dissemination of false or misleading information and (2) transactions that deceive or would be likely to mislead market participants by distorting the price-setting mechanism of financial instruments. The development of new products and technologies increases the incentives, means, and opportunities for market manipulation. Additionally, the increasing complexity and sophistication of the technologies used for communicating with market participants have created new avenues for manipulation.

Information-Based Manipulation

Information-based manipulation includes, but is not limited to, spreading false rumors to induce trading by others. For example, members and candidates must refrain from "pumping up" the price of an investment by issuing misleading positive information or overly optimistic projections of a security's worth only to later "dump" the investment (i.e., sell it) once the price, fueled by the misleading information's effect on other market participants, reaches an artificially high level.

Transaction-Based Manipulation

Transaction-based manipulation involves instances where a member or candidate knew or should have known that his or her actions could affect the pricing of a security. This type of manipulation includes, but is not limited to, the following:

- transactions that artificially affect prices or volume to give the impression of activity or price movement in a financial instrument, which represent a diversion from the expectations of a fair and efficient market, and
- securing a controlling, dominant position in a financial instrument to exploit and manipulate the price of a related derivative and/or the underlying asset.

Standard II(B) is not intended to preclude transactions undertaken on legitimate trading strategies based on perceived market inefficiencies. The intent of the action is critical to determining whether it is a violation of this standard.

Application of the Standard

Example 1 (Independent Analysis and Company Promotion):

The principal owner of Financial Information Services (FIS) entered into an agreement with two microcap companies to promote the companies' stock in exchange for stock and cash compensation. The principal owner caused FIS to disseminate e-mails, design and maintain several websites, and distribute an online investment newsletter—all of which recommended investment in the two companies. The systematic publication of purportedly independent analyses and recommendations containing inaccurate and highly promotional and speculative statements increased public investment in the companies and led to dramatically higher stock prices.

> *Comment*: The principal owner of FIS violated Standard II(B) by using inaccurate reporting and misleading information under the guise of independent analysis to artificially increase the stock price of the companies. Furthermore, the principal owner violated Standard V(A)–Diligence and Reasonable Basis by not having a reasonable and adequate basis for recommending the two companies and violated Standard VI(A)–Disclosure of Conflicts by not disclosing to investors the compensation agreements (which constituted a conflict of interest).

Example 2 (Personal Trading Practices and Price):

John Gray is a private investor in Belgium who bought a large position several years ago in Fame Pharmaceuticals, a German small-cap security with limited average trading volume. He has now decided to significantly reduce his holdings owing to the poor price performance. Gray is worried that the low trading volume for the stock may cause the price to decline further as he attempts to sell his large position.

Gray devises a plan to divide his holdings into multiple accounts in different brokerage firms and private banks in the names of family members, friends, and even a private religious institution. He then creates a rumor campaign on various blogs and social media outlets promoting the company.

Gray begins to buy and sell the stock using the accounts in hopes of raising the trading volume and the price. He conducts the trades through multiple brokers, selling slightly larger positions than he bought on a tactical schedule, and over time, he is able to reduce his holding as desired without negatively affecting the sale price.

> *Comment*: John violated Standard II(B) by fraudulently creating the appearance that there was a greater investor interest in the stock through the online rumors. Additionally, through his trading strategy, he created the

appearance that there was greater liquidity in the stock than actually existed. He was able to manipulate the price through both misinformation and trading practices.

Example 3 (Creating Artificial Price Volatility):

Matthew Murphy is an analyst at Divisadero Securities & Co., which has a significant number of hedge funds among its most important brokerage clients. Some of the hedge funds hold short positions on Wirewolf Semiconductor. Two trading days before the publication of a quarter-end report, Murphy alerts his sales force that he is about to issue a research report on Wirewolf that will include the following opinions:

- quarterly revenues are likely to fall short of management's guidance,
- earnings will be as much as 5 cents per share (or more than 10%) below consensus, and
- Wirewolf's highly respected chief financial officer may be about to join another company.

Knowing that Wirewolf has already entered its declared quarter-end "quiet period" before reporting earnings (and thus would be reluctant to respond to rumors), Murphy times the release of his research report specifically to sensationalize the negative aspects of the message in order to create significant downward pressure on Wirewolf's stock—to the distinct advantage of Divisadero's hedge fund clients. The report's conclusions are based on speculation, not on fact. The next day, the research report is broadcast to all of Divisadero's clients and to the usual newswire services.

Before Wirewolf's investor-relations department can assess the damage on the final trading day of the quarter and refute Murphy's report, its stock opens trading sharply lower, allowing Divisadero's clients to cover their short positions at substantial gains.

> *Comment*: Murphy violated Standard II(B) by aiming to create artificial price volatility designed to have a material impact on the price of an issuer's stock. Moreover, by lacking an adequate basis for the recommendation, Murphy also violated Standard V(A)–Diligence and Reasonable Basis.

Example 4 (Personal Trading and Volume):

Rajesh Sekar manages two funds—an equity fund and a balanced fund—whose equity components are supposed to be managed in accordance with the same model. According to that model, the funds' holdings in stock of Digital Design Inc. (DD) are excessive. Reduction of the DD holdings would not be easy, however, because the stock has low liquidity in the stock market. Sekar decides to start trading larger portions of DD stock back and forth between his two funds to slowly increase the price; he believes market participants will see growing volume and increasing price and become interested in the stock. If other investors are willing to buy the DD stock because of such interest, then Sekar will be able to get rid of at least some of his overweight position without inducing price decreases. In this way, the whole transaction will be for the benefit of fund participants, even if additional brokers' commissions are incurred.

> *Comment*: Sekar's plan would be beneficial for his funds' participants but is based on artificial distortion of both trading volume and the price of the DD stock and thus constitutes a violation of Standard II(B).

Example 5 ("Pump-Priming" Strategy):

ACME Futures Exchange is launching a new bond futures contract. To convince investors, traders, arbitrageurs, hedgers, and so on, to use its contract, the exchange attempts to demonstrate that it has the best liquidity. To do so, it enters into agreements with members in which they commit to a substantial minimum trading volume on the new contract over a specific period in exchange for substantial reductions of their regular commissions.

> *Comment*: The formal liquidity of a market is determined by the obligations set on market makers, but the actual liquidity of a market is better estimated by the actual trading volume and bid–ask spreads. Attempts to mislead participants about the actual liquidity of the market constitute a violation of Standard II(B). In this example, investors have been intentionally misled to believe they chose the most liquid instrument for some specific purpose, but they could eventually see the actual liquidity of the contract significantly reduced after the term of the agreement expires. If the ACME Futures Exchange fully discloses its agreement with members to boost transactions over some initial launch period, it will not violate Standard II(B). ACME's intent is not to harm investors but, on the contrary, to give them a better service. For that purpose, it may engage in a liquidity-pumping strategy, but the strategy must be disclosed.

Example 6 (Creating Artificial Price Volatility):

Emily Gordon, an analyst of household products companies, is employed by a research boutique, Picador & Co. Based on information that she has gathered during a trip through Latin America, she believes that Hygene, Inc., a major marketer of personal care products, has generated better-than-expected sales from its new product initiatives in South America. After modestly boosting her projections for revenue and for gross profit margin in her worksheet models for Hygene, Gordon estimates that her earnings projection of US$2.00 per diluted share for the current year may be as much as 5% too low. She contacts the chief financial officer (CFO) of Hygene to try to gain confirmation of her findings from her trip and to get some feedback regarding her revised models. The CFO declines to comment and reiterates management's most recent guidance of US$1.95–US$2.05 for the year.

Gordon decides to try to force a comment from the company by telling Picador & Co. clients who follow a momentum investment style that consensus earnings projections for Hygene are much too low; she explains that she is considering raising her published estimate by an ambitious US$0.15 to US$2.15 per share. She believes that when word of an unrealistically high earnings projection filters back to Hygene's investor-relations department, the company will feel compelled to update its earnings guidance. Meanwhile, Gordon hopes that she is at least correct with respect to the earnings direction and that she will help clients who act on her insights to profit from a quick gain by trading on her advice.

> *Comment*: By exaggerating her earnings projections in order to try to fuel a quick gain in Hygene's stock price, Gordon is in violation of Standard II(B). Furthermore, by virtue of previewing her intentions of revising upward her earnings projections to only a select group of clients, she is in violation of Standard III(B)–Fair Dealing. However, it would have been acceptable for Gordon to write a report that
>
> - framed her earnings projection in a range of possible outcomes,

- outlined clearly the assumptions used in her Hygene models that took into consideration the findings from her trip through Latin America, and
- was distributed to all Picador & Co. clients in an equitable manner.

Example 7 (Pump and Dump Strategy):

In an effort to pump up the price of his holdings in Moosehead & Belfast Railroad Company, Steve Weinberg logs on to several investor chat rooms on the internet to start rumors that the company is about to expand its rail network in anticipation of receiving a large contract for shipping lumber.

> *Comment*: Weinberg has violated Standard II(B) by disseminating false information about Moosehead & Belfast with the intent to mislead market participants.

Example 8 (Manipulating Model Inputs):

Bill Mandeville supervises a structured financing team for Superior Investment Bank. His responsibilities include packaging new structured investment products and managing Superior's relationship with relevant rating agencies. To achieve the best rating possible, Mandeville uses mostly positive scenarios as model inputs—scenarios that reflect minimal downside risk in the assets underlying the structured products. The resulting output statistics in the rating request and underwriting prospectus support the idea that the new structured products have minimal potential downside risk. Additionally, Mandeville's compensation from Superior is partially based on both the level of the rating assigned and the successful sale of new structured investment products but does not have a link to the long-term performance of the instruments.

Mandeville is extremely successful and leads Superior as the top originator of structured investment products for the next two years. In the third year, the economy experiences difficulties and the values of the assets underlying structured products significantly decline. The subsequent defaults lead to major turmoil in the capital markets, the demise of Superior Investment Bank, and the loss of Mandeville's employment.

> *Comment*: Mandeville manipulates the inputs of a model to minimize associated risk to achieve higher ratings. His understanding of structured products allows him to skillfully decide which inputs to include in support of the desired rating and price. This information manipulation for short-term gain, which is in violation of Standard II(B), ultimately causes significant damage to many parties and the capital markets as a whole. Mandeville should have realized that promoting a rating and price with inaccurate information could cause not only a loss of price confidence in the particular structured product but also a loss of investor trust in the system. Such loss of confidence affects the ability of the capital markets to operate efficiently.

Example 9 (Information Manipulation):

Allen King is a performance analyst for Torrey Investment Funds. King believes that the portfolio manager for the firm's small- and microcap equity fund dislikes him because the manager never offers him tickets to the local baseball team's games but does offer tickets to other employees. To incite a potential regulatory review of the manager, King creates user profiles on several online forums under the portfolio manager's name and starts rumors about potential mergers for several of the smaller companies in the portfolio. As the prices of these companies' stocks increase, the portfolio manager sells the position, which leads to an investigation by the regulator as King desired.

> *Comment*: King has violated Standard II(B) even though he did not personally profit from the market's reaction to the rumor. In posting the false information, King misleads others into believing the companies were likely to be acquired. Although his intent was to create trouble for the portfolio manager, his actions clearly manipulated the factual information that was available to the market.

STANDARD III: DUTIES TO CLIENTS

Standard III(A) Loyalty, Prudence, and Care

> Members and Candidates have a duty of loyalty to their clients and must act with reasonable care and exercise prudent judgment. Members and Candidates must act for the benefit of their clients and place their clients' interests before their employer's or their own interests.

Guidance

Highlights:

- *Understanding the Application of Loyalty, Prudence, and Care*
- *Identifying the Actual Investment Client*
- *Developing the Client's Portfolio*
- *Soft Commission Policies*
- *Proxy Voting Policies*

Standard III(A) clarifies that client interests are paramount. A member's or candidate's responsibility to a client includes a duty of loyalty and a duty to exercise reasonable care. Investment actions must be carried out for the sole benefit of the client and in a manner the member or candidate believes, given the known facts and circumstances, to be in the best interest of the client. Members and candidates must exercise the same level of prudence, judgment, and care that they would apply in the management and disposition of their own interests in similar circumstances.

Prudence requires caution and discretion. The exercise of prudence by investment professionals requires that they act with the care, skill, and diligence that a reasonable person acting in a like capacity and familiar with such matters would use. In the context of managing a client's portfolio, prudence requires following the investment parameters set forth by the client and balancing risk and return. Acting with care requires members and candidates to act in a prudent and judicious manner in avoiding harm to clients.

Standard III(A) sets minimum expectations for members and candidates when fulfilling their responsibilities to their clients. Regulatory and legal requirements for such duties can vary across the investment industry depending on a variety of factors, including job function of the investment professional, the existence of an adviser/client relationship, and the nature of the recommendations being offered. From the perspective of the end user of financial services, these different standards can be arcane and confusing, leaving investors unsure of what level of service to expect from investment professionals they employ. The single standard of conduct described in Standard III(A) benefits investors by establishing a benchmark for the duties of loyalty, prudence, and care and clarifies that all CFA Institute members and candidates, regardless of job title, local laws, or cultural differences, are required to comply with these fundamental responsibilities. Investors hiring members or candidates who must adhere to the duty of loyalty, prudence, and care set forth in this standard can be confident that these responsibilities are a requirement regardless of any legally imposed fiduciary duties.

Standard III(A), however, is not a substitute for a member's or candidate's legal or regulatory obligations. As stated in Standard I(A), members and candidates must abide by the most strict requirements imposed on them by regulators or the Code and Standards, including any legally imposed fiduciary duty. Members and candidates must also be aware of whether they have "custody" or effective control of client assets. If so, a heightened level of responsibility arises. Members and candidates are considered to have custody if they have any direct or indirect access to client funds. Members and candidates must manage any pool of assets in their control in accordance with the terms of the governing documents (such as trust documents and investment management agreements), which are the primary determinant of the manager's powers and duties. Whenever their actions are contrary to provisions of those instruments or applicable law, members and candidates are at risk of violating Standard III(A).

Understanding the Application of Loyalty, Prudence, and Care

Standard III(A) establishes a minimum benchmark for the duties of loyalty, prudence, and care that are required of all members and candidates regardless of whether a legal fiduciary duty applies. Although fiduciary duty often encompasses the principles of loyalty, prudence, and care, Standard III(A) does not render all members and candidates fiduciaries. The responsibilities of members and candidates for fulfilling their obligations under this standard depend greatly on the nature of their professional responsibilities and the relationships they have with clients. The conduct of members and candidates may or may not rise to the level of being a fiduciary, depending on the type of client, whether the member or candidate is giving investment advice, and the many facts and circumstances surrounding a particular transaction or client relationship.

Fiduciary duties are often imposed by law or regulation when an individual or institution is charged with the duty of acting for the benefit of another party, such as managing investment assets. The duty required in fiduciary relationships exceeds what is acceptable in many other business relationships because a fiduciary is in an enhanced position of trust. Although members and candidates must comply with any legally imposed fiduciary duty, the Code and Standards neither impose such a legal responsibility nor require all members or candidates to act as fiduciaries. However, Standard III(A) requires members and candidates to work in the client's best interest no matter what the job function.

A member or candidate who does not provide advisory services to a client but who acts only as a trade execution professional must prudently work in the client's interest when completing requested trades. Acting in the client's best interest requires these professionals to use their skills and diligence to execute trades in the most favorable terms that can be achieved. Members and candidates operating in such positions must use care to operate within the parameters set by the client's trading instructions.

Members and candidates may also operate in a blended environment where they execute client trades and offer advice on a limited set of investment options. The extent of the advisory arrangement and limitations should be outlined in the agreement with the client at the outset of the relationship. For instance, members and candidates should inform clients that the advice provided will be limited to the propriety products of the firm and not include other products available on the market. Clients who want access to a wider range of investment products would have the information necessary to decide not to engage with members or candidates working under these restrictions.

Members and candidates operating in this blended context would comply with their obligations by recommending the allowable products that are consistent with the client's objectives and risk tolerance. They would exercise care through diligently aligning the client's needs with the attributes of the products being recommended. Members and candidates should place the client's interests first by disregarding any firm or personal interest in motivating a recommended transaction.

There is a large variety of professional relationships that members and candidates have with their clients. Standard III(A) requires them to fulfill the obligations outlined explicitly or implicitly in the client agreements to the best of their abilities and with loyalty, prudence, and care. Whether a member or candidate is structuring a new securitization transaction, completing a credit rating analysis, or leading a public company, he or she must work with prudence and care in delivering the agreed-on services.

Identifying the Actual Investment Client

The first step for members and candidates in fulfilling their duty of loyalty to clients is to determine the identity of the "client" to whom the duty of loyalty is owed. In the context of an investment manager managing the personal assets of an individual, the client is easily identified. When the manager is responsible for the portfolios of pension plans or trusts, however, the client is not the person or entity who hires the manager but, rather, the beneficiaries of the plan or trust. The duty of loyalty is owed to the ultimate beneficiaries.

In some situations, an actual client or group of beneficiaries may not exist. Members and candidates managing a fund to an index or an expected mandate owe the duty of loyalty, prudence, and care to invest in a manner consistent with the stated mandate. The decisions of a fund's manager, although benefiting all fund investors, do not have to be based on an individual investor's requirements and risk profile. Client loyalty and care for those investing in the fund are the responsibility of members and candidates who have an advisory relationship with those individuals.

Situations involving potential conflicts of interest with respect to responsibilities to clients may be extremely complex because they may involve a number of competing interests. The duty of loyalty, prudence, and care applies to a large number of persons in varying capacities, but the exact duties may differ in many respects in accord with the relationship with each client or each type of account in which the assets are managed. Members and candidates must not only put their obligations to clients first in all dealings but also endeavor to avoid all real or potential conflicts of interest.

Members and candidates with positions whose responsibilities do not include direct investment management also have "clients" that must be considered. Just as there are various types of advisory relationships, members and candidates must look at their roles and responsibilities when making a determination of who their clients are. Sometimes the client is easily identifiable; such is the case in the relationship between a company executive and the firm's public shareholders. At other times, the client may be the investing public as a whole, in which case the goals of independence and objectivity of research surpass the goal of loyalty to a single organization.

Developing the Client's Portfolio

The duty of loyalty, prudence, and care owed to the individual client is especially important because the professional investment manager typically possesses greater knowledge in the investment arena than the client does. This disparity places the individual client in a vulnerable position; the client must trust the manager. The manager in these situations should ensure that the client's objectives and expectations for the performance of the account are realistic and suitable to the client's circumstances and that the risks involved are appropriate. In most circumstances, recommended investment strategies should relate to the long-term objectives and circumstances of the client.

Particular care must be taken to detect whether the goals of the investment manager or the firm in conducting business, selling products, and executing security transactions potentially conflict with the best interests and objectives of the client. When members and candidates cannot avoid potential conflicts between their firm and clients' interests, they must provide clear and factual disclosures of the circumstances to the clients.

Members and candidates must follow any guidelines set by their clients for the management of their assets. Some clients, such as charitable organizations and pension plans, have strict investment policies that limit investment options to certain types or classes of investment or prohibit investment in certain securities. Other organizations have aggressive policies that do not prohibit investments by type but, instead, set criteria on the basis of the portfolio's total risk and return.

Investment decisions must be judged in the context of the total portfolio rather than by individual investment within the portfolio. The member's or candidate's duty is satisfied with respect to a particular investment if the individual has thoroughly considered the investment's place in the overall portfolio, the risk of loss and opportunity for gains, tax implications, and the diversification, liquidity, cash flow, and overall return requirements of the assets or the portion of the assets for which the manager is responsible.

Soft Commission Policies

An investment manager often has discretion over the selection of brokers executing transactions. Conflicts may arise when an investment manager uses client brokerage to purchase research services, a practice commonly called "soft dollars" or "soft commissions." A member or candidate who pays a higher brokerage commission than he or she would normally pay to allow for the purchase of goods or services, without corresponding benefit to the client, violates the duty of loyalty to the client.

From time to time, a client will direct a manager to use the client's brokerage to purchase goods or services for the client, a practice that is commonly called "directed brokerage." Because brokerage commission is an asset of the client and is used to benefit that client, not the manager, such a practice does not violate any duty of loyalty. However, a member or candidate is obligated to seek "best price" and "best execution" and be assured by the client that the goods or services purchased from the brokerage will benefit the account beneficiaries. "Best execution" refers to a trading process that seeks to maximize the value of the client's portfolio within the client's stated investment objectives and constraints. In addition, the member or candidate should disclose to the client that the client may not be getting best execution from the directed brokerage.

Proxy Voting Policies

The duty of loyalty, prudence, and care may apply in a number of situations facing the investment professional besides those related directly to investing assets.

Part of a member's or candidate's duty of loyalty includes voting proxies in an informed and responsible manner. Proxies have economic value to a client, and members and candidates must ensure that they properly safeguard and maximize this value. An investment manager who fails to vote, casts a vote without considering the impact of the question, or votes blindly with management on nonroutine governance issues (e.g., a change in company capitalization) may violate this standard. Voting of proxies is an integral part of the management of investments.

A cost–benefit analysis may show that voting all proxies may not benefit the client, so voting proxies may not be necessary in all instances. Members and candidates should disclose to clients their proxy voting policies.

Recommended Procedures for Compliance

Regular Account Information

Members and candidates with control of client assets (1) should submit to each client, at least quarterly, an itemized statement showing the funds and securities in the custody or possession of the member or candidate plus all debits, credits, and transactions that occurred during the period, (2) should disclose to the client where the assets are to be maintained, as well as where or when they are moved, and (3) should separate the client's assets from any other party's assets, including the member's or candidate's own assets.

Client Approval

If a member or candidate is uncertain about the appropriate course of action with respect to a client, the member or candidate should consider what he or she would expect or demand if the member or candidate were the client. If in doubt, a member or candidate should disclose the questionable matter in writing to the client and obtain client approval.

Firm Policies

Members and candidates should address and encourage their firms to address the following topics when drafting the statements or manuals containing their policies and procedures regarding responsibilities to clients:

- *Follow all applicable rules and laws*: Members and candidates must follow all legal requirements and applicable provisions of the Code and Standards.
- *Establish the investment objectives of the client*: Make a reasonable inquiry into a client's investment experience, risk and return objectives, and financial constraints prior to making investment recommendations or taking investment actions.
- *Consider all the information when taking actions*: When taking investment actions, members and candidates must consider the appropriateness and suitability of the investment relative to (1) the client's needs and circumstances, (2) the investment's basic characteristics, and (3) the basic characteristics of the total portfolio.
- *Diversify*: Members and candidates should diversify investments to reduce the risk of loss, unless diversification is not consistent with plan guidelines or is contrary to the account objectives.
- *Carry out regular reviews*: Members and candidates should establish regular review schedules to ensure that the investments held in the account adhere to the terms of the governing documents.
- *Deal fairly with all clients with respect to investment actions*: Members and candidates must not favor some clients over others and should establish policies for allocating trades and disseminating investment recommendations.
- *Disclose conflicts of interest*: Members and candidates must disclose all actual and potential conflicts of interest so that clients can evaluate those conflicts.
- *Disclose compensation arrangements*: Members and candidates should make their clients aware of all forms of manager compensation.

- *Vote proxies*: In most cases, members and candidates should determine who is authorized to vote shares and vote proxies in the best interests of the clients and ultimate beneficiaries.
- *Maintain confidentiality*: Members and candidates must preserve the confidentiality of client information.
- *Seek best execution*: Unless directed by the client as ultimate beneficiary, members and candidates must seek best execution for their clients. (Best execution is defined in the preceding text.)
- *Place client interests first*: Members and candidates must serve the best interests of clients.

Application of the Standard

Example 1 (Identifying the Client—Plan Participants):

First Country Bank serves as trustee for the Miller Company's pension plan. Miller is the target of a hostile takeover attempt by Newton, Inc. In attempting to ward off Newton, Miller's managers persuade Julian Wiley, an investment manager at First Country Bank, to purchase Miller common stock in the open market for the employee pension plan. Miller's officials indicate that such action would be favorably received and would probably result in other accounts being placed with the bank. Although Wiley believes the stock is overvalued and would not ordinarily buy it, he purchases the stock to support Miller's managers, to maintain Miller's good favor toward the bank, and to realize additional new business. The heavy stock purchases cause Miller's market price to rise to such a level that Newton retracts its takeover bid.

> *Comment*: Standard III(A) requires that a member or candidate, in evaluating a takeover bid, act prudently and solely in the interests of plan participants and beneficiaries. To meet this requirement, a member or candidate must carefully evaluate the long-term prospects of the company against the short-term prospects presented by the takeover offer and by the ability to invest elsewhere. In this instance, Wiley, acting on behalf of his employer, which was the trustee for a pension plan, clearly violated Standard III(A). He used the pension plan to perpetuate existing management, perhaps to the detriment of plan participants and the company's shareholders, and to benefit himself. Wiley's responsibilities to the plan participants and beneficiaries should have taken precedence over any ties of his bank to corporate managers and over his self-interest. Wiley had a duty to examine the takeover offer on its own merits and to make an independent decision. The guiding principle is the appropriateness of the investment decision to the pension plan, not whether the decision benefited Wiley or the company that hired him.

Example 2 (Client Commission Practices):

JNI, a successful investment counseling firm, serves as investment manager for the pension plans of several large regionally based companies. Its trading activities generate a significant amount of commission-related business. JNI uses the brokerage and research services of many firms, but most of its trading activity is handled through a large brokerage company, Thompson, Inc., because the executives of the two firms have a close friendship. Thompson's commission structure is high in comparison with charges for similar brokerage services from other firms. JNI considers Thompson's research services and execution capabilities average. In exchange for JNI directing its brokerage to Thompson, Thompson absorbs a number of JNI overhead expenses, including those for rent.

> *Comment*: JNI executives are breaching their responsibilities by using client brokerage for services that do not benefit JNI clients and by not obtaining best price and best execution for their clients. Because JNI executives are not upholding their duty of loyalty, they are violating Standard III(A).

Example 3 (Brokerage Arrangements):

Charlotte Everett, a struggling independent investment adviser, serves as investment manager for the pension plans of several companies. One of her brokers, Scott Company, is close to consummating management agreements with prospective new clients whereby Everett would manage the new client accounts and trade the accounts exclusively through Scott. One of Everett's existing clients, Crayton Corporation, has directed Everett to place securities transactions for Crayton's account exclusively through Scott. But to induce Scott to exert efforts to send more new accounts to her, Everett also directs transactions to Scott from other clients without their knowledge.

> *Comment*: Everett has an obligation at all times to seek best price and best execution on all trades. Everett may direct new client trades exclusively through Scott Company as long as Everett receives best price and execution on the trades or receives a written statement from new clients that she is *not* to seek best price and execution and that they are aware of the consequence for their accounts. Everett may trade other accounts through Scott as a reward for directing clients to Everett only if the accounts receive best price and execution and the practice is disclosed to the accounts. Because Everett does not disclose the directed trading, Everett has violated Standard III(A).

Example 4 (Brokerage Arrangements):

Emilie Rome is a trust officer for Paget Trust Company. Rome's supervisor is responsible for reviewing Rome's trust account transactions and her monthly reports of personal stock transactions. Rome has been using Nathan Gray, a broker, almost exclusively for trust account brokerage transactions. When Gray makes a market in stocks, he has been giving Rome a lower price for personal purchases and a higher price for sales than he gives to Rome's trust accounts and other investors.

> *Comment*: Rome is violating her duty of loyalty to the bank's trust accounts by using Gray for brokerage transactions simply because Gray trades Rome's personal account on favorable terms. Rome is placing her own interests before those of her clients.

Example 5 (Client Commission Practices):

Lauren Parker, an analyst with Provo Advisors, covers South American equities for her firm. She likes to travel to the markets for which she is responsible and decides to go on a trip to Chile, Argentina, and Brazil. The trip is sponsored by SouthAM, Inc., a research firm with a small broker/dealer affiliate that uses the clearing facilities of a larger New York brokerage house. SouthAM specializes in arranging South American trips for analysts during which they can meet with central bank officials, government ministers, local economists, and senior executives of corporations. SouthAM accepts commission dollars at a ratio of 2 to 1 against the hard-dollar costs of the research fee for the trip. Parker is not sure that SouthAM's execution is competitive, but without informing her supervisor, she directs the trading desk at Provo to start giving commission business to SouthAM so she can take the trip. SouthAM has conveniently timed the briefing trip to coincide with the beginning of Carnival season, so Parker also decides to spend five days of vacation in Rio de Janeiro at the end of the trip. Parker uses commission dollars to pay for the five days of hotel expenses.

> *Comment*: Parker is violating Standard III(A) by not exercising her duty of loyalty to her clients. She should have determined whether the commissions charged by SouthAM are reasonable in relation to the benefit of the research provided by the trip. She also should have determined whether best execution and prices could be received from SouthAM. In addition, the five extra days are not part of the research effort because they do not assist in the investment decision making. Thus, the hotel expenses for the five days should not be paid for with client assets.

Example 6 (Excessive Trading):

Vida Knauss manages the portfolios of a number of high-net-worth individuals. A major part of her investment management fee is based on trading commissions. Knauss engages in extensive trading for each of her clients to ensure that she attains the minimum commission level set by her firm. Although the securities purchased and sold for the clients are appropriate and fall within the acceptable asset classes for the clients, the amount of trading for each account exceeds what is necessary to accomplish the client's investment objectives.

> *Comment*: Knauss has violated Standard III(A) because she is using the assets of her clients to benefit her firm and herself.

Example 7 (Managing Family Accounts):

Adam Dill recently joined New Investments Asset Managers. To assist Dill in building a book of clients, both his father and brother opened new fee-paying accounts. Dill followed all the firm's procedures in noting his relationships with these clients and in developing their investment policy statements.

After several years, the number of Dill's clients has grown, but he still manages the original accounts of his family members. An IPO is coming to market that is a suitable investment for many of his clients, including his brother. Dill does not receive the amount of stock he requested, so to avoid any appearance of a conflict of interest, he does not allocate any shares to his brother's account.

> *Comment*: Dill has violated Standard III(A) because he is not acting for the benefit of his brother's account as well as his other accounts. The brother's account is a regular fee-paying account comparable to the accounts of his other clients. By not allocating the shares proportionately across *all* accounts for which he thought the IPO was suitable, Dill is disadvantaging specific clients.
>
> Dill would have been correct in not allocating shares to his brother's account if that account was being managed outside the normal fee structure of the firm.

Example 8 (Identifying the Client):

Donna Hensley has been hired by a law firm to testify as an expert witness. Although the testimony is intended to represent impartial advice, she is concerned that her work may have negative consequences for the law firm. If the law firm is Hensley's client, how does she ensure that her testimony will not violate the required duty of loyalty, prudence, and care to one's client?

Comment: In this situation, the law firm represents Hensley's employer and the aspect of "who is the client" is not well defined. When acting as an expert witness, Hensley is bound by the standard of independence and objectivity in the same manner as an independent research analyst would be bound. Hensley must not let the law firm influence the testimony she provides in the legal proceedings.

Example 9 (Identifying the Client):

Jon Miller is a mutual fund portfolio manager. The fund is focused on the global financial services sector. Wanda Spears is a private wealth manager in the same city as Miller and is a friend of Miller. At a local CFA Institute society meeting, Spears mentions to Miller that her new client is an investor in Miller's fund. She states that the two of them now share a responsibility to this client.

Comment: Spears' statement is not totally correct. Because she provides the advisory services to her new client, she alone is bound by the duty of loyalty to this client. Miller's responsibility is to manage the fund according to the investment policy statement of the fund. His actions should not be influenced by the needs of any particular fund investor.

Example 10 (Client Loyalty):

After providing client account investment performance to the external-facing departments but prior to it being finalized for release to clients, Teresa Nguyen, an investment performance analyst, notices the reporting system missed a trade. Correcting the omission resulted in a large loss for a client that had previously placed the firm on "watch" for potential termination owing to underperformance in prior periods. Nguyen knows this news is unpleasant but informs the appropriate individuals that the report needs to be updated before releasing it to the client.

Comment: Nguyen's actions align with the requirements of Standard III(A). Even though the correction may lead to the firm's termination by the client, withholding information on errors would not be in the best interest of the client.

Example 11 (Execution-Only Responsibilities):

Baftija Sulejman recently became a candidate in the CFA Program. He is a broker who executes client-directed trades for several high-net-worth individuals. Sulejman does not provide any investment advice and only executes the trading decisions made by clients. He is concerned that the Code and Standards impose a fiduciary duty on him in his dealing with clients and sends an e-mail to the CFA Ethics Helpdesk (ethics@cfainstitute.org) to seek guidance on this issue.

Comment: In this instance, Sulejman serves in an execution-only capacity and his duty of loyalty, prudence, and care is centered on the skill and diligence used when executing trades—namely, by seeking best execution and making trades within the parameters set by the clients (instructions on quantity, price, timing, etc.). Acting in the best interests of the client dictates that trades are executed on the most favorable terms that can be achieved for the client. Given this job function, the requirements of the Code and Standards for loyalty, prudence, and care clearly do not impose a fiduciary duty.

Standard III(B) Fair Dealing

Members and Candidates must deal fairly and objectively with all clients when providing investment analysis, making investment recommendations, taking investment action, or engaging in other professional activities.

Guidance

Highlights:

- *Investment Recommendations*
- *Investment Action*

Standard III(B) requires members and candidates to treat all clients fairly when disseminating investment recommendations or making material changes to prior investment recommendations or when taking investment action with regard to general purchases, new issues, or secondary offerings. Only through the fair treatment of all parties can the investment management profession maintain the confidence of the investing public.

When an investment adviser has multiple clients, the potential exists for the adviser to favor one client over another. This favoritism may take various forms—from the quality and timing of services provided to the allocation of investment opportunities.

The term “fairly” implies that the member or candidate must take care not to discriminate against any clients when disseminating investment recommendations or taking investment action. Standard III(B) does not state “equally” because members and candidates could not possibly reach all clients at exactly the same time—whether by printed mail, telephone (including text messaging), computer (including internet updates and e-mail distribution), facsimile (fax), or wire. Each client has unique needs, investment criteria, and investment objectives, so not all investment opportunities are suitable for all clients. In addition, members and candidates may provide more personal, specialized, or in-depth service to clients who are willing to pay for premium services through higher management fees or higher levels of brokerage. Members and candidates may differentiate their services to clients, but different levels of service must not disadvantage or negatively affect clients. In addition, the different service levels should be disclosed to clients and prospective clients and should be available to everyone (i.e., different service levels should not be offered selectively).

Standard III(B) covers conduct in two broadly defined categories—investment recommendations and investment action.

Investment Recommendations

The first category of conduct involves members and candidates whose primary function is the preparation of investment recommendations to be disseminated either to the public or within a firm for the use of others in making investment decisions. This group includes members and candidates employed by investment counseling, advisory, or consulting firms as well as banks, brokerage firms, and insurance companies. The criterion is that the member’s or candidate’s primary responsibility is the preparation of recommendations to be acted on by others, including those in the member’s or candidate’s organization.

An investment recommendation is any opinion expressed by a member or candidate in regard to purchasing, selling, or holding a given security or other investment. The opinion may be disseminated to customers or clients through an initial detailed research report, through a brief update report, by addition to or deletion from a list of recommended securities, or simply by oral communication. A recommendation that is distributed to anyone outside the organization is considered a communication for general distribution under Standard III(B).

Standard III(B) addresses the manner in which investment recommendations or changes in prior recommendations are disseminated to clients. Each member or candidate is obligated to ensure that information is disseminated in such a manner that all clients have a fair opportunity to act on every recommendation. Communicating with all clients on a uniform basis presents practical problems for members and candidates because of differences in timing and methods of communication with various types of customers and clients. Members and candidates should encourage their firms to design an equitable system to prevent selective or discriminatory disclosure and should inform clients about what kind of communications they will receive.

The duty to clients imposed by Standard III(B) may be more critical when members or candidates change their recommendations than when they make initial recommendations. Material changes in a member's or candidate's prior investment recommendations because of subsequent research should be communicated to all current clients; particular care should be taken that the information reaches those clients who the member or candidate knows have acted on or been affected by the earlier advice. Clients who do not know that the member or candidate has changed a recommendation and who, therefore, place orders contrary to a current recommendation should be advised of the changed recommendation before the order is accepted.

Investment Action

The second category of conduct includes those members and candidates whose primary function is taking investment action (portfolio management) on the basis of recommendations prepared internally or received from external sources. Investment action, like investment recommendations, can affect market value. Consequently, Standard III(B) requires that members or candidates treat all clients fairly in light of their investment objectives and circumstances. For example, when making investments in new offerings or in secondary financings, members and candidates should distribute the issues to all customers for whom the investments are appropriate in a manner consistent with the policies of the firm for allocating blocks of stock. If the issue is oversubscribed, then the issue should be prorated to all subscribers. This action should be taken on a round-lot basis to avoid odd-lot distributions. In addition, if the issue is oversubscribed, members and candidates should forgo any sales to themselves or their immediate families in order to free up additional shares for clients. If the investment professional's family-member accounts are managed similarly to the accounts of other clients of the firm, however, the family-member accounts should not be excluded from buying such shares.

Members and candidates must make every effort to treat all individual and institutional clients in a fair and impartial manner. A member or candidate may have multiple relationships with an institution; for example, the member or candidate may be a corporate trustee, pension fund manager, manager of funds for individuals employed by the customer, loan originator, or creditor. A member or candidate must exercise care to treat all clients fairly.

Members and candidates should disclose to clients and prospective clients the documented allocation procedures they or their firms have in place and how the procedures would affect the client or prospect. The disclosure should be clear and complete so that the client can make an informed investment decision. Even when complete disclosure is made, however, members and candidates must put client interests ahead of their own. A member's or candidate's duty of fairness and loyalty to clients can never be overridden by client consent to patently unfair allocation procedures.

Treating clients fairly also means that members and candidates should not take advantage of their position in the industry to the detriment of clients. For instance, in the context of IPOs, members and candidates must make bona fide public distributions of "hot issue" securities (defined as securities of a public offering that are trading at a premium in the secondary market whenever such trading commences because of the great demand for the securities). Members and candidates are prohibited from withholding such securities for their own benefit and must not use such securities as a reward or incentive to gain benefit.

Recommended Procedures for Compliance

Develop Firm Policies

Although Standard III(B) refers to a member's or candidate's responsibility to deal fairly and objectively with clients, members and candidates should also encourage their firms to establish compliance procedures requiring all employees who disseminate investment recommendations or take investment actions to treat customers and clients fairly. At the very least, a member or candidate should recommend appropriate procedures to management if none are in place. And the member or candidate should make management aware of possible violations of fair-dealing practices within the firm when they come to the attention of the member or candidate.

The extent of the formality and complexity of such compliance procedures depends on the nature and size of the organization and the type of securities involved. An investment adviser who is a sole proprietor and handles only discretionary accounts might not disseminate recommendations to the public, but that adviser should have formal written procedures to ensure that all clients receive fair investment action.

Good business practice dictates that initial recommendations be made available to all customers who indicate an interest. Although a member or candidate need not communicate a recommendation to all customers, the selection process by which customers receive information should be based on suitability and known interest, not on any preferred or favored status. A common practice to assure fair dealing is to communicate recommendations simultaneously within the firm and to customers.

Members and candidates should consider the following points when establishing fair-dealing compliance procedures:

- *Limit the number of people involved*: Members and candidates should make reasonable efforts to limit the number of people who are privy to the fact that a recommendation is going to be disseminated.
- *Shorten the time frame between decision and dissemination*: Members and candidates should make reasonable efforts to limit the amount of time that elapses between the decision to make an investment recommendation and the time the actual recommendation is disseminated. If a detailed institutional recommendation that might take two or three weeks to publish is in preparation, a short summary report including the conclusion might be published in advance. In an organization where both a research committee and an investment policy committee must approve a recommendation, the meetings should be held on the same day if possible. The process of reviewing reports and printing and mailing them, faxing them, or distributing them by e-mail necessarily involves the passage of time, sometimes long periods of time. In large firms with extensive review processes, the time factor is usually not within the control of the analyst who prepares the report. Thus, many firms and their analysts communicate

to customers and firm personnel the new or changed recommendations by an update or "flash" report. The communication technique might be fax, e-mail, wire, or short written report.

- *Publish guidelines for pre-dissemination behavior*: Members and candidates should encourage firms to develop guidelines that prohibit personnel who have prior knowledge of an investment recommendation from discussing or taking any action on the pending recommendation.
- *Simultaneous dissemination*: Members and candidates should establish procedures for the timing of dissemination of investment recommendations so that all clients are treated fairly—that is, are informed at approximately the same time. For example, if a firm is going to announce a new recommendation, supervisory personnel should time the announcement to avoid placing any client or group of clients at an unfair advantage relative to other clients. A communication to all branch offices should be sent at the time of the general announcement. (When appropriate, the firm should accompany the announcement of a new recommendation with a statement that trading restrictions for the firm's employees are now in effect. The trading restrictions should stay in effect until the recommendation is widely distributed to all relevant clients.) Once this distribution has occurred, the member or candidate may follow up separately with individual clients, but members and candidates should not give favored clients advance information when such advance notification may disadvantage other clients.
- *Maintain a list of clients and their holdings*: Members and candidates should maintain a list of all clients and the securities or other investments each client holds in order to facilitate notification of customers or clients of a change in an investment recommendation. If a particular security or other investment is to be sold, such a list can be used to ensure that all holders are treated fairly in the liquidation of that particular investment.
- *Develop and document trade allocation procedures*: When formulating procedures for allocating trades, members and candidates should develop a set of guiding principles that ensure
  - fairness to advisory clients, both in priority of execution of orders and in the allocation of the price obtained in execution of block orders or trades,
  - timeliness and efficiency in the execution of orders, and
  - accuracy of the member's or candidate's records as to trade orders and client account positions.

With these principles in mind, members and candidates should develop or encourage their firm to develop written allocation procedures, with particular attention to procedures for block trades and new issues. Procedures to consider are as follows:

- requiring orders and modifications or cancellations of orders to be documented and time stamped;
- processing and executing orders on a first-in, first-out basis with consideration of bundling orders for efficiency as appropriate for the asset class or the security;
- developing a policy to address such issues as calculating execution prices and "partial fills" when trades are grouped, or in a block, for efficiency;
- giving all client accounts participating in a block trade the same execution price and charging the same commission;

- when the full amount of the block order is not executed, allocating partially executed orders among the participating client accounts pro rata on the basis of order size while not going below an established minimum lot size for some securities (e.g., bonds); and
- when allocating trades for new issues, obtaining advance indications of interest, allocating securities by client (rather than portfolio manager), and providing a method for calculating allocations.

Disclose Trade Allocation Procedures

Members and candidates should disclose to clients and prospective clients how they select accounts to participate in an order and how they determine the amount of securities each account will buy or sell. Trade allocation procedures must be fair and equitable, and disclosure of inequitable allocation methods does not relieve the member or candidate of this obligation.

Establish Systematic Account Review

Member and candidate supervisors should review each account on a regular basis to ensure that no client or customer is being given preferential treatment and that the investment actions taken for each account are suitable for each account's objectives. Because investments should be based on individual needs and circumstances, an investment manager may have good reasons for placing a given security or other investment in one account while selling it from another account and should fully document the reasons behind both sides of the transaction. Members and candidates should encourage firms to establish review procedures, however, to detect whether trading in one account is being used to benefit a favored client.

Disclose Levels of Service

Members and candidates should disclose to all clients whether the organization offers different levels of service to clients for the same fee or different fees. Different levels of service should not be offered to clients selectively.

Application of the Standard

Example 1 (Selective Disclosure):

Bradley Ames, a well-known and respected analyst, follows the computer industry. In the course of his research, he finds that a small, relatively unknown company whose shares are traded over the counter has just signed significant contracts with some of the companies he follows. After a considerable amount of investigation, Ames decides to write a research report on the small company and recommend purchase of its shares. While the report is being reviewed by the company for factual accuracy, Ames schedules a luncheon with several of his best clients to discuss the company. At the luncheon, he mentions the purchase recommendation scheduled to be sent early the following week to all the firm's clients.

> *Comment*: Ames has violated Standard III(B) by disseminating the purchase recommendation to the clients with whom he has lunch a week before the recommendation is sent to all clients.

Example 2 (Fair Dealing between Funds):

Spencer Rivers, president of XYZ Corporation, moves his company's growth-oriented pension fund to a particular bank primarily because of the excellent investment performance achieved by the bank's commingled fund for the prior five-year period.

Later, Rivers compares the results of his pension fund with those of the bank's commingled fund. He is startled to learn that, even though the two accounts have the same investment objectives and similar portfolios, his company's pension fund has significantly underperformed the bank's commingled fund. Questioning this result at his next meeting with the pension fund's manager, Rivers is told that, as a matter of policy, when a new security is placed on the recommended list, Morgan Jackson, the pension fund manager, first purchases the security for the commingled account and then purchases it on a pro rata basis for all other pension fund accounts. Similarly, when a sale is recommended, the security is sold first from the commingled account and then sold on a pro rata basis from all other accounts. Rivers also learns that if the bank cannot get enough shares (especially of hot issues) to be meaningful to all the accounts, its policy is to place the new issues only in the commingled account.

Seeing that Rivers is neither satisfied nor pleased by the explanation, Jackson quickly adds that nondiscretionary pension accounts and personal trust accounts have a lower priority on purchase and sale recommendations than discretionary pension fund accounts. Furthermore, Jackson states, the company's pension fund had the opportunity to invest up to 5% in the commingled fund.

> *Comment*: The bank's policy does not treat all customers fairly, and Jackson has violated her duty to her clients by giving priority to the growth-oriented commingled fund over all other funds and to discretionary accounts over nondiscretionary accounts. Jackson must execute orders on a systematic basis that is fair to all clients. In addition, trade allocation procedures should be disclosed to all clients when they become clients. Of course, in this case, disclosure of the bank's policy would not change the fact that the policy is unfair.

Example 3 (Fair Dealing and IPO Distribution):

Dominic Morris works for a small regional securities firm. His work consists of corporate finance activities and investing for institutional clients. Arena, Ltd., is planning to go public. The partners have secured rights to buy an arena football league franchise and are planning to use the funds from the issue to complete the purchase. Because arena football is the current rage, Morris believes he has a hot issue on his hands. He has quietly negotiated some options for himself for helping convince Arena to do the financing through his securities firm. When he seeks expressions of interest, the institutional buyers oversubscribe the issue. Morris, assuming that the institutions have the financial clout to drive the stock up, then fills all orders (including his own) and decreases the institutional blocks.

> *Comment*: Morris has violated Standard III(B) by not treating all customers fairly. He should not have taken any shares himself and should have prorated the shares offered among all clients. In addition, he should have disclosed to his firm and to his clients that he received options as part of the deal [see Standard VI(A)–Disclosure of Conflicts].

Example 4 (Fair Dealing and Transaction Allocation):

Eleanor Preston, the chief investment officer of Porter Williams Investments (PWI), a medium-size money management firm, has been trying to retain a client, Colby Company. Management at Colby, which accounts for almost half of PWI's revenues, recently told Preston that if the performance of its account did not improve, it would find a new money manager. Shortly after this threat, Preston purchases mortgage-backed securities (MBSs) for several accounts, including Colby's. Preston is busy with a number of transactions that day, so she fails to allocate the trades immediately or write up the trade tickets. A few days later, when Preston is allocating trades, she notes that some of the MBSs have significantly increased in price and some have dropped. Preston decides to allocate the profitable trades to Colby and spread the losing trades among several other PWI accounts.

> *Comment*: Preston has violated Standard III(B) by failing to deal fairly with her clients in taking these investment actions. Preston should have allocated the trades prior to executing the orders, or she should have had a systematic approach to allocating the trades, such as pro rata, as soon as practical after they were executed. Among other things, Preston must disclose to the client that the adviser may act as broker for, receive commissions from, and have a potential conflict of interest regarding both parties in agency cross-transactions. After the disclosure, she should obtain from the client consent authorizing such transactions in advance.

Example 5 (Selective Disclosure):

Saunders Industrial Waste Management (SIWM) publicly indicates to analysts that it is comfortable with the somewhat disappointing earnings-per-share projection of US$1.16 for the quarter. Bernard Roberts, an analyst at Coffey Investments, is confident that SIWM management has understated the forecasted earnings so that the real announcement will cause an “upside surprise” and boost the price of SIWM stock. The “whisper number” (rumored) estimate based on extensive research and discussed among knowledgeable analysts is higher than US$1.16. Roberts repeats the US$1.16 figure in his research report to all Coffey clients but informally tells his large clients that he expects the earnings per share to be higher, making SIWM a good buy.

> *Comment*: By not sharing his opinion regarding the potential for a significant upside earnings surprise with all clients, Roberts is not treating all clients fairly and has violated Standard III(B).

Example 6 (Additional Services for Select Clients):

Jenpin Weng uses e-mail to issue a new recommendation to all his clients. He then calls his three largest institutional clients to discuss the recommendation in detail.

> *Comment*: Weng has not violated Standard III(B) because he widely disseminated the recommendation and provided the information to all his clients prior to discussing it with a select few. Weng’s largest clients received additional personal service because they presumably pay higher fees or because they have a large amount of assets under Weng’s management. If Weng had discussed the report with a select group of clients prior to distributing it to all his clients, he would have violated Standard III(B).

Example 7 (Minimum Lot Allocations):

Lynn Hampton is a well-respected private wealth manager in her community with a diversified client base. She determines that a new 10-year bond being offered by Healthy Pharmaceuticals is appropriate for five of her clients. Three clients request to purchase US$10,000 each, and the other two request US$50,000 each. The minimum lot size is established at US$5,000, and the issue is oversubscribed at the time of placement. Her firm’s policy is that odd-lot allocations, especially those below the minimum, should be avoided because they may affect the liquidity of the security at the time of sale.

Hampton is informed she will receive only US$55,000 of the offering for all accounts. Hampton distributes the bond investments as follows: The three accounts that requested US$10,000 are allocated US$5,000 each, and the two accounts that requested US$50,000 are allocated US$20,000 each.

> *Comment*: Hampton has not violated Standard III(B), even though the distribution is not on a completely pro rata basis because of the required minimum lot size. With the total allocation being significantly below the amount requested, Hampton ensured that each client received at least the minimum lot size of the issue. This approach allowed the clients to efficiently sell the bond later if necessary.

Example 8 (Excessive Trading):

Ling Chan manages the accounts for many pension plans, including the plan of his father's employer. Chan developed similar but not identical investment policies for each client, so the investment portfolios are rarely the same. To minimize the cost to his father's pension plan, he intentionally trades more frequently in the accounts of other clients to ensure the required brokerage is incurred to continue receiving free research for use by all the pensions.

> *Comment*: Chan is violating Standard III(B) because his trading actions are disadvantaging his clients to enhance a relationship with a preferred client. All clients are benefiting from the research being provided and should incur their fair portion of the costs. This does not mean that additional trading should occur if a client has not paid an equal portion of the commission; trading should occur only as required by the strategy.

Example 9 (Limited Social Media Disclosures):

Mary Burdette was recently hired by Fundamental Investment Management (FIM) as a junior auto industry analyst. Burdette is expected to expand the social media presence of the firm because she is active with various networks, including Facebook, LinkedIn, and Twitter. Although Burdette's supervisor, Joe Graf, has never used social media, he encourages Burdette to explore opportunities to increase FIM's online presence and ability to share content, communicate, and broadcast information to clients. In response to Graf's encouragement, Burdette is working on a proposal detailing the advantages of getting FIM onto Twitter in addition to launching a company Facebook page.

As part of her auto industry research for FIM, Burdette is completing a report on the financial impact of Sun Drive Auto Ltd.'s new solar technology for compact automobiles. This research report will be her first for FIM, and she believes Sun Drive's technology could revolutionize the auto industry. In her excitement, Burdette sends a quick tweet to FIM Twitter followers summarizing her "buy" recommendation for Sun Drive Auto stock.

> *Comment*: Burdette has violated Standard III(B) by sending an investment recommendation to a select group of contacts prior to distributing it to all clients. Burdette must make sure she has received the appropriate training about FIM's policies and procedures, including the appropriate business use of personal social media networks before engaging in such activities.
>
> See Standard IV(C) for guidance related to the duties of the supervisor.

Example 10 (Fair Dealing between Clients):

Paul Rove, performance analyst for Alpha-Beta Investment Management, is describing to the firm's chief investment officer (CIO) two new reports he would like to develop to assist the firm in meeting its obligations to treat clients fairly. Because many of the firm's clients have similar investment objectives and portfolios, Rove suggests a report detailing securities owned across several clients and the percentage of the portfolio the security represents. The second report would compare the monthly performance of portfolios with similar strategies. The outliers within each report would be submitted to the CIO for review.

> *Comment*: As a performance analyst, Rove likely has little direct contact with clients and thus has limited opportunity to treat clients differently. The recommended reports comply with Standard III(B) while helping the firm conduct after-the-fact reviews of how effectively the firm's advisers are dealing with their clients' portfolios. Reports that monitor the fair treatment of clients are an important oversight tool to ensure that clients are treated fairly.

Standard III(C) Suitability

1 When Members and Candidates are in an advisory relationship with a client, they must:

- **a** Make a reasonable inquiry into a client's or prospective client's investment experience, risk and return objectives, and financial constraints prior to making any investment recommendation or taking investment action and must reassess and update this information regularly.
- **b** Determine that an investment is suitable to the client's financial situation and consistent with the client's written objectives, mandates, and constraints before making an investment recommendation or taking investment action.
- **c** Judge the suitability of investments in the context of the client's total portfolio.

2 When Members and Candidates are responsible for managing a portfolio to a specific mandate, strategy, or style, they must make only investment recommendations or take only investment actions that are consistent with the stated objectives and constraints of the portfolio.

Guidance

Highlights:

- *Developing an Investment Policy*
- *Understanding the Client's Risk Profile*
- *Updating an Investment Policy*
- *The Need for Diversification*
- *Addressing Unsolicited Trading Requests*
- *Managing to an Index or Mandate*

Standard III(C) requires that members and candidates who are in an investment advisory relationship with clients consider carefully the needs, circumstances, and objectives of the clients when determining the appropriateness and suitability of a given investment or course of investment action. An appropriate suitability determination will not, however, prevent some investments or investment actions from losing value.

In judging the suitability of a potential investment, the member or candidate should review many aspects of the client's knowledge, experience related to investing, and financial situation. These aspects include, but are not limited to, the risk profile of the investment as compared with the constraints of the client, the impact of the investment on the diversity of the portfolio, and whether the client has the means or net worth to assume the associated risk. The investment professional's determination of suitability should reflect only the investment recommendations or actions that a prudent person would be willing to undertake. Not every investment opportunity will be suitable for every portfolio, regardless of the potential return being offered.

The responsibilities of members and candidates to gather information and make a suitability analysis prior to making a recommendation or taking investment action fall on those members and candidates who provide investment advice in the course of an advisory relationship with a client. Other members and candidates may be simply executing specific instructions for retail clients when buying or selling securities, such as shares in mutual funds. These members and candidates and some others, such as sell-side analysts, may not have the opportunity to judge the suitability of a particular investment for the ultimate client.

Developing an Investment Policy

When an advisory relationship exists, members and candidates must gather client information at the inception of the relationship. Such information includes the client's financial circumstances, personal data (such as age and occupation) that are relevant to investment decisions, attitudes toward risk, and objectives in investing. This information should be incorporated into a written investment policy statement (IPS) that addresses the client's risk tolerance, return requirements, and all investment constraints (including time horizon, liquidity needs, tax concerns, legal and regulatory factors, and unique circumstances). Without identifying such client factors, members and candidates cannot judge whether a particular investment or strategy is suitable for a particular client. The IPS also should identify and describe the roles and responsibilities of the parties to the advisory relationship and investment process, as well as schedules for review and evaluation of the IPS. After formulating long-term capital market expectations, members and candidates can assist in developing an appropriate strategic asset allocation and investment program for the client, whether these are presented in separate documents or incorporated in the IPS or in appendices to the IPS.

Understanding the Client's Risk Profile

One of the most important factors to be considered in matching appropriateness and suitability of an investment with a client's needs and circumstances is measuring that client's tolerance for risk. The investment professional must consider the possibilities of rapidly changing investment environments and their likely impact on a client's holdings, both individual securities and the collective portfolio. The risk of many investment strategies can and should be analyzed and quantified in advance.

The use of synthetic investment vehicles and derivative investment products has introduced particular issues of risk. Members and candidates should pay careful attention to the leverage inherent in many of these vehicles or products when considering them for use in a client's investment program. Such leverage and limited liquidity, depending on the degree to which they are hedged, bear directly on the issue of suitability for the client.

Updating an Investment Policy

Updating the IPS should be repeated at least annually and also prior to material changes to any specific investment recommendations or decisions on behalf of the client. The effort to determine the needs and circumstances of each client is not a one-time occurrence. Investment recommendations or decisions are usually part of an ongoing process that takes into account the diversity and changing nature of portfolio and client characteristics. The passage of time is bound to produce changes that are important with respect to investment objectives.

For an individual client, important changes might include the number of dependents, personal tax status, health, liquidity needs, risk tolerance, amount of wealth beyond that represented in the portfolio, and extent to which compensation and other income provide for current income needs. With respect to an institutional client, such changes might relate to the magnitude of unfunded liabilities in a pension fund, the withdrawal privileges in an employee savings plan, or the distribution requirements of a charitable foundation. Without efforts to update information concerning client factors, one or more factors could change without the investment manager's knowledge.

Suitability review can be done most effectively when the client fully discloses his or her complete financial portfolio, including those portions not managed by the member or candidate. If clients withhold information about their financial portfolios, the suitability analysis conducted by members and candidates cannot be expected to be complete; it must be based on the information provided.

The Need for Diversification

The investment profession has long recognized that combining several different investments is likely to provide a more acceptable level of risk exposure than having all assets in a single investment. The unique characteristics (or risks) of an individual investment may become partially or entirely neutralized when it is combined with other individual investments within a portfolio. Some reasonable amount of diversification is thus the norm for many portfolios, especially those managed by individuals or institutions that have some degree of legal fiduciary responsibility.

An investment with high relative risk on its own may be a suitable investment in the context of the entire portfolio or when the client's stated objectives contemplate speculative or risky investments. The manager may be responsible for only a portion of the client's total portfolio, or the client may not have provided a full financial picture. Members and candidates can be responsible for assessing the suitability of an investment only on the basis of the information and criteria actually provided by the client.

Addressing Unsolicited Trading Requests

Members and candidates may receive requests from a client for trades that do not properly align with the risk and return objectives outlined in the client's investment policy statement. These transaction requests may be based on the client's individual biases or professional experience. Members and candidates will need to make reasonable efforts to balance their clients' trading requests with their responsibilities to follow the agreed-on investment policy statement.

In cases of unsolicited trade requests that a member or candidate knows are unsuitable for a client, the member or candidate should refrain from making the trade until he or she discusses the concerns with the client. The discussions and resulting actions may encompass a variety of scenarios depending on how the requested unsuitable investment relates to the client's full portfolio.

Many times, an unsolicited request may be expected to have only a minimum impact on the entire portfolio because the size of the requested trade is small or the trade would result in a limited change to the portfolio's risk profile. In discussing the trade, the member or candidate should focus on educating the investor on how the request deviates from the current policy statement. Following the discussion, the member or candidate may follow his or her firm's policies regarding the necessary client approval for executing unsuitable trades. At a minimum, the client should acknowledge the discussion and accept the conditions that make the recommendation unsuitable.

Should the unsolicited request be expected to have a material impact on the portfolio, the member or candidate should use this opportunity to update the investment policy statement. Doing so would allow the client to fully understand the potential effect of the requested trade on his or her current goals or risk levels.

Members and candidates may have some clients who decline to modify their policy statements while insisting an unsolicited trade be made. In such instances, members or candidates will need to evaluate the effectiveness of their services to the client. The options available to the members or candidates will depend on the services provided by their employer. Some firms may allow for the trade to be executed in a new unmanaged account. If alternative options are not available, members and candidates ultimately will need to determine whether they should continue the advisory arrangement with the client.

Managing to an Index or Mandate

Some members and candidates do not manage money for individuals but are responsible for managing a fund to an index or an expected mandate. The responsibility of these members and candidates is to invest in a manner consistent with the stated mandate. For example, a member or candidate who serves as the fund manager for a large-cap income fund would not be following the fund mandate by investing heavily in small-cap or start-up companies whose stock is speculative in nature. Members and candidates who manage pooled assets to a specific mandate are not responsible for determining the suitability of the *fund* as an investment for investors who may be purchasing shares in the fund. The responsibility for determining the suitability of an investment for clients can be conferred only on members and candidates who have an advisory relationship with clients.

Recommended Procedures for Compliance

Investment Policy Statement

To fulfill the basic provisions of Standard III(C), a member or candidate should put the needs and circumstances of each client and the client's investment objectives into a written investment policy statement. In formulating an investment policy for the client, the member or candidate should take the following into consideration:

- client identification—(1) type and nature of client, (2) the existence of separate beneficiaries, and (3) approximate portion of total client assets that the member or candidate is managing;
- investor objectives—(1) return objectives (income, growth in principal, maintenance of purchasing power) and (2) risk tolerance (suitability, stability of values);
- investor constraints—(1) liquidity needs, (2) expected cash flows (patterns of additions and/or withdrawals), (3) investable funds (assets and liabilities or other commitments), (4) time horizon, (5) tax considerations, (6) regulatory and legal circumstances, (7) investor preferences, prohibitions, circumstances, and unique needs, and (8) proxy voting responsibilities and guidance; and
- performance measurement benchmarks.

Regular Updates

The investor's objectives and constraints should be maintained and reviewed periodically to reflect any changes in the client's circumstances. Members and candidates should regularly compare client constraints with capital market expectations to arrive at an appropriate asset allocation. Changes in either factor may result in a fundamental change in asset allocation. Annual review is reasonable unless business or other reasons, such as a major change in market conditions, dictate more frequent review. Members and candidates should document attempts to carry out such a review if circumstances prevent it.

Suitability Test Policies

With the increase in regulatory required suitability tests, members and candidates should encourage their firms to develop related policies and procedures. The procedures will differ according to the size of the firm and the scope of the services offered to its clients.

The test procedures should require the investment professional to look beyond the potential return of the investment and include the following:

- an analysis of the impact on the portfolio's diversification,
- a comparison of the investment risks with the client's assessed risk tolerance, and
- the fit of the investment with the required investment strategy.

Application of the Standard

Example 1 (Investment Suitability—Risk Profile):

Caleb Smith, an investment adviser, has two clients: Larry Robertson, 60 years old, and Gabriel Lanai, 40 years old. Both clients earn roughly the same salary, but Robertson has a much higher risk tolerance because he has a large asset base. Robertson is willing to invest part of his assets very aggressively; Lanai wants only to achieve a steady rate of return with low volatility to pay for his children's education. Smith recommends investing 20% of both portfolios in zero-yield, small-cap, high-technology equity issues.

> *Comment*: In Robertson's case, the investment may be appropriate because of his financial circumstances and aggressive investment position, but this investment is not suitable for Lanai. Smith is violating Standard III(C) by applying Robertson's investment strategy to Lanai because the two clients' financial circumstances and objectives differ.

Example 2 (Investment Suitability—Entire Portfolio):

Jessica McDowell, an investment adviser, suggests to Brian Crosby, a risk-averse client, that covered call options be used in his equity portfolio. The purpose would be to enhance Crosby's income and partially offset any untimely depreciation in the portfolio's value should the stock market or other circumstances affect his holdings unfavorably. McDowell educates Crosby about all possible outcomes, including the risk of incurring an added tax liability if a stock rises in price and is called away and, conversely, the risk of his holdings losing protection on the downside if prices drop sharply.

> *Comment*: When determining suitability of an investment, the primary focus should be the characteristics of the client's entire portfolio, not the characteristics of single securities on an issue-by-issue basis. The basic characteristics of the entire portfolio will largely determine whether investment

recommendations are taking client factors into account. Therefore, the most important aspects of a particular investment are those that will affect the characteristics of the total portfolio. In this case, McDowell properly considers the investment in the context of the entire portfolio and thoroughly explains the investment to the client.

Example 3 (IPS Updating):

In a regular meeting with client Seth Jones, the portfolio managers at Blue Chip Investment Advisors are careful to allow some time to review his current needs and circumstances. In doing so, they learn that some significant changes have recently taken place in his life. A wealthy uncle left Jones an inheritance that increased his net worth fourfold, to US$1 million.

> *Comment*: The inheritance has significantly increased Jones's ability (and possibly his willingness) to assume risk and has diminished the average yield required to meet his current income needs. Jones's financial circumstances have definitely changed, so Blue Chip managers must update Jones's investment policy statement to reflect how his investment objectives have changed. Accordingly, the Blue Chip portfolio managers should consider a somewhat higher equity ratio for his portfolio than was called for by the previous circumstances, and the managers' specific common stock recommendations might be heavily tilted toward low-yield, growth-oriented issues.

Example 4 (Following an Investment Mandate):

Louis Perkowski manages a high-income mutual fund. He purchases zero-dividend stock in a financial services company because he believes the stock is undervalued and is in a potential growth industry, which makes it an attractive investment.

> *Comment*: A zero-dividend stock does not seem to fit the mandate of the fund that Perkowski is managing. Unless Perkowski's investment fits within the mandate or is within the realm of allowable investments the fund has made clear in its disclosures, Perkowski has violated Standard III(C).

Example 5 (IPS Requirements and Limitations):

Max Gubler, chief investment officer of a property/casualty insurance subsidiary of a large financial conglomerate, wants to improve the diversification of the subsidiary's investment portfolio and increase its returns. The subsidiary's investment policy statement provides for highly liquid investments, such as large-cap equities and government, supranational, and corporate bonds with a minimum credit rating of AA and maturity of no more than five years. In a recent presentation, a venture capital group offered very attractive prospective returns on some of its private equity funds that provide seed capital to ventures. An exit strategy was already contemplated, but investors would have to observe a minimum three-year lockup period and a subsequent laddered exit option for a maximum of one-third of their shares per year. Gubler does not want to miss this opportunity. After extensive analysis, with the intent to optimize the return on the equity assets within the subsidiary's current portfolio, he invests 4% in this seed fund, leaving the portfolio's total equity exposure still well below its upper limit.

> *Comment*: Gubler is violating Standard III(A)–Loyalty, Prudence, and Care as well as Standard III(C). His new investment locks up part of the subsidiary's assets for at least three years and up to as many as five years and possibly beyond. The IPS requires investments in highly liquid investments and describes accepted asset classes; private equity investments with

a lockup period certainly do not qualify. Even without a lockup period, an asset class with only an occasional, and thus implicitly illiquid, market may not be suitable for the portfolio. Although an IPS typically describes objectives and constraints in great detail, the manager must also make every effort to understand the client's business and circumstances. Doing so should enable the manager to recognize, understand, and discuss with the client other factors that may be or may become material in the investment management process.

Example 6 (Submanager and IPS Reviews):

Paul Ostrowski's investment management business has grown significantly over the past couple of years, and some clients want to diversify internationally. Ostrowski decides to find a submanager to handle the expected international investments. Because this will be his first subadviser, Ostrowski uses the CFA Institute model "request for proposal" to design a questionnaire for his search. By his deadline, he receives seven completed questionnaires from a variety of domestic and international firms trying to gain his business. Ostrowski reviews all the applications in detail and decides to select the firm that charges the lowest fees because doing so will have the least impact on his firm's bottom line.

Comment: When selecting an external manager or subadviser, Ostrowski needs to ensure that the new manager's services are appropriate for his clients. This due diligence includes comparing the risk profile of the clients with the investment strategy of the manager. In basing the decision on the fee structure alone, Ostrowski may be violating Standard III(C).

When clients ask to diversify into international products, it is an appropriate time to review and update the clients' IPSs. Ostrowski's review may determine that the risk of international investments modifies the risk profiles of the clients or does not represent an appropriate investment.

See also Standard V(A)–Diligence and Reasonable Basis for further discussion of the review process needed in selecting appropriate submanagers.

Example 7 (Investment Suitability—Risk Profile):

Samantha Snead, a portfolio manager for Thomas Investment Counsel, Inc., specializes in managing public retirement funds and defined benefit pension plan accounts, all of which have long-term investment objectives. A year ago, Snead's employer, in an attempt to motivate and retain key investment professionals, introduced a bonus compensation system that rewards portfolio managers on the basis of quarterly performance relative to their peers and to certain benchmark indexes. In an attempt to improve the short-term performance of her accounts, Snead changes her investment strategy and purchases several high-beta stocks for client portfolios. These purchases are seemingly contrary to the clients' investment policy statements. Following their purchase, an officer of Griffin Corporation, one of Snead's pension fund clients, asks why Griffin Corporation's portfolio seems to be dominated by high-beta stocks of companies that often appear among the most actively traded issues. No change in objective or strategy has been recommended by Snead during the year.

Comment: Snead violated Standard III(C) by investing the clients' assets in high-beta stocks. These high-risk investments are contrary to the long-term risk profile established in the clients' IPSs. Snead has changed the investment strategy of the clients in an attempt to reap short-term rewards offered by her firm's new compensation arrangement, not in response to changes in clients' investment policy statements.

See also Standard VI(A)–Disclosure of Conflicts.

Example 8 (Investment Suitability):

Andre Shrub owns and operates Conduit, an investment advisory firm. Prior to opening Conduit, Shrub was an account manager with Elite Investment, a hedge fund managed by his good friend Adam Reed. To attract clients to a new Conduit fund, Shrub offers lower-than-normal management fees. He can do so because the fund consists of two top-performing funds managed by Reed. Given his personal friendship with Reed and the prior performance record of these two funds, Shrub believes this new fund is a winning combination for all parties. Clients quickly invest with Conduit to gain access to the Elite funds. No one is turned away because Conduit is seeking to expand its assets under management.

> *Comment*: Shrub has violated Standard III(C) because the risk profile of the new fund may not be suitable for every client. As an investment adviser, Shrub needs to establish an investment policy statement for each client and recommend only investments that match each client's risk and return profile in the IPS. Shrub is required to act as more than a simple sales agent for Elite.
>
> Although Shrub cannot disobey the direct request of a client to purchase a specific security, he should fully discuss the risks of a planned purchase and provide reasons why it might not be suitable for a client. This requirement may lead members and candidates to decline new customers if those customers' requested investment decisions are significantly out of line with their stated requirements.
>
> See also Standard V(A)–Diligence and Reasonable Basis.

Standard III(D) Performance Presentation

When communicating investment performance information, Members and Candidates must make reasonable efforts to ensure that it is fair, accurate, and complete.

Guidance

Standard III(D) requires members and candidates to provide credible performance information to clients and prospective clients and to avoid misstating performance or misleading clients and prospective clients about the investment performance of members or candidates or their firms. This standard encourages full disclosure of investment performance data to clients and prospective clients.

Standard III(D) covers any practice that would lead to misrepresentation of a member's or candidate's performance record, whether the practice involves performance presentation or performance measurement. This standard prohibits misrepresentations of past performance or reasonably expected performance. A member or candidate must give a fair and complete presentation of performance information whenever communicating data with respect to the performance history of individual accounts, composites or groups of accounts, or composites of an analyst's or firm's performance results. Furthermore, members and candidates should not state or imply that clients will obtain or benefit from a rate of return that was generated in the past.

The requirements of this standard are not limited to members and candidates managing separate accounts. Whenever a member or candidate provides performance information for which the manager is claiming responsibility, such as for pooled funds, the history must be accurate. Research analysts promoting the success or accuracy of their recommendations must ensure that their claims are fair, accurate, and complete.

If the presentation is brief, the member or candidate must make available to clients and prospects, on request, the detailed information supporting that communication. Best practice dictates that brief presentations include a reference to the limited nature of the information provided.

Recommended Procedures for Compliance

Apply the GIPS Standards

For members and candidates who are showing the performance history of the assets they manage, compliance with the GIPS standards is the best method to meet their obligations under Standard III(D). Members and candidates should encourage their firms to comply with the GIPS standards.

Compliance without Applying GIPS Standards

Members and candidates can also meet their obligations under Standard III(D) by

- considering the knowledge and sophistication of the audience to whom a performance presentation is addressed,
- presenting the performance of the weighted composite of similar portfolios rather than using a single representative account,
- including terminated accounts as part of performance history with a clear indication of when the accounts were terminated,
- including disclosures that fully explain the performance results being reported (for example, stating, when appropriate, that results are simulated when model results are used, clearly indicating when the performance record is that of a prior entity, or disclosing whether the performance is gross of fees, net of fees, or after tax), and
- maintaining the data and records used to calculate the performance being presented.

Application of the Standard

Example 1 (Performance Calculation and Length of Time):

Kyle Taylor of Taylor Trust Company, noting the performance of Taylor's common trust fund for the past two years, states in a brochure sent to his potential clients, "You can expect steady 25% annual compound growth of the value of your investments over the year." Taylor Trust's common trust fund did increase at the rate of 25% per year for the past year, which mirrored the increase of the entire market. The fund has never averaged that growth for more than one year, however, and the average rate of growth of all of its trust accounts for five years is 5% per year.

> *Comment*: Taylor's brochure is in violation of Standard III(D). Taylor should have disclosed that the 25% growth occurred only in one year. Additionally, Taylor did not include client accounts other than those in the firm's common trust fund. A general claim of firm performance should take into account the performance of all categories of accounts. Finally, by

stating that clients can expect a steady 25% annual compound growth rate, Taylor is also violating Standard I(C)–Misrepresentation, which prohibits assurances or guarantees regarding an investment.

Example 2 (Performance Calculation and Asset Weighting):

Anna Judd, a senior partner of Alexander Capital Management, circulates a performance report for the capital appreciation accounts for the years 1988 through 2004. The firm claims compliance with the GIPS standards. Returns are not calculated in accordance with the requirements of the GIPS standards, however, because the composites are not asset weighted.

> *Comment*: Judd is in violation of Standard III(D). When claiming compliance with the GIPS standards, firms must meet *all* of the requirements, make mandatory disclosures, and meet any other requirements that apply to that firm's specific situation. Judd's violation is not from any misuse of the data but from a false claim of GIPS compliance.

Example 3 (Performance Presentation and Prior Fund/Employer):

Aaron McCoy is vice president and managing partner of the equity investment group of Mastermind Financial Advisors, a new business. Mastermind recruited McCoy because he had a proven six-year track record with G&P Financial. In developing Mastermind's advertising and marketing campaign, McCoy prepares an advertisement that includes the equity investment performance he achieved at G&P Financial. The advertisement for Mastermind does not identify the equity performance as being earned while at G&P. The advertisement is distributed to existing clients and prospective clients of Mastermind.

> *Comment*: McCoy has violated Standard III(D) by distributing an advertisement that contains material misrepresentations about the historical performance of Mastermind. Standard III(D) requires that members and candidates make every reasonable effort to ensure that performance information is a fair, accurate, and complete representation of an individual's or firm's performance. As a general matter, this standard does not prohibit showing past performance of funds managed at a prior firm as part of a performance track record as long as showing that record is accompanied by appropriate disclosures about where the performance took place and the person's specific role in achieving that performance. If McCoy chooses to use his past performance from G&P in Mastermind's advertising, he should make full disclosure of the source of the historical performance.

Example 4 (Performance Presentation and Simulated Results):

Jed Davis has developed a mutual fund selection product based on historical information from the 1990–95 period. Davis tested his methodology by applying it retroactively to data from the 1996–2003 period, thus producing simulated performance results for those years. In January 2004, Davis's employer decided to offer the product and Davis began promoting it through trade journal advertisements and direct dissemination to clients. The advertisements included the performance results for the 1996–2003 period but did not indicate that the results were simulated.

> *Comment*: Davis violated Standard III(D) by failing to clearly identify simulated performance results. Standard III(D) prohibits members and candidates from making any statements that misrepresent the performance achieved by them or their firms and requires members and candidates

to make every reasonable effort to ensure that performance information presented to clients is fair, accurate, and complete. Use of simulated results should be accompanied by full disclosure as to the source of the performance data, including the fact that the results from 1995 through 2003 were the result of applying the model retroactively to that time period.

Example 5 (Performance Calculation and Selected Accounts Only):

In a presentation prepared for prospective clients, William Kilmer shows the rates of return realized over a five-year period by a "composite" of his firm's discretionary accounts that have a "balanced" objective. This composite, however, consisted of only a few of the accounts that met the balanced criterion set by the firm, excluded accounts under a certain asset level without disclosing the fact of their exclusion, and included accounts that did not have the balanced mandate because those accounts would boost the investment results. In addition, to achieve better results, Kilmer manipulated the narrow range of accounts included in the composite by changing the accounts that made up the composite over time.

Comment: Kilmer violated Standard III(D) by misrepresenting the facts in the promotional material sent to prospective clients, distorting his firm's performance record, and failing to include disclosures that would have clarified the presentation.

Example 6 (Performance Attribution Changes):

Art Purell is reviewing the quarterly performance attribution reports for distribution to clients. Purell works for an investment management firm with a bottom-up, fundamentals-driven investment process that seeks to add value through stock selection. The attribution methodology currently compares each stock with its sector. The attribution report indicates that the value added this quarter came from asset allocation and that stock selection contributed negatively to the calculated return.

Through running several different scenarios, Purell discovers that calculating attribution by comparing each stock with its industry and then rolling the effect to the sector level improves the appearance of the manager's stock selection activities. Because the firm defines the attribution terms and the results better reflect the stated strategy, Purell recommends that the client reports should use the revised methodology.

Comment: Modifying the attribution methodology without proper notifications to clients would fail to meet the requirements of Standard III(D). Purrell's recommendation is being done solely for the interest of the firm to improve its perceived ability to meet the stated investment strategy. Such changes are unfair to clients and obscure the facts regarding the firm's abilities.

Had Purell believed the new methodology offered improvements to the original model, then he would have needed to report the results of both calculations to the client. The report should also include the reasons why the new methodology is preferred, which would allow the client to make a meaningful comparison to prior results and provide a basis for comparing future attributions.

Example 7 (Performance Calculation Methodology Disclosure):

While developing a new reporting package for existing clients, Alisha Singh, a performance analyst, discovers that her company's new system automatically calculates both time-weighted and money-weighted returns. She asks the head of client services and retention which value would be preferred given that the firm has various investment strategies that include bonds, equities, securities without leverage, and alternatives. Singh is told not to label the return value so that the firm may show whichever value is greatest for the period.

> *Comment*: Following these instructions would lead to Singh violating Standard III(D). In reporting inconsistent return values, Singh would not be providing complete information to the firm's clients. Full information is provided when clients have sufficient information to judge the performance generated by the firm.

Example 8 (Performance Calculation Methodology Disclosure):

Richmond Equity Investors manages a long–short equity fund in which clients can trade once a week (on Fridays). For transparency reasons, a daily net asset value of the fund is calculated by Richmond. The monthly fact sheets of the fund report month-to-date and year-to-date performance. Richmond publishes the performance based on the higher of the last trading day of the month (typically, not the last business day) or the last business day of the month as determined by Richmond. The fact sheet mentions only that the data are as of the end of the month, without giving the exact date. Maggie Clark, the investment performance analyst in charge of the calculations, is concerned about the frequent changes and asks her supervisor whether they are appropriate.

> *Comment*: Clark's actions in questioning the changing performance metric comply with Standard III(D). She has shown concern that these changes are not presenting an accurate and complete picture of the performance generated.

Standard III(E) Preservation of Confidentiality

Members and Candidates must keep information about current, former, and prospective clients confidential unless:

1. The information concerns illegal activities on the part of the client;
2. Disclosure is required by law; or
3. The client or prospective client permits disclosure of the information.

Guidance

Highlights:

- *Status of Client*
- *Compliance with Laws*
- *Electronic Information and Security*
- *Professional Conduct Investigations by CFA Institute*

Standard III(E) requires that members and candidates preserve the confidentiality of information communicated to them by their clients, prospective clients, and former clients. This standard is applicable when (1) the member or candidate receives information because of his or her special ability to conduct a portion of the client's business or personal affairs and (2) the member or candidate receives information that arises from or is relevant to that portion of the client's business that is the subject of the special or confidential relationship. If disclosure of the information is required by law or the information concerns illegal activities by the client, however, the member or candidate may have an obligation to report the activities to the appropriate authorities.

Status of Client

This standard protects the confidentiality of client information even if the person or entity is no longer a client of the member or candidate. Therefore, members and candidates must continue to maintain the confidentiality of client records even after the client relationship has ended. If a client or former client expressly authorizes the member or candidate to disclose information, however, the member or candidate may follow the terms of the authorization and provide the information.

Compliance with Laws

As a general matter, members and candidates must comply with applicable law. If applicable law requires disclosure of client information in certain circumstances, members and candidates must comply with the law. Similarly, if applicable law requires members and candidates to maintain confidentiality, even if the information concerns illegal activities on the part of the client, members and candidates should not disclose such information. Additionally, applicable laws, such as inter-departmental communication restrictions within financial institutions, can impose limitations on information flow about a client within an entity that may lead to a violation of confidentiality. When in doubt, members and candidates should consult with their employer's compliance personnel or legal counsel before disclosing confidential information about clients.

Electronic Information and Security

Because of the ever-increasing volume of electronically stored information, members and candidates need to be particularly aware of possible accidental disclosures. Many employers have strict policies about how to electronically communicate sensitive client information and store client information on personal laptops, mobile devices, or portable disk/flash drives. In recent years, regulatory authorities have imposed stricter data security laws applying to the use of mobile remote digital communication, including the use of social media, that must be considered. Standard III(E) does not require members or candidates to become experts in information security technology, but they should have a thorough understanding of the policies of their employer. The size and operations of the firm will lead to differing policies for ensuring the security of confidential information maintained within the firm. Members and candidates should encourage their firm to conduct regular periodic training on confidentiality procedures for all firm personnel, including portfolio associates, receptionists, and other non-investment staff who have routine direct contact with clients and their records.

Professional Conduct Investigations by CFA Institute

The requirements of Standard III(E) are not intended to prevent members and candidates from cooperating with an investigation by the CFA Institute Professional Conduct Program (PCP). When permissible under applicable law, members and candidates shall consider the PCP an extension of themselves when requested to provide information about a client in support of a PCP investigation into their own conduct. Members and candidates are encouraged to cooperate with investigations into the conduct of others. Any information turned over to the PCP is kept in the strictest confidence. Members and candidates will not be considered in violation of this standard by forwarding confidential information to the PCP.

Recommended Procedures for Compliance

The simplest, most conservative, and most effective way to comply with Standard III(E) is to avoid disclosing any information received from a client except to authorized fellow employees who are also working for the client. In some instances, however, a member or candidate may want to disclose information received from clients that is outside the scope of the confidential relationship and does not involve illegal activities. Before making such a disclosure, a member or candidate should ask the following:

- In what context was the information disclosed? If disclosed in a discussion of work being performed for the client, is the information relevant to the work?
- Is the information background material that, if disclosed, will enable the member or candidate to improve service to the client?

Members and candidates need to understand and follow their firm's electronic information communication and storage procedures. If the firm does not have procedures in place, members and candidates should encourage the development of procedures that appropriately reflect the firm's size and business operations.

Communicating with Clients

Technological changes are constantly enhancing the methods that are used to communicate with clients and prospective clients. Members and candidates should make reasonable efforts to ensure that firm-supported communication methods and compliance procedures follow practices designed for preventing accidental distribution of confidential information. Given the rate at which technology changes, a regular review of privacy protection measures is encouraged.

Members and candidates should be diligent in discussing with clients the appropriate methods for providing confidential information. It is important to convey to clients that not all firm-sponsored resources may be appropriate for such communications.

Application of the Standard

Example 1 (Possessing Confidential Information):

Sarah Connor, a financial analyst employed by Johnson Investment Counselors, Inc., provides investment advice to the trustees of City Medical Center. The trustees have given her a number of internal reports concerning City Medical's needs for physical plant renovation and expansion. They have asked Connor to recommend investments that would generate capital appreciation in endowment funds to meet projected capital expenditures. Connor is approached by a local businessman, Thomas Kasey, who is considering a substantial contribution either to City Medical Center or to another local hospital. Kasey wants to find out the building plans of both institutions before making a decision, but he does not want to speak to the trustees.

> *Comment*: The trustees gave Connor the internal reports so she could advise them on how to manage their endowment funds. Because the information in the reports is clearly both confidential and within the scope of the confidential relationship, Standard III(E) requires that Connor refuse to divulge information to Kasey.

Example 2 (Disclosing Confidential Information):

Lynn Moody is an investment officer at the Lester Trust Company. She has an advisory customer who has talked to her about giving approximately US$50,000 to charity to reduce her income taxes. Moody is also treasurer of the Home for Indigent Widows (HIW), which is planning its annual giving campaign. HIW hopes to expand its list of prospects, particularly those capable of substantial gifts. Moody recommends that HIW's vice president for corporate gifts call on her customer and ask for a donation in the US$50,000 range.

> *Comment*: Even though the attempt to help the Home for Indigent Widows was well intended, Moody violated Standard III(E) by revealing confidential information about her client.

Example 3 (Disclosing Possible Illegal Activity):

Government officials approach Casey Samuel, the portfolio manager for Garcia Company's pension plan, to examine pension fund records. They tell her that Garcia's corporate tax returns are being audited and the pension fund is being reviewed. Two days earlier, Samuel had learned in a regular investment review with Garcia officers that potentially excessive and improper charges were being made to the pension plan by Garcia. Samuel consults her employer's general counsel and is advised that Garcia has probably violated tax and fiduciary regulations and laws.

> *Comment*: Samuel should inform her supervisor of these activities, and her employer should take steps, with Garcia, to remedy the violations. If that approach is not successful, Samuel and her employer should seek advice of legal counsel to determine the appropriate steps to be taken. Samuel may well have a duty to disclose the evidence she has of the continuing legal violations and to resign as asset manager for Garcia.

Example 4 (Disclosing Possible Illegal Activity):

David Bradford manages money for a family-owned real estate development corporation. He also manages the individual portfolios of several of the family members and officers of the corporation, including the chief financial officer (CFO). Based on the financial records of the corporation and some questionable practices of the CFO that Bradford has observed, Bradford believes that the CFO is embezzling money from the corporation and putting it into his personal investment account.

> *Comment*: Bradford should check with his firm's compliance department or appropriate legal counsel to determine whether applicable securities regulations require reporting the CFO's financial records.

Example 5 (Accidental Disclosure of Confidential Information):

Lynn Moody is an investment officer at the Lester Trust Company (LTC). She has stewardship of a significant number of individually managed taxable accounts. In addition to receiving quarterly written reports, about a dozen high-net-worth individuals have indicated to Moody a willingness to receive communications about overall economic and financial market outlooks directly from her by way of a social media platform. Under the direction of her firm's technology and compliance departments, she established a new group page on an existing social media platform specifically for her clients. In the instructions provided to clients, Moody asked them to "join" the group so they may be granted access to the posted content. The instructions also advised clients that all comments posted would be available to the public and thus the platform was not an appropriate method for communicating personal or confidential information.

Six months later, in early January, Moody posted LTC's year-end "Market Outlook." The report outlined a new asset allocation strategy that the firm is adding to its recommendations in the new year. Moody introduced the publication with a note informing her clients that she would be discussing the changes with them individually in their upcoming meetings.

One of Moody's clients responded directly on the group page that his family recently experienced a major change in their financial profile. The client described highly personal and confidential details of the event. Unfortunately, all clients that were part of the group were also able to read the detailed posting until Moody was able to have the comment removed.

> *Comment*: Moody has taken reasonable steps for protecting the confidentiality of client information while using the social media platform. She provided instructions clarifying that all information posted to the site would be publically viewable to all group members and warned against using this method for communicating confidential information. The accidental disclosure of confidential information by a client is not under Moody's control. Her actions to remove the information promptly once she became aware further align with Standard III(E).
>
> In understanding the potential sensitivity clients express surrounding the confidentiality of personal information, this event highlights a need for further training. Moody might advocate for additional warnings or controls for clients when they consider using social media platforms for two-way communications.

STANDARD IV: DUTIES TO EMPLOYERS

Standard IV(A) Loyalty

> In matters related to their employment, Members and Candidates must act for the benefit of their employer and not deprive their employer of the advantage of their skills and abilities, divulge confidential information, or otherwise cause harm to their employer.

Guidance

Highlights:

- *Employer Responsibilities*
- *Independent Practice*
- *Leaving an Employer*
- *Use of Social Media*
- *Whistleblowing*
- *Nature of Employment*

Standard IV(A) requires members and candidates to protect the interests of their firm by refraining from any conduct that would injure the firm, deprive it of profit, or deprive it of the member's or candidate's skills and ability. Members and candidates must always place the interests of clients above the interests of their employer but should also consider the effects of their conduct on the sustainability and integrity of the employer firm. In matters related to their employment, members and candidates must not engage in conduct that harms the interests of their employer. Implicit in this standard is the obligation of members and candidates to comply with the policies and procedures established by their employers that govern the employer–employee relationship—to the extent that such policies and procedures do not conflict with applicable laws, rules, or regulations or the Code and Standards.

This standard is not meant to be a blanket requirement to place employer interests ahead of personal interests in all matters. The standard does not require members and candidates to subordinate important personal and family obligations to their work. Members and candidates should enter into a dialogue with their employer about balancing personal and employment obligations when personal matters may interfere with their work on a regular or significant basis.

Employer Responsibilities

The employer–employee relationship imposes duties and responsibilities on both parties. Employers must recognize the duties and responsibilities that they owe to their employees if they expect to have content and productive employees.

Members and candidates are encouraged to provide their employer with a copy of the Code and Standards. These materials will inform the employer of the responsibilities of a CFA Institute member or a candidate in the CFA Program. The Code and Standards also serve as a basis for questioning employer policies and practices that conflict with these responsibilities.

Employers are not obligated to adhere to the Code and Standards. In expecting to retain competent employees who are members and candidates, however, they should not develop conflicting policies and procedures. The employer is responsible for a positive working environment, which includes an ethical workplace. Senior management has the additional responsibility to devise compensation structures and incentive arrangements that do not encourage unethical behavior.

Independent Practice

Included in Standard IV(A) is the requirement that members and candidates abstain from independent competitive activity that could conflict with the interests of their employer. Although Standard IV(A) does not preclude members or candidates from entering into an independent business while still employed, members and candidates who plan to engage in independent practice for compensation must notify their employer and describe the types of services they will render to prospective independent clients, the expected duration of the services, and the compensation for the services. Members and candidates should not render services until they receive consent from their employer to all of the terms of the arrangement. “Practice” means any service that the employer currently makes available for remuneration. “Undertaking independent practice” means engaging in competitive business, as opposed to making preparations to begin such practice.

Leaving an Employer

When members and candidates are planning to leave their current employer, they must continue to act in the employer’s best interest. They must not engage in any activities that would conflict with this duty until their resignation becomes effective. It is difficult to define specific guidelines for those members and candidates who are planning to compete with their employer as part of a new venture. The circumstances of each situation must be reviewed to distinguish permissible preparations from violations of duty. Activities that might constitute a violation, especially in combination, include the following:

- misappropriation of trade secrets,
- misuse of confidential information,
- solicitation of the employer's clients prior to cessation of employment,
- self-dealing (appropriating for one's own property a business opportunity or information belonging to one's employer), and
- misappropriation of clients or client lists.

A departing employee is generally free to make arrangements or preparations to go into a competitive business before terminating the relationship with his or her employer as long as such preparations do not breach the employee's duty of loyalty. A member or candidate who is contemplating seeking other employment must not contact existing clients or potential clients prior to leaving his or her employer for purposes of soliciting their business for the new employer. Once notice is provided to the employer of the intent to resign, the member or candidate must follow the employer's policies and procedures related to notifying clients of his or her planned departure. In addition, the member or candidate must not take records or files to a new employer without the written permission of the previous employer.

Once an employee has left the firm, the skills and experience that an employee obtained while employed are not "confidential" or "privileged" information. Similarly, simple knowledge of the names and existence of former clients is generally not confidential information unless deemed such by an agreement or by law. Standard IV(A) does not prohibit experience or knowledge gained at one employer from being used at another employer. Firm records or work performed on behalf of the firm that is stored in paper copy or electronically for the member's or candidate's convenience while employed, however, should be erased or returned to the employer unless the firm gives permission to keep those records after employment ends.

The standard does not prohibit former employees from contacting clients of their previous firm as long as the contact information does not come from the records of the former employer or violate an applicable "noncompete agreement." Members and candidates are free to use public information after departing to contact former clients without violating Standard IV(A) as long as there is no specific agreement not to do so.

Employers often require employees to sign noncompete agreements that preclude a departing employee from engaging in certain conduct. Members and candidates should take care to review the terms of any such agreement when leaving their employer to determine what, if any, conduct those agreements may prohibit.

In some markets, there are agreements between employers within an industry that outline information that departing employees are permitted to take upon resignation, such as the "Protocol for Broker Recruiting" in the United States. These agreements ease individuals' transition between firms that have agreed to follow the outlined procedures. Members and candidates who move between firms that sign such agreements may rely on the protections provided as long as they faithfully adhere to all the procedures outlined.

For example, under the agreement between many US brokers, individuals are allowed to take some general client contact information when departing. To be protected, a copy of the information the individual is taking must be provided to the local management team for review. Additionally, the specific client information may only be used by the departing employee and not others employed by the new firm.

Use of Social Media

The growth in various online networking platforms, such as LinkedIn, Twitter, and Facebook (commonly referred to as social media platforms), is providing new opportunities and challenges for businesses. Members and candidates should understand and abide by all applicable firm policies and regulations as to the acceptable use of social media platforms to interact with clients and prospective clients. This is especially important when a member or candidate is planning to leave an employer.

Social media use makes determining how and when departure notification is delivered to clients more complex. Members and candidates may have developed profiles on these platforms that include connections with individuals who are clients of the firm, as well as individuals unrelated to their employer. Communications through social media platforms that potentially reach current clients should adhere to the employer's policies and procedures regarding notification of departing employees.

Social media connections with clients are also raising questions concerning the differences between public information and firm property. Specific accounts and user profiles of members and candidates may be created for solely professional reasons, including firm-approved accounts for client engagements. Such firm-approved business-related accounts would be considered part of the firm's assets, thus requiring members and candidates to transfer or delete the accounts as directed by their firm's policies and procedures. Best practice for members and candidates is to maintain separate accounts for their personal and professional social media activities. Members and candidates should discuss with their employers how profiles should be treated when a single account includes personal connections and also is used to conduct aspects of their professional activities.

Whistleblowing

A member's or candidate's personal interests, as well as the interests of his or her employer, are secondary to protecting the integrity of capital markets and the interests of clients. Therefore, circumstances may arise (e.g., when an employer is engaged in illegal or unethical activity) in which members and candidates must act contrary to their employer's interests in order to comply with their duties to the market and clients. In such instances, activities that would normally violate a member's or candidate's duty to his or her employer (such as contradicting employer instructions, violating certain policies and procedures, or preserving a record by copying employer records) may be justified. Such action would be permitted only if the intent is clearly aimed at protecting clients or the integrity of the market, not for personal gain.

Nature of Employment

A wide variety of business relationships exists within the investment industry. For instance, a member or candidate may be an employee or an independent contractor. Members and candidates must determine whether they are employees or independent contractors in order to determine the applicability of Standard IV(A). This issue will be decided largely by the degree of control exercised by the employing entity over the member or candidate. Factors determining control include whether the member's or candidate's hours, work location, and other parameters of the job are set; whether facilities are provided to the member or candidate; whether the member's or candidate's expenses are reimbursed; whether the member or candidate seeks work from other employers; and the number of clients or employers the member or candidate works for.

A member's or candidate's duties within an independent contractor relationship are governed by the oral or written agreement between the member and the client. Members and candidates should take care to define clearly the scope of their responsibilities and the expectations of each client within the context of each relationship. Once a member or candidate establishes a relationship with a client, the member or candidate has a duty to abide by the terms of the agreement.

Recommended Procedures for Compliance

Employers may establish codes of conduct and operating procedures for their employees to follow. Members and candidates should fully understand the policies to ensure that they are not in conflict with the Code and Standards. The following topics identify policies that members and candidates should encourage their firms to adopt if the policies are not currently in place.

Competition Policy

A member or candidate must understand any restrictions placed by the employer on offering similar services outside the firm while employed by the firm. The policy may outline the procedures for requesting approval to undertake the outside service or may be a strict prohibition of such service. If a member's or candidate's employer elects to have its employees sign a noncompete agreement as part of the employment agreement, the member or candidate should ensure that the details are clear and fully explained prior to signing the agreement.

Termination Policy

Members and candidates should clearly understand the termination policies of their employer. Termination policies should establish clear procedures regarding the resignation process, including addressing how the termination will be disclosed to clients and staff and whether updates posted through social media platforms will be allowed. The firm's policy may also outline the procedures for transferring ongoing research and account management responsibilities. Finally, the procedures should address agreements that allow departing employees to remove specific client-related information upon resignation.

Incident-Reporting Procedures

Members and candidates should be aware of their firm's policies related to whistleblowing and encourage their firm to adopt industry best practices in this area. Many firms are required by regulatory mandates to establish confidential and anonymous reporting procedures that allow employees to report potentially unethical and illegal activities in the firm.

Employee Classification

Members and candidates should understand their status within their employer firm. Firms are encouraged to adopt a standardized classification structure (e.g., part time, full time, outside contractor) for their employees and indicate how each of the firm's policies applies to each employee class.

Application of the Standard

Example 1 (Soliciting Former Clients):

Samuel Magee manages pension accounts for Trust Assets, Inc., but has become frustrated with the working environment and has been offered a position with Fiduciary Management. Before resigning from Trust Assets, Magee asks four big accounts to leave that firm and open accounts with Fiduciary. Magee also persuades several prospective clients to sign agreements with Fiduciary Management. Magee had previously made presentations to these prospects on behalf of Trust Assets.

> *Comment*: Magee violated the employee–employer principle requiring him to act solely for his employer's benefit. Magee's duty is to Trust Assets as long as he is employed there. The solicitation of Trust Assets' current clients and prospective clients is unethical and violates Standard IV(A).

Example 2 (Former Employer's Documents and Files):

James Hightower has been employed by Jason Investment Management Corporation for 15 years. He began as an analyst but assumed increasing responsibilities and is now a senior portfolio manager and a member of the firm's investment policy committee. Hightower has decided to leave Jason Investment and start his own investment management business. He has been careful not to tell any of Jason's clients that he is leaving; he does not want to be accused of breaching his duty to Jason by soliciting Jason's clients before his departure. Hightower is planning to copy and take with him the following documents and information he developed or worked on while at Jason: (1) the client list, with addresses, telephone numbers, and other pertinent client information; (2) client account statements; (3) sample marketing presentations to prospective clients containing Jason's performance record; (4) Jason's recommended list of securities; (5) computer models to determine asset allocations for accounts with various objectives; (6) computer models for stock selection; and (7) personal computer spreadsheets for Hightower's major corporate recommendations, which he developed when he was an analyst.

> *Comment*: Except with the consent of their employer, departing members and candidates may not take employer property, which includes books, records, reports, and other materials, because taking such materials may interfere with their employer's business opportunities. Taking any employer records, even those the member or candidate prepared, violates Standard IV(A). Employer records include items stored in hard copy or any other medium (e.g., home computers, portable storage devices, cell phones).

Example 3 (Addressing Rumors):

Reuben Winston manages all-equity portfolios at Target Asset Management (TAM), a large, established investment counselor. Ten years previously, Philpott & Company, which manages a family of global bond mutual funds, acquired TAM in a diversification move. After the merger, the combined operations prospered in the fixed-income business but the equity management business at TAM languished. Lately, a few of the equity pension accounts that had been with TAM before the merger have terminated their relationships with TAM. One day, Winston finds on his voice mail the following message from a concerned client: "Hey! I just heard that Philpott is close to announcing the sale of your firm's equity management business to Rugged Life. What is going on?" Not being aware of any such deal, Winston and his associates are stunned. Their internal inquiries are met with denials from Philpott management, but the rumors persist. Feeling left in the dark, Winston contemplates leading an employee buyout of TAM's equity management business.

> *Comment*: An employee-led buyout of TAM's equity asset management business would be consistent with Standard IV(A) because it would rest on the permission of the employer and, ultimately, the clients. In this case,

> however, in which employees suspect the senior managers or principals are not truthful or forthcoming, Winston should consult legal counsel to determine appropriate action.

Example 4 (Ownership of Completed Prior Work):

Laura Clay, who is unemployed, wants part-time consulting work while seeking a full-time analyst position. During an interview at Bradley Associates, a large institutional asset manager, Clay is told that the firm has no immediate research openings but would be willing to pay her a flat fee to complete a study of the wireless communications industry within a given period of time. Clay would be allowed unlimited access to Bradley's research files and would be welcome to come to the offices and use whatever support facilities are available during normal working hours. Bradley's research director does not seek any exclusivity for Clay's output, and the two agree to the arrangement on a handshake. As Clay nears completion of the study, she is offered an analyst job in the research department of Winston & Company, a brokerage firm, and she is pondering submitting the draft of her wireless study for publication by Winston.

> *Comment*: Although she is under no written contractual obligation to Bradley, Clay has an obligation to let Bradley act on the output of her study before Winston & Company or Clay uses the information to their advantage. That is, unless Bradley gives permission to Clay and waives its rights to her wireless report, Clay would be in violation of Standard IV(A) if she were to immediately recommend to Winston the same transactions recommended in the report to Bradley. Furthermore, Clay must not take from Bradley any research file material or other property that she may have used.

Example 5 (Ownership of Completed Prior Work):

Emma Madeline, a recent college graduate and a candidate in the CFA Program, spends her summer as an unpaid intern at Murdoch and Lowell. The senior managers at Murdoch are attempting to bring the firm into compliance with the GIPS standards, and Madeline is assigned to assist in its efforts. Two months into her internship, Madeline applies for a job at McMillan & Company, which has plans to become GIPS compliant. Madeline accepts the job with McMillan. Before leaving Murdoch, she copies the firm's software that she helped develop because she believes this software will assist her in her new position.

> *Comment*: Even though Madeline does not receive monetary compensation for her services at Murdoch, she has used firm resources in creating the software and is considered an employee because she receives compensation and benefits in the form of work experience and knowledge. By copying the software, Madeline violated Standard IV(A) because she misappropriated Murdoch's property without permission.

Example 6 (Soliciting Former Clients):

Dennis Elliot has hired Sam Chisolm, who previously worked for a competing firm. Chisolm left his former firm after 18 years of employment. When Chisolm begins working for Elliot, he wants to contact his former clients because he knows them well and is certain that many will follow him to his new employer. Is Chisolm in violation of Standard IV(A) if he contacts his former clients?

> *Comment*: Because client records are the property of the firm, contacting former clients for any reason through the use of client lists or other information taken from a former employer without permission would be a

violation of Standard IV(A). In addition, the nature and extent of the contact with former clients may be governed by the terms of any noncompete agreement signed by the employee and the former employer that covers contact with former clients after employment.

Simple knowledge of the names and existence of former clients is not confidential information, just as skills or experience that an employee obtains while employed are not "confidential" or "privileged" information. The Code and Standards do not impose a prohibition on the use of experience or knowledge gained at one employer from being used at another employer. The Code and Standards also do not prohibit former employees from contacting clients of their previous firm, in the absence of a noncompete agreement. Members and candidates are free to use public information about their former firm after departing to contact former clients without violating Standard IV(A).

In the absence of a noncompete agreement, as long as Chisolm maintains his duty of loyalty to his employer before joining Elliot's firm, does not take steps to solicit clients until he has left his former firm, and does not use material from his former employer without its permission after he has left, he is not in violation of the Code and Standards.

Example 7 (Starting a New Firm):

Geraldine Allen currently works at a registered investment company as an equity analyst. Without notice to her employer, she registers with government authorities to start an investment company that will compete with her employer, but she does not actively seek clients. Does registration of this competing company with the appropriate regulatory authorities constitute a violation of Standard IV(A)?

Comment: Allen's preparation for the new business by registering with the regulatory authorities does not conflict with the work for her employer if the preparations have been done on Allen's own time outside the office and if Allen will not be soliciting clients for the business or otherwise operating the new company until she has left her current employer.

Example 8 (Competing with Current Employer):

Several employees are planning to depart their current employer within a few weeks and have been careful to not engage in any activities that would conflict with their duty to their current employer. They have just learned that one of their employer's clients has undertaken a request for proposal (RFP) to review and possibly hire a new investment consultant. The RFP has been sent to the employer and all of its competitors. The group believes that the new entity to be formed would be qualified to respond to the RFP and be eligible for the business. The RFP submission period is likely to conclude before the employees' resignations are effective. Is it permissible for the group of departing employees to respond to the RFP for their anticipated new firm?

Comment: A group of employees responding to an RFP that their employer is also responding to would lead to direct competition between the employees and the employer. Such conduct violates Standard IV(A) unless the group of employees receives permission from their employer as well as the entity sending out the RFP.

Example 9 (Externally Compensated Assignments):

Alfonso Mota is a research analyst with Tyson Investments. He works part time as a mayor for his hometown, a position for which he receives compensation. Must Mota seek permission from Tyson to serve as mayor?

> *Comment*: If Mota's mayoral duties are so extensive and time-consuming that they might detract from his ability to fulfill his responsibilities at Tyson, he should discuss his outside activities with his employer and come to a mutual agreement regarding how to manage his personal commitments with his responsibilities to his employer.

Example 10 (Soliciting Former Clients):

After leaving her employer, Shawna McQuillen establishes her own money management business. While with her former employer, she did not sign a noncompete agreement that would have prevented her from soliciting former clients. Upon her departure, she does not take any of her client lists or contact information and she clears her personal computer of any employer records, including client contact information. She obtains the phone numbers of her former clients through public records and contacts them to solicit their business.

> *Comment*: McQuillen is not in violation of Standard IV(A) because she has not used information or records from her former employer and is not prevented by an agreement with her former employer from soliciting her former clients.

Example 11 (Whistleblowing Actions):

Meredith Rasmussen works on a buy-side trading desk and concentrates on in-house trades for a hedge fund subsidiary managed by a team at the investment management firm. The hedge fund has been very successful and is marketed globally by the firm. From her experience as the trader for much of the activity of the fund, Rasmussen has become quite knowledgeable about the hedge fund's strategy, tactics, and performance. When a distinct break in the market occurs, however, and many of the securities involved in the hedge fund's strategy decline markedly in value, Rasmussen observes that the reported performance of the hedge fund does not reflect this decline. In her experience, the lack of any effect is a very unlikely occurrence. She approaches the head of trading about her concern and is told that she should not ask any questions and that the fund is big and successful and is not her concern. She is fairly sure something is not right, so she contacts the compliance officer, who also tells her to stay away from the issue of this hedge fund's reporting.

> *Comment*: Rasmussen has clearly come upon an error in policies, procedures, and compliance practices in the firm's operations. Having been unsuccessful in finding a resolution with her supervisor and the compliance officer, Rasmussen should consult the firm's whistleblowing policy to determine the appropriate next step toward informing management of her concerns. The potentially unethical actions of the investment management division are appropriate grounds for further disclosure, so Rasmussen's whistleblowing would not represent a violation of Standard IV(A).
>
> See also Standard I(D)–Misconduct and Standard IV(C)–Responsibilities of Supervisors.

Example 12 (Soliciting Former Clients):

Angel Crome has been a private banker for YBSafe Bank for the past eight years. She has been very successful and built a considerable client portfolio during that time but is extremely frustrated by the recent loss of reputation by her current employer and subsequent client insecurity. A locally renowned headhunter contacted Crome a few days ago and offered her an interesting job with a competing private bank. This bank offers a substantial signing bonus for advisers with their own client portfolios. Crome figures that she can solicit at least 70% of her clients to follow her and gladly enters into the new employment contract.

> *Comment*: Crome may contact former clients upon termination of her employment with YBSafe Bank, but she is prohibited from using client records built by and kept with her in her capacity as an employee of YBSafe Bank. Client lists are proprietary information of her former employer and must not be used for her or her new employer's benefit. The use of written, electronic, or any other form of records other than publicly available information to contact her former clients at YBSafe Bank will be a violation of Standard IV(A).

Example 13 (Notification of Code and Standards):

Krista Smith is a relatively new assistant trader for the fixed-income desk of a major investment bank. She is on a team responsible for structuring collateralized debt obligations (CDOs) made up of securities in the inventory of the trading desk. At a meeting of the team, senior executives explain the opportunity to eventually separate the CDO into various risk-rated tranches to be sold to the clients of the firm. After the senior executives leave the meeting, the head trader announces various responsibilities of each member of the team and then says, "This is a good time to unload some of the junk we have been stuck with for a while and disguise it with ratings and a thick, unreadable prospectus, so don't be shy in putting this CDO together. Just kidding." Smith is worried by this remark and asks some of her colleagues what the head trader meant. They all respond that he was just kidding but that there is some truth in the remark because the CDO is seen by management as an opportunity to improve the quality of the securities in the firm's inventory.

Concerned about the ethical environment of the workplace, Smith decides to talk to her supervisor about her concerns and provides the head trader with a copy of the Code and Standards. Smith discusses the principle of placing the client above the interest of the firm and the possibility that the development of the new CDO will not adhere to this responsibility. The head trader assures Smith that the appropriate analysis will be conducted when determining the appropriate securities for collateral. Furthermore, the ratings are assigned by an independent firm and the prospectus will include full and factual disclosures. Smith is reassured by the meeting, but she also reviews the company's procedures and requirements for reporting potential violations of company policy and securities laws.

> *Comment*: Smith's review of the company policies and procedures for reporting violations allows her to be prepared to report through the appropriate whistleblower process if she decides that the CDO development process involves unethical actions by others. Smith's actions comply with the Code and Standards principles of placing the client's interests first and being loyal to her employer. In providing her supervisor with a copy of the Code and Standards, Smith is highlighting the high level of ethical conduct she is required to adhere to in her professional activities.

Example 14 (Leaving an Employer):

Laura Webb just left her position as portfolio analyst at Research Systems, Inc. (RSI). Her employment contract included a non-solicitation agreement that requires her to wait two years before soliciting RSI clients for any investment-related services. Upon leaving, Webb was informed that RSI would contact clients immediately about her departure and introduce her replacement.

While working at RSI, Webb connected with clients, other industry associates, and friends through her LinkedIn network. Her business and personal relationships were intermingled because she considered many of her clients to be personal friends. Realizing that her LinkedIn network would be a valuable resource for new employment opportunities, she updated her profile several days following her departure from RSI. LinkedIn automatically sent a notification to Webb's entire network that her employment status had been changed in her profile.

> *Comment*: Prior to her departure, Webb should have discussed any client information contained in her social media networks. By updating her LinkedIn profile after RSI notified clients and after her employment ended, she has appropriately placed her employer's interests ahead of her own personal interests. In addition, she has not violated the non-solicitation agreement with RSI, unless it prohibited any contact with clients during the two-year period.

Example 15 (Confidential Firm Information):

Sanjay Gupta is a research analyst at Naram Investment Management (NIM). NIM uses a team-based research process to develop recommendations on investment opportunities covered by the team members. Gupta, like others, provides commentary for NIM's clients through the company blog, which is posted weekly on the NIM password-protected website. According to NIM's policy, every contribution to the website must be approved by the company's compliance department before posting. Any opinions expressed on the website are disclosed as representing the perspective of NIM.

Gupta also writes a personal blog to share his experiences with friends and family. As with most blogs, Gupta's personal blog is widely available to interested readers through various internet search engines. Occasionally, when he disagrees with the team-based research opinions of NIM, Gupta uses his personal blog to express his own opinions as a counterpoint to the commentary posted on the NIM website. Gupta believes this provides his readers with a more complete perspective on these investment opportunities.

> *Comment*: Gupta is in violation of Standard IV(A) for disclosing confidential firm information through his personal blog. The recommendations on the firm's blog to clients are not freely available across the internet, but his personal blog post indirectly provides the firm's recommendations.
>
> Additionally, by posting research commentary on his personal blog, Gupta is using firm resources for his personal advantage. To comply with Standard IV(A), members and candidates must receive consent from their employer prior to using company resources.

Standard IV(B) Additional Compensation Arrangements

Members and Candidates must not accept gifts, benefits, compensation, or consideration that competes with or might reasonably be expected to create a conflict of interest with their employer's interest unless they obtain written consent from all parties involved.

Guidance

Standard IV(B) requires members and candidates to obtain permission from their employer before accepting compensation or other benefits from third parties for the services rendered to the employer or for any services that might create a conflict with their employer's interest. Compensation and benefits include direct compensation by the client and any indirect compensation or other benefits received from third parties. "Written consent" includes any form of communication that can be documented (for example, communication via e-mail that can be retrieved and documented).

Members and candidates must obtain permission for additional compensation/benefits because such arrangements may affect loyalties and objectivity and create potential conflicts of interest. Disclosure allows an employer to consider the outside arrangements when evaluating the actions and motivations of members and candidates. Moreover, the employer is entitled to have full knowledge of all compensation/benefit arrangements so as to be able to assess the true cost of the services members or candidates are providing.

There may be instances in which a member or candidate is hired by an employer on a "part-time" basis. "Part-time" status applies to employees who do not commit the full number of hours required for a normal work week. Members and candidates should discuss possible limitations to their abilities to provide services that may be competitive with their employer during the negotiation and hiring process. The requirements of Standard IV(B) would be applicable to limitations identified at that time.

Recommended Procedures for Compliance

Members and candidates should make an immediate written report to their supervisor and compliance officer specifying any compensation they propose to receive for services in addition to the compensation or benefits received from their employer. The details of the report should be confirmed by the party offering the additional compensation, including performance incentives offered by clients. This written report should state the terms of any agreement under which a member or candidate will receive additional compensation; "terms" include the nature of the compensation, the approximate amount of compensation, and the duration of the agreement.

Application of the Standard

Example 1 (Notification of Client Bonus Compensation):

Geoff Whitman, a portfolio analyst for Adams Trust Company, manages the account of Carol Cochran, a client. Whitman is paid a salary by his employer, and Cochran pays the trust company a standard fee based on the market value of assets in her portfolio. Cochran proposes to Whitman that "any year that my portfolio achieves at least a 15% return before taxes, you and your wife can fly to Monaco at my expense and use my condominium during the third week of January." Whitman does not inform his employer of the arrangement and vacations in Monaco the following January as Cochran's guest.

> *Comment*: Whitman violated Standard IV(B) by failing to inform his employer in writing of this supplemental, contingent compensation arrangement. The nature of the arrangement could have resulted in partiality to Cochran's account, which could have detracted from Whitman's performance with respect to other accounts he handles for Adams Trust. Whitman must obtain the consent of his employer to accept such a supplemental benefit.

Example 2 (Notification of Outside Compensation):

Terry Jones sits on the board of directors of Exercise Unlimited, Inc. In return for his services on the board, Jones receives unlimited membership privileges for his family at all Exercise Unlimited facilities. Jones purchases Exercise Unlimited stock for the client accounts for which it is appropriate. Jones does not disclose this arrangement to his employer because he does not receive monetary compensation for his services to the board.

> *Comment*: Jones has violated Standard IV(B) by failing to disclose to his employer benefits received in exchange for his services on the board of directors. The nonmonetary compensation may create a conflict of interest in the same manner as being paid to serve as a director.

Example 3 (Prior Approval for Outside Compensation):

Jonathan Hollis is an analyst of oil-and-gas companies for Specialty Investment Management. He is currently recommending the purchase of ABC Oil Company shares and has published a long, well-thought-out research report to substantiate his recommendation. Several weeks after publishing the report, Hollis receives a call from the investor-relations office of ABC Oil saying that Thomas Andrews, CEO of the company, saw the report and really liked the analyst's grasp of the business and his company. The investor-relations officer invites Hollis to visit ABC Oil to discuss the industry further. ABC Oil offers to send a company plane to pick Hollis up and arrange for his accommodations while visiting. Hollis, after gaining the appropriate approvals, accepts the meeting with the CEO but declines the offered travel arrangements.

Several weeks later, Andrews and Hollis meet to discuss the oil business and Hollis's report. Following the meeting, Hollis joins Andrews and the investment relations officer for dinner at an upscale restaurant near ABC Oil's headquarters.

Upon returning to Specialty Investment Management, Hollis provides a full review of the meeting to the director of research, including a disclosure of the dinner attended.

> *Comment*: Hollis's actions did not violate Standard IV(B). Through gaining approval before accepting the meeting and declining the offered travel arrangements, Hollis sought to avoid any potential conflicts of interest between his company and ABC Oil. Because the location of the dinner was not available prior to arrival and Hollis notified his company of the dinner upon his return, accepting the dinner should not impair his objectivity. By disclosing the dinner, Hollis has enabled Specialty Investment Management to assess whether it has any impact on future reports and recommendations by Hollis related to ABC Oil.

Standard IV(C) Responsibilities of Supervisors

Members and Candidates must make reasonable efforts to ensure that anyone subject to their supervision or authority complies with applicable laws, rules, regulations, and the Code and Standards.

Guidance

Highlights:

- *System for Supervision*
- *Supervision Includes Detection*

Standard IV(C) states that members and candidates must promote actions by all employees under their supervision and authority to comply with applicable laws, rules, regulations, and firm policies and the Code and Standards.

Any investment professional who has employees subject to her or his control or influence—whether or not the employees are CFA Institute members, CFA charterholders, or candidates in the CFA Program—exercises supervisory responsibility. Members and candidates acting as supervisors must also have in-depth knowledge of the Code and Standards so that they can apply this knowledge in discharging their supervisory responsibilities.

The conduct that constitutes reasonable supervision in a particular case depends on the number of employees supervised and the work performed by those employees. Members and candidates with oversight responsibilities for large numbers of employees may not be able to personally evaluate the conduct of these employees on a continuing basis. These members and candidates may delegate supervisory duties to subordinates who directly oversee the other employees. A member's or candidate's responsibilities under Standard IV(C) include instructing those subordinates to whom supervision is delegated about methods to promote compliance, including preventing and detecting violations of laws, rules, regulations, firm policies, and the Code and Standards.

At a minimum, Standard IV(C) requires that members and candidates with supervisory responsibility make reasonable efforts to prevent and detect violations by ensuring the establishment of effective compliance systems. However, an effective compliance system goes beyond enacting a code of ethics, establishing policies and procedures to achieve compliance with the code and applicable law, and reviewing employee actions to determine whether they are following the rules.

To be effective supervisors, members and candidates should implement education and training programs on a recurring or regular basis for employees under their supervision. Such programs will assist the employees with meeting their professional obligations to practice in an ethical manner within the applicable legal system. Further, establishing incentives—monetary or otherwise—for employees not only to meet business goals but also to reward ethical behavior offers supervisors another way to assist employees in complying with their legal and ethical obligations.

Often, especially in large organizations, members and candidates may have supervisory responsibility but not the authority to establish or modify firm-wide compliance policies and procedures or incentive structures. Such limitations should not prevent a member or candidate from working with his or her own superiors and within the firm structure to develop and implement effective compliance tools, including but not limited to:

- a code of ethics,
- compliance policies and procedures,
- education and training programs,
- an incentive structure that rewards ethical conduct, and
- adoption of firm-wide best practice standards (e.g., the GIPS standards, the CFA Institute Asset Manager Code of Professional Conduct).

A member or candidate with supervisory responsibility should bring an inadequate compliance system to the attention of the firm's senior managers and recommend corrective action. If the member or candidate clearly cannot discharge supervisory responsibilities because of the absence of a compliance system or because of an inadequate compliance system, the member or candidate should decline in writing to accept supervisory responsibility until the firm adopts reasonable procedures to allow adequate exercise of supervisory responsibility.

System for Supervision

Members and candidates with supervisory responsibility must understand what constitutes an adequate compliance system for their firms and make reasonable efforts to see that appropriate compliance procedures are established, documented, communicated to covered personnel, and followed. "Adequate" procedures are those designed to meet industry standards, regulatory requirements, the requirements of the Code and Standards, and the circumstances of the firm. Once compliance procedures are established, the supervisor must also make reasonable efforts to ensure that the procedures are monitored and enforced.

To be effective, compliance procedures must be in place prior to the occurrence of a violation of the law or the Code and Standards. Although compliance procedures cannot be designed to anticipate every potential violation, they should be designed to anticipate the activities most likely to result in misconduct. Compliance programs must be appropriate for the size and nature of the organization. The member or candidate should review model compliance procedures or other industry programs to ensure that the firm's procedures meet the minimum industry standards.

Once a supervisor learns that an employee has violated or may have violated the law or the Code and Standards, the supervisor must promptly initiate an assessment to determine the extent of the wrongdoing. Relying on an employee's statements about the extent of the violation or assurances that the wrongdoing will not reoccur is not enough. Reporting the misconduct up the chain of command and warning the employee to cease the activity are also not enough. Pending the outcome of the investigation, a supervisor should take steps to ensure that the violation will not be repeated, such as placing limits on the employee's activities or increasing the monitoring of the employee's activities.

Supervision Includes Detection

Members and candidates with supervisory responsibility must also make reasonable efforts to detect violations of laws, rules, regulations, firm policies, and the Code and Standards. The supervisors exercise reasonable supervision by establishing and implementing written compliance procedures and ensuring that those procedures are followed through periodic review. If a member or candidate has adopted reasonable procedures and taken steps to institute an effective compliance program, then the member or candidate may not be in violation of Standard IV(C) if he or she does not detect violations that occur despite these efforts. The fact that violations do occur may indicate, however, that the compliance procedures are inadequate. In addition, in some cases, merely enacting such procedures may not be sufficient to fulfill the duty required by Standard IV(C). A member or candidate may be in violation of Standard IV(C) if he or she knows or should know that the procedures designed to promote compliance, including detecting and preventing violations, are not being followed.

Recommended Procedures for Compliance

Codes of Ethics or Compliance Procedures

Members and candidates are encouraged to recommend that their employers adopt a code of ethics. Adoption of a code of ethics is critical to establishing a strong ethical foundation for investment advisory firms and their employees. Codes of ethics formally emphasize and reinforce the client loyalty responsibilities of investment firm personnel, protect investing clients by deterring misconduct, and protect the firm's reputation for integrity.

There is a distinction, however, between codes of ethics and the specific policies and procedures needed to ensure compliance with the codes and with securities laws and regulations. Although both are important, codes of ethics should consist of fundamental, principle-based ethical and fiduciary concepts that are applicable to all of the firm's employees. In this way, firms can best convey to employees and clients the ethical ideals that investment advisers strive to achieve. These concepts need to be implemented, however, by detailed, firm-wide compliance policies and procedures. Compliance procedures assist the firm's personnel in fulfilling the responsibilities enumerated in the code of ethics and make probable that the ideals expressed in the code of ethics will be adhered to in the day-to-day operation of the firm.

Stand-alone codes of ethics should be written in plain language and should address general fiduciary concepts. They should be unencumbered by numerous detailed procedures. Codes presented in this way are the most effective in stressing to employees that they are in positions of trust and must act with integrity at all times. Mingling compliance procedures in the firm's code of ethics goes against the goal of reinforcing the ethical obligations of employees.

Separating the code of ethics from compliance procedures will also reduce, if not eliminate, the legal terminology and "boilerplate" language that can make the underlying ethical principles incomprehensible to the average person. Above all, to ensure the creation of a culture of ethics and integrity rather than one that merely focuses on following the rules, the principles in the code of ethics must be stated in a way that is accessible and understandable to everyone in the firm.

Members and candidates should encourage their employers to provide their codes of ethics to clients. In this case also, a simple, straightforward code of ethics will be best understood by clients. Unencumbered by the compliance procedures, the code of ethics will be effective in conveying that the firm is committed to conducting business in an ethical manner and in the best interests of the clients.

Adequate Compliance Procedures

A supervisor complies with Standard IV(C) by identifying situations in which legal violations or violations of the Code and Standards are likely to occur and by establishing and enforcing compliance procedures to prevent such violations. Adequate compliance procedures should

- be contained in a clearly written and accessible manual that is tailored to the firm's operations,
- be drafted so that the procedures are easy to understand,

- designate a compliance officer whose authority and responsibility are clearly defined and who has the necessary resources and authority to implement the firm's compliance procedures,
- describe the hierarchy of supervision and assign duties among supervisors,
- implement a system of checks and balances,
- outline the scope of the procedures,
- outline procedures to document the monitoring and testing of compliance procedures,
- outline permissible conduct, and
- delineate procedures for reporting violations and sanctions.

Once a compliance program is in place, a supervisor should

- disseminate the contents of the program to appropriate personnel,
- periodically update procedures to ensure that the measures are adequate under the law,
- continually educate personnel regarding the compliance procedures,
- issue periodic reminders of the procedures to appropriate personnel,
- incorporate a professional conduct evaluation as part of an employee's performance review,
- review the actions of employees to ensure compliance and identify violators, and
- take the necessary steps to enforce the procedures once a violation has occurred.

Once a violation is discovered, a supervisor should

- respond promptly,
- conduct a thorough investigation of the activities to determine the scope of the wrongdoing,
- increase supervision or place appropriate limitations on the wrongdoer pending the outcome of the investigation, and
- review procedures for potential changes necessary to prevent future violations from occurring.

Implementation of Compliance Education and Training

No amount of ethics education and awareness will deter someone determined to commit fraud for personal enrichment. But the vast majority of investment professionals strive to achieve personal success with dedicated service to their clients and employers.

Regular ethics and compliance training, in conjunction with adoption of a code of ethics, is critical to investment firms seeking to establish a strong culture of integrity and to provide an environment in which employees routinely engage in ethical conduct in compliance with the law. Training and education assist individuals in both recognizing areas that are prone to ethical and legal pitfalls and identifying those circumstances and influences that can impair ethical judgment.

By implementing educational programs, supervisors can train their subordinates to put into practice what the firm's code of ethics requires. Education helps employees make the link between legal and ethical conduct and the long-term success of the business; a strong culture of compliance signals to clients and potential clients that the firm has truly embraced ethical conduct as fundamental to the firm's mission to serve its clients.

Establish an Appropriate Incentive Structure

Even if individuals want to make the right choices and follow an ethical course of conduct and are aware of the obstacles that may trip them up, they can still be influenced to act improperly by a corporate culture that embraces a "succeed at all costs" mentality, stresses results regardless of the methods used to achieve those results, and does not reward ethical behavior. Supervisors can reinforce an individual's natural desire to "do the right thing" by building a culture of integrity in the workplace.

Supervisors and firms must look closely at their incentive structure to determine whether the structure encourages profits and returns at the expense of ethically appropriate conduct. Reward structures may turn a blind eye to how desired outcomes are achieved and encourage dysfunctional or counterproductive behavior. Only when compensation and incentives are firmly tied to client interests and *how* outcomes are achieved, rather than *how much* is generated for the firm, will employees work to achieve a culture of integrity.

Application of the Standard

Example 1 (Supervising Research Activities):

Jane Mattock, senior vice president and head of the research department of H&V, Inc., a regional brokerage firm, has decided to change her recommendation for Timber Products from buy to sell. In line with H&V's procedures, she orally advises certain other H&V executives of her proposed actions before the report is prepared for publication. As a result of Mattock's conversation with Dieter Frampton, one of the H&V executives accountable to Mattock, Frampton immediately sells Timber's stock from his own account and from certain discretionary client accounts. In addition, other personnel inform certain institutional customers of the changed recommendation before it is printed and disseminated to all H&V customers who have received previous Timber reports.

> *Comment*: Mattock has violated Standard IV(C) by failing to reasonably and adequately supervise the actions of those accountable to her. She did not prevent or establish reasonable procedures designed to prevent dissemination of or trading on the information by those who knew of her changed recommendation. She must ensure that her firm has procedures for reviewing or recording any trading in the stock of a corporation that has been the subject of an unpublished change in recommendation. Adequate procedures would have informed the subordinates of their duties and detected sales by Frampton and selected customers.

Example 2 (Supervising Research Activities):

Deion Miller is the research director for Jamestown Investment Programs. The portfolio managers have become critical of Miller and his staff because the Jamestown portfolios do not include any stock that has been the subject of a merger or tender offer. Georgia Ginn, a member of Miller's staff, tells Miller that she has been studying a local company, Excelsior, Inc., and recommends its purchase. Ginn adds that the company has been widely rumored to be the subject of a merger study by a well-known conglomerate and discussions between them are under way. At Miller's request, Ginn prepares a memo recommending the stock. Miller passes along Ginn's memo to the portfolio managers prior to leaving for vacation, and he notes that he has not reviewed the memo. As a result of the memo, the portfolio managers buy Excelsior stock immediately. The day Miller returns to the office, he learns that Ginn's only sources for the report were her brother, who is an acquisitions analyst with Acme Industries, the "well-known conglomerate," and that the merger discussions were planned but not held.

Comment: Miller violated Standard IV(C) by not exercising reasonable supervision when he disseminated the memo without checking to ensure that Ginn had a reasonable and adequate basis for her recommendations and that Ginn was not relying on material nonpublic information.

Example 3 (Supervising Trading Activities):

David Edwards, a trainee trader at Wheeler & Company, a major national brokerage firm, assists a customer in paying for the securities of Highland, Inc., by using anticipated profits from the immediate sale of the same securities. Despite the fact that Highland is not on Wheeler's recommended list, a large volume of its stock is traded through Wheeler in this manner. Roberta Ann Mason is a Wheeler vice president responsible for supervising compliance with the securities laws in the trading department. Part of her compensation from Wheeler is based on commission revenues from the trading department. Although she notices the increased trading activity, she does nothing to investigate or halt it.

Comment: Mason's failure to adequately review and investigate purchase orders in Highland stock executed by Edwards and her failure to supervise the trainee's activities violate Standard IV(C). Supervisors should be especially sensitive to actual or potential conflicts between their own self-interests and their supervisory responsibilities.

Example 4 (Supervising Trading Activities and Record Keeping):

Samantha Tabbing is senior vice president and portfolio manager for Crozet, Inc., a registered investment advisory and registered broker/dealer firm. She reports to Charles Henry, the president of Crozet. Crozet serves as the investment adviser and principal underwriter for ABC and XYZ public mutual funds. The two funds' prospectuses allow Crozet to trade financial futures for the funds for the limited purpose of hedging against market risks. Henry, extremely impressed by Tabbing's performance in the past two years, directs Tabbing to act as portfolio manager for the funds. For the benefit of its employees, Crozet has also organized the Crozet Employee Profit-Sharing Plan (CEPSP), a defined contribution retirement plan. Henry assigns Tabbing to manage 20% of the assets of CEPSP. Tabbing's investment objective for her portion of CEPSP's assets is aggressive growth. Unbeknownst to Henry, Tabbing frequently places S&P 500 Index purchase and sale orders for the funds and the CEPSP without providing the futures commission merchants (FCMs) who take the orders with any prior or simultaneous designation of the account for which the trade has been placed. Frequently, neither Tabbing nor anyone else at Crozet completes an internal trade ticket to record the time an order was placed or the specific account for which the order was intended. FCMs often designate a specific account only after the trade, when Tabbing provides such designation. Crozet has no written operating procedures or compliance manual concerning its futures trading, and its compliance department does not review such trading. After observing the market's movement, Tabbing assigns to CEPSP the S&P 500 positions with more favorable execution prices and assigns positions with less favorable execution prices to the funds.

Comment: Henry violated Standard IV(C) by failing to adequately supervise Tabbing with respect to her S&P 500 trading. Henry further violated Standard IV(C) by failing to establish record-keeping and reporting procedures to prevent or detect Tabbing's violations. Henry must make a reasonable effort to determine that adequate compliance procedures covering all employee trading activity are established, documented, communicated, and followed.

Example 5 (Accepting Responsibility):

Meredith Rasmussen works on a buy-side trading desk and concentrates on in-house trades for a hedge fund subsidiary managed by a team at the investment management firm. The hedge fund has been very successful and is marketed globally by the firm. From her experience as the trader for much of the activity of the fund, Rasmussen has become quite knowledgeable about the hedge fund's strategy, tactics, and performance. When a distinct break in the market occurs and many of the securities involved in the hedge fund's strategy decline markedly in value, however, Rasmussen observes that the reported performance of the hedge fund does not at all reflect this decline. From her experience, this lack of an effect is a very unlikely occurrence. She approaches the head of trading about her concern and is told that she should not ask any questions and that the fund is too big and successful and is not her concern. She is fairly sure something is not right, so she contacts the compliance officer and is again told to stay away from the hedge fund reporting issue.

> *Comment*: Rasmussen has clearly come upon an error in policies, procedures, and compliance practices within the firm's operations. According to Standard IV(C), the supervisor and the compliance officer have the responsibility to review the concerns brought forth by Rasmussen. Supervisors have the responsibility of establishing and encouraging an ethical culture in the firm. The dismissal of Rasmussen's question violates Standard IV(C) and undermines the firm's ethical operations.
>
> See also Standard I(D)–Misconduct and, for guidance on whistleblowing, Standard IV(A)–Loyalty.

Example 6 (Inadequate Procedures):

Brendan Witt, a former junior sell-side technology analyst, decided to return to school to earn an MBA. To keep his research skills and industry knowledge sharp, Witt accepted a position with On-line and Informed, an independent internet-based research company. The position requires the publication of a recommendation and report on a different company every month. Initially, Witt is a regular contributor of new research and a participant in the associated discussion boards that generally have positive comments on the technology sector. Over time, his ability to manage his educational requirements and his work requirements begin to conflict with one another. Knowing a recommendation is due the next day for On-line, Witt creates a report based on a few news articles and what the conventional wisdom of the markets has deemed the "hot" security of the day.

> *Comment*: Allowing the report submitted by Witt to be posted highlights a lack of compliance procedures by the research firm. Witt's supervisor needs to work with the management of On-line to develop an appropriate review process to ensure that all contracted analysts comply with the requirements.
>
> See also Standard V(A)–Diligence and Reasonable Basis because it relates to Witt's responsibility for substantiating a recommendation.

Example 7 (Inadequate Supervision):

Michael Papis is the chief investment officer of his state's retirement fund. The fund has always used outside advisers for the real estate allocation, and this information is clearly presented in all fund communications. Thomas Nagle, a recognized sell-side research analyst and Papis's business school classmate, recently left the investment bank he worked for to start his own asset management firm, Accessible Real Estate. Nagle is trying to build his assets under management and contacts Papis about gaining some of the retirement fund's allocation. In the previous few years, the performance of the retirement fund's real estate investments was in line with the fund's benchmark but was not extraordinary. Papis decides to help out his old friend and also to seek better returns by moving the real estate allocation to Accessible. The only notice of the change in adviser appears in the next annual report in the listing of associated advisers.

> *Comment*: Papis's actions highlight the need for supervision and review at all levels in an organization. His responsibilities may include the selection of external advisers, but the decision to change advisers appears arbitrary. Members and candidates should ensure that their firm has appropriate policies and procedures in place to detect inappropriate actions, such as the action taken by Papis.
>
> See also Standard V(A)–Diligence and Reasonable Basis, Standard V(B)–Communication with Clients and Prospective Clients, and Standard VI(A)–Disclosure of Conflicts.

Example 8 (Supervising Research Activities):

Mary Burdette was recently hired by Fundamental Investment Management (FIM) as a junior auto industry analyst. Burdette is expected to expand the social media presence of the firm because she is active with various networks, including Facebook, LinkedIn, and Twitter. Although Burdette's supervisor, Joe Graf, has never used social media, he encourages Burdette to explore opportunities to increase FIM's online presence and ability to share content, communicate, and broadcast information to clients. In response to Graf's encouragement, Burdette is working on a proposal detailing the advantages of getting FIM onto Twitter in addition to launching a company Facebook page.

As part of her auto industry research for FIM, Burdette is completing a report on the financial impact of Sun Drive Auto Ltd.'s new solar technology for compact automobiles. This research report will be her first for FIM, and she believes Sun Drive's technology could revolutionize the auto industry. In her excitement, Burdette sends a quick tweet to FIM Twitter followers summarizing her "buy" recommendation for Sun Drive Auto stock.

> *Comment*: Graf has violated Standard IV(C) by failing to reasonably supervise Burdette with respect to the contents of her tweet. He did not establish reasonable procedures to prevent the unauthorized dissemination of company research through social media networks. Graf must make sure all employees receive regular training about FIM's policies and procedures, including the appropriate business use of personal social media networks.
>
> See Standard III(B) for additional guidance.

Example 9 (Supervising Research Activities):

Chen Wang leads the research department at YYRA Retirement Planning Specialists. Chen supervises a team of 10 analysts in a fast-paced and understaffed organization. He is responsible for coordinating the firm's approved process to review all reports before they are provided to the portfolio management team for use in rebalancing client portfolios.

One of Chen's direct reports, Huang Mei, covers the banking industry. Chen must submit the latest updates to the portfolio management team tomorrow morning. Huang has yet to submit her research report on ZYX Bank because she is uncomfortable providing a "buy" or "sell" opinion of ZYX on the basis of the completed analysis. Pressed for time and concerned that Chen will reject a "hold" recommendation, she researches various websites and blogs on the banking sector for whatever she can find on ZYX. One independent blogger provides a new interpretation of the recently reported data Huang has analyzed and concludes with a strong "sell" recommendation for ZYX. She is impressed by the originality and resourcefulness of this blogger's report.

Very late in the evening, Huang submits her report and "sell" recommendation to Chen without any reference to the independent blogger's report. Given the late time of the submission and the competence of Huang's prior work, Chen compiles this report with the recommendations from each of the other analysts and meets with the portfolio managers to discuss implementation.

> *Comment*: Chen has violated Standard IV(C) by neglecting to reasonably and adequately follow the firm's approved review process for Huang's research report. The delayed submission and the quality of prior work do not remove Chen's requirement to uphold the designated review process. A member or candidate with supervisory responsibility must make reasonable efforts to see that appropriate procedures are established, documented, communicated to covered personnel, and followed.

STANDARD V: INVESTMENT ANALYSIS, RECOMMENDATIONS, AND ACTIONS

Standard V(A) Diligence and Reasonable Basis

Members and Candidates must:

1. Exercise diligence, independence, and thoroughness in analyzing investments, making investment recommendations, and taking investment actions.
2. Have a reasonable and adequate basis, supported by appropriate research and investigation, for any investment analysis, recommendation, or action.

Guidance

Highlights:

- *Defining Diligence and Reasonable Basis*
- *Using Secondary or Third-Party Research*
- *Using Quantitatively Oriented Research*
- *Developing Quantitatively Oriented Techniques*
- *Selecting External Advisers and Subadvisers*
- *Group Research and Decision Making*

The application of Standard V(A) depends on the investment philosophy the member, candidate, or firm is following, the role of the member or candidate in the investment decision-making process, and the support and resources provided by the member's or candidate's employer. These factors will dictate the nature of the diligence and thoroughness of the research and the level of investigation required by Standard V(A).

The requirements for issuing conclusions based on research will vary in relation to the member's or candidate's role in the investment decision-making process, but the member or candidate must make reasonable efforts to cover all pertinent issues when arriving at a recommendation. Members and candidates enhance transparency by providing or offering to provide supporting information to clients when recommending a purchase or sale or when changing a recommendation.

Defining Diligence and Reasonable Basis

Every investment decision is based on a set of facts known and understood at the time. Clients turn to members and candidates for advice and expect these advisers to have more information and knowledge than they do. This information and knowledge is the basis from which members and candidates apply their professional judgment in taking investment actions and making recommendations.

At a basic level, clients want assurance that members and candidates are putting forth the necessary effort to support the recommendations they are making. Communicating the level and thoroughness of the information reviewed before the member or candidate makes a judgment allows clients to understand the reasonableness of the recommended investment actions.

As with determining the suitability of an investment for the client, the necessary level of research and analysis will differ with the product, security, or service being offered. In providing an investment service, members and candidates typically use a variety of resources, including company reports, third-party research, and results from quantitative models. A reasonable basis is formed through a balance of these resources appropriate for the security or decision being analyzed.

The following list provides some, but definitely not all, examples of attributes to consider while forming the basis for a recommendation:

- global, regional, and country macroeconomic conditions,
- a company's operating and financial history,
- the industry's and sector's current conditions and the stage of the business cycle,
- a mutual fund's fee structure and management history,
- the output and potential limitations of quantitative models,
- the quality of the assets included in a securitization, and
- the appropriateness of selected peer-group comparisons.

Even though an investment recommendation may be well informed, downside risk remains for any investment. Members and candidates can base their decisions only on the information available at the time decisions are made. The steps taken in developing a diligent and reasonable recommendation should minimize unexpected downside events.

Using Secondary or Third-Party Research

If members and candidates rely on secondary or third-party research, they must make reasonable and diligent efforts to determine whether such research is sound. Secondary research is defined as research conducted by someone else in the member's or candidate's firm. Third-party research is research conducted by entities outside the member's or candidate's firm, such as a brokerage firm, bank, or research firm. If a member or candidate has reason to suspect that either secondary or third-party research or information comes from a source that lacks a sound basis, the member or candidate must not rely on that information.

Members and candidates should make reasonable enquiries into the source and accuracy of all data used in completing their investment analysis and recommendations. The sources of the information and data will influence the level of the review a member or candidate must undertake. Information and data taken from internet sources, such as personal blogs, independent research aggregation websites, or social media websites, likely require a greater level of review than information from more established research organizations.

Criteria that a member or candidate can use in forming an opinion on whether research is sound include the following:

- assumptions used,
- rigor of the analysis performed,
- date/timeliness of the research, and
- evaluation of the objectivity and independence of the recommendations.

A member or candidate may rely on others in his or her firm to determine whether secondary or third-party research is sound and use the information in good faith unless the member or candidate has reason to question its validity or the processes and procedures used by those responsible for the research. For example, a portfolio manager may not have a choice of a data source because the firm's senior managers conducted due diligence to determine which vendor would provide services; the member or candidate can use the information in good faith assuming the due diligence process was deemed adequate.

A member or candidate should verify that the firm has a policy about the timely and consistent review of approved research providers to ensure that the quality of the research continues to meet the necessary standards. If such a policy is not in place at the firm, the member or candidate should encourage the development and adoption of a formal review practice.

Using Quantitatively Oriented Research

Standard V(A) applies to the rapidly expanding use of quantitatively oriented research models and processes, such as computer-generated modeling, screening, and ranking of investment securities; the creation or valuation of derivative instruments; and quantitative portfolio construction techniques. These models and processes are being used for much more than the back testing of investment strategies, especially with continually advancing technology and techniques. The continued broad development of quantitative methods and models is an important part of capital market developments.

Members and candidates need to have an understanding of the parameters used in models and quantitative research that are incorporated into their investment recommendations. Although they are not required to become experts in every technical aspect of the models, they must understand the assumptions and limitations inherent in any model and how the results were used in the decision-making process.

The reliance on and potential limitations of financial models became clear through the investment crisis that unfolded in 2007 and 2008. In some cases, the financial models used to value specific securities and related derivative products did not adequately demonstrate the level of associated risks. Members and candidates should make reasonable efforts to test the output of investment models and other pre-programed analytical tools they use. Such validation should occur before incorporating the process into their methods, models, or analyses.

Although not every model can test for every factor or outcome, members and candidates should ensure that their analyses incorporate a broad range of assumptions sufficient to capture the underlying characteristics of investments. The omission from the analysis of potentially negative outcomes or of levels of risk outside the norm may misrepresent the true economic value of an investment. The possible scenarios for analysis should include factors that are likely to have a substantial influence on the investment value and may include extremely positive and negative scenarios.

Developing Quantitatively Oriented Techniques

Individuals who create new quantitative models and services must exhibit a higher level of diligence in reviewing new products than the individuals who ultimately use the analytical output. Members and candidates involved in the development and oversight of quantitatively oriented models, methods, and algorithms must understand the technical aspects of the products they provide to clients. A thorough testing of the model and resulting analysis should be completed prior to product distribution.

Members and candidates need to consider the source and time horizon of the data used as inputs in financial models. The information from many commercially available databases may not effectively incorporate both positive and negative market cycles. In the development of a recommendation, the member or candidate may need to test the models by using volatility and performance expectations that represent scenarios outside the observable databases. In reviewing the computer models or the resulting output, members and candidates need to pay particular attention to the assumptions used in the analysis and the rigor of the analysis to ensure that the model incorporates a wide range of possible input expectations, including negative market events.

Selecting External Advisers and Subadvisers

Financial instruments and asset allocation techniques continue to develop and evolve. This progression has led to the use of specialized managers to invest in specific asset classes or diversification strategies that complement a firm's in-house expertise. Standard V(A) applies to the level of review necessary in selecting an external adviser or subadviser to manage a specifically mandated allocation. Members and candidates must review managers as diligently as they review individual funds and securities.

Members and candidates who are directly involved with the use of external advisers need to ensure that their firms have standardized criteria for reviewing these selected external advisers and managers. Such criteria would include, but would not be limited to, the following:

- reviewing the adviser's established code of ethics,
- understanding the adviser's compliance and internal control procedures,
- assessing the quality of the published return information, and
- reviewing the adviser's investment process and adherence to its stated strategy.

Codes, standards, and guides to best practice published by CFA Institute provide members and candidates with examples of acceptable practices for external advisers and advice in selecting a new adviser. The following guides are available at the CFA Institute website (www.cfainstitute.org): Asset Manager Code of Professional Conduct, Global Investment Performance Standards, and Model Request for Proposal (for equity, credit, or real estate managers).

Group Research and Decision Making

Commonly, members and candidates are part of a group or team that is collectively responsible for producing investment analysis or research. The conclusions or recommendations of the group report represent the consensus of the group and are not necessarily the views of the member or candidate, even though the name of the member or candidate is included on the report. In some instances, a member or candidate will not agree with the view of the group. If, however, the member or candidate believes that the consensus opinion has a reasonable and adequate basis and is independent and objective, the member or candidate need not decline to be identified with the report. If the member or candidate is confident in the process, the member or candidate does not need to dissociate from the report even if it does not reflect his or her opinion.

Recommended Procedures for Compliance

Members and candidates should encourage their firms to consider the following policies and procedures to support the principles of Standard V(A):

- Establish a policy requiring that research reports, credit ratings, and investment recommendations have a basis that can be substantiated as reasonable and adequate. An individual employee (a supervisory analyst) or a group of employees (a review committee) should be appointed to review and approve such items prior to external circulation to determine whether the criteria established in the policy have been met.
- Develop detailed, written guidance for analysts (research, investment, or credit), supervisory analysts, and review committees that establishes the due diligence procedures for judging whether a particular recommendation has a reasonable and adequate basis.
- Develop measurable criteria for assessing the quality of research, the reasonableness and adequacy of the basis for any recommendation or rating, and the accuracy of recommendations over time. In some cases, firms may consider implementing compensation arrangements that depend on these measurable criteria and that are applied consistently to all related analysts.
- Develop detailed, written guidance that establishes minimum levels of scenario testing of all computer-based models used in developing, rating, and evaluating financial instruments. The policy should contain criteria related to the breadth of the scenarios tested, the accuracy of the output over time, and the analysis of cash flow sensitivity to inputs.
- Develop measurable criteria for assessing outside providers, including the quality of information being provided, the reasonableness and adequacy of the provider's collection practices, and the accuracy of the information over time. The established policy should outline how often the provider's products are reviewed.
- Adopt a standardized set of criteria for evaluating the adequacy of external advisers. The policy should include how often and on what basis the allocation of funds to the adviser will be reviewed.

Application of the Standard

Example 1 (Sufficient Due Diligence):

Helen Hawke manages the corporate finance department of Sarkozi Securities, Ltd. The firm is anticipating that the government will soon close a tax loophole that currently allows oil-and-gas exploration companies to pass on drilling expenses to holders of a certain class of shares. Because market demand for this tax-advantaged class of stock is currently high, Sarkozi convinces several companies to undertake new equity financings at once, before the loophole closes. Time is of the essence, but Sarkozi lacks sufficient resources to conduct adequate research on all the prospective issuing companies. Hawke decides to estimate the IPO prices on the basis of the relative size of each company and to justify the pricing later when her staff has time.

> *Comment*: Sarkozi should have taken on only the work that it could adequately handle. By categorizing the issuers by general size, Hawke has bypassed researching all the other relevant aspects that should be considered when pricing new issues and thus has not performed sufficient due diligence. Such an omission can result in investors purchasing shares at prices that have no actual basis. Hawke has violated Standard V(A).

Example 2 (Sufficient Scenario Testing):

Babu Dhaliwal works for Heinrich Brokerage in the corporate finance group. He has just persuaded Feggans Resources, Ltd., to allow his firm to do a secondary equity financing at Feggans Resources' current stock price. Because the stock has been trading at higher multiples than similar companies with equivalent production, Dhaliwal presses the Feggans Resources managers to project what would be the maximum production they could achieve in an optimal scenario. Based on these numbers, he is able to justify the price his firm will be asking for the secondary issue. During a sales pitch to the brokers, Dhaliwal then uses these numbers as the base-case production levels that Feggans Resources will achieve.

> *Comment*: When presenting information to the brokers, Dhaliwal should have given a range of production scenarios and the probability of Feggans Resources achieving each level. By giving the maximum production level as the likely level of production, he has misrepresented the chances of achieving that production level and seriously misled the brokers. Dhaliwal has violated Standard V(A).

Example 3 (Developing a Reasonable Basis):

Brendan Witt, a former junior sell-side technology analyst, decided to return to school to earn an MBA. To keep his research skills and industry knowledge sharp, Witt accepted a position with On-line and Informed, an independent internet-based research company. The position requires the publication of a recommendation and report on a different company every month. Initially, Witt is a regular contributor of new research and a participant in the associated discussion boards that generally have positive comments on the technology sector. Over time, his ability to manage his educational requirements and his work requirements begin to conflict with one another. Knowing a recommendation is due the next day for On-line, Witt creates a report based on a few news articles and what the conventional wisdom of the markets has deemed the "hot" security of the day.

> *Comment*: Witt's knowledge of and exuberance for technology stocks, a few news articles, and the conventional wisdom of the markets do not constitute, without more information, a reasonable and adequate basis for a stock recommendation that is supported by appropriate research and investigation. Therefore, Witt has violated Standard V(A).
>
> See also Standard IV(C)–Responsibilities of Supervisors because it relates to the firm's inadequate procedures.

Example 4 (Timely Client Updates):

Kristen Chandler is an investment consultant in the London office of Dalton Securities, a major global investment consultant firm. One of her UK pension funds has decided to appoint a specialist US equity manager. Dalton's global manager of research relies on local consultants to cover managers within their regions and, after conducting thorough due diligence, puts their views and ratings in Dalton's manager database. Chandler accesses Dalton's global manager research database and conducts a screen of all US equity managers on the basis of a match with the client's desired philosophy/style, performance, and tracking-error targets. She selects the five managers that meet these criteria and puts them in a briefing report that is delivered to the client 10 days later. Between the time of Chandler's database search and the delivery of the report to the client, Chandler is told that Dalton has updated the database with the information that one of the firms that Chandler has recommended for consideration lost its chief investment officer, the head of its US equity research, and the majority of its portfolio managers on the US equity product—all of whom have left to establish their own firm. Chandler does not revise her report with this updated information.

> *Comment*: Chandler has failed to satisfy the requirement of Standard V(A). Although Dalton updated the manager ratings to reflect the personnel turnover at one of the firms, Chandler did not update her report to reflect the new information.

Example 5 (Group Research Opinions):

Evelyn Mastakis is a junior analyst who has been asked by her firm to write a research report predicting the expected interest rate for residential mortgages over the next six months. Mastakis submits her report to the fixed-income investment committee of her firm for review, as required by firm procedures. Although some committee members support Mastakis's conclusion, the majority of the committee disagrees with her conclusion, and the report is significantly changed to indicate that interest rates are likely to increase more than originally predicted by Mastakis. Should Mastakis ask that her name be taken off the report when it is disseminated?

> *Comment*: The results of research are not always clear, and different people may have different opinions based on the same factual evidence. In this case, the committee may have valid reasons for issuing a report that differs from the analyst's original research. The firm can issue a report that is different from the original report of an analyst as long as there is a reasonable and adequate basis for its conclusions.
>
> Generally, analysts must write research reports that reflect their own opinion and can ask the firm not to put their name on reports that ultimately differ from that opinion. When the work is a group effort, however, not all members of the team may agree with all aspects of the report. Ultimately, members and candidates can ask to have their names removed from the report, but if they are satisfied that the process has produced results or conclusions that have a reasonable and adequate basis, members and candidates do not have to dissociate from the report even when they do not agree with its contents. If Mastakis is confident in the process, she does not need to dissociate from the report even if it does not reflect her opinion.

Example 6 (Reliance on Third-Party Research):

Gary McDermott runs a two-person investment management firm. McDermott's firm subscribes to a service from a large investment research firm that provides research reports. McDermott's firm makes investment recommendations on the basis of these reports.

> *Comment*: Members and candidates can rely on third-party research but must make reasonable and diligent efforts to determine that such research is sound. If McDermott undertakes due diligence efforts on a regular basis to ensure that the research produced by the large firm is objective and reasonably based, McDermott can rely on that research when making investment recommendations to clients.

Example 7 (Due Diligence in Submanager Selection):

Paul Ostrowski's business has grown significantly over the past couple of years, and some clients want to diversify internationally. Ostrowski decides to find a submanager to handle the expected international investments. Because this will be his first subadviser, Ostrowski uses the CFA Institute model "request for proposal" to design a questionnaire for his search. By his deadline, he receives seven completed questionnaires from a variety of domestic and international firms trying to gain his business. Ostrowski reviews all the applications in detail and decides to select the firm that charges the lowest fees because doing so will have the least impact on his firm's bottom line.

> *Comment*: The selection of an external adviser or subadviser should be based on a full and complete review of the adviser's services, performance history, and cost structure. In basing the decision on the fee structure alone, Ostrowski may be violating Standard V(A).
>
> See also Standard III(C)–Suitability because it relates to the ability of the selected adviser to meet the needs of the clients.

Example 8 (Sufficient Due Diligence):

Michael Papis is the chief investment officer of his state's retirement fund. The fund has always used outside advisers for the real estate allocation, and this information is clearly presented in all fund communications. Thomas Nagle, a recognized sell-side research analyst and Papis's business school classmate, recently left the investment bank he worked for to start his own asset management firm, Accessible Real Estate. Nagle is trying to build his assets under management and contacts Papis about gaining some of the retirement fund's allocation. In the previous few years, the performance of the retirement fund's real estate investments was in line with the fund's benchmark but was not extraordinary. Papis decides to help out his old friend and also to seek better returns by moving the real estate allocation to Accessible. The only notice of the change in adviser appears in the next annual report in the listing of associated advisers.

> *Comment*: Papis violated Standard V(A). His responsibilities may include the selection of the external advisers, but the decision to change advisers appears to have been arbitrary. If Papis was dissatisfied with the current real estate adviser, he should have conducted a proper solicitation to select the most appropriate adviser.
>
> See also Standard IV(C)–Responsibilities of Supervisors, Standard V(B)–Communication with Clients and Prospective Clients, and Standard VI(A)–Disclosure of Conflicts.

Example 9 (Sufficient Due Diligence):

Andre Shrub owns and operates Conduit, an investment advisory firm. Prior to opening Conduit, Shrub was an account manager with Elite Investment, a hedge fund managed by his good friend Adam Reed. To attract clients to a new Conduit fund, Shrub offers lower-than-normal management fees. He can do so because the fund consists of two top-performing funds managed by Reed. Given his personal friendship with Reed and the prior performance record of these two funds, Shrub believes this new fund is a winning combination for all parties. Clients quickly invest with Conduit to gain access to the Elite funds. No one is turned away because Conduit is seeking to expand its assets under management.

> *Comment*: Shrub violated Standard V(A) by not conducting a thorough analysis of the funds managed by Reed before developing the new Conduit fund. Shrub's reliance on his personal relationship with Reed and his prior knowledge of Elite are insufficient justification for the investments. The funds may be appropriately considered, but a full review of their operating procedures, reporting practices, and transparency are some elements of the necessary due diligence.

See also Standard III(C)–Suitability.

Example 10 (Sufficient Due Diligence):

Bob Thompson has been doing research for the portfolio manager of the fixed-income department. His assignment is to do sensitivity analysis on securitized subprime mortgages. He has discussed with the manager possible scenarios to use to calculate expected returns. A key assumption in such calculations is housing price appreciation (HPA) because it drives "prepays" (prepayments of mortgages) and losses. Thompson is concerned with the significant appreciation experienced over the previous five years as a result of the increased availability of funds from subprime mortgages. Thompson insists that the analysis should include a scenario run with −10% for Year 1, −5% for Year 2, and then (to project a worst-case scenario) 0% for Years 3 through 5. The manager replies that these assumptions are too dire because there has never been a time in their available database when HPA was negative.

Thompson conducts his research to better understand the risks inherent in these securities and evaluates these securities in the worst-case scenario, a less likely but possible environment. Based on the results of the enhanced scenarios, Thompson does not recommend the purchase of the securitization. Against the general market trends, the manager follows Thompson's recommendation and does not invest. The following year, the housing market collapses. In avoiding the subprime investments, the manager's portfolio outperforms its peer group that year.

> *Comment*: Thompson's actions in running the scenario test with inputs beyond the historical trends available in the firm's databases adhere to the principles of Standard V(A). His concerns over recent trends provide a sound basis for further analysis. Thompson understands the limitations of his model, when combined with the limited available historical information, to accurately predict the performance of the funds if market conditions change negatively.
>
> See also Standard I(B)–Independence and Objectivity.

Example 11 (Use of Quantitatively Oriented Models):

Espacia Liakos works in sales for Hellenica Securities, a firm specializing in developing intricate derivative strategies to profit from particular views on market expectations. One of her clients is Eugenie Carapalis, who has become convinced that commodity prices will become more volatile over the coming months. Carapalis asks Liakos to quickly engineer a strategy that will benefit from this expectation. Liakos turns to Hellenica's modeling group to fulfill this request. Because of the tight deadline, the modeling group outsources parts of the work to several trusted third parties. Liakos implements the disparate components of the strategy as the firms complete them.

Within a month, Carapalis is proven correct: Volatility across a range of commodities increases sharply. But her derivatives position with Hellenica returns huge losses, and the losses increase daily. Liakos investigates and realizes that although each of the various components of the strategy had been validated, they had never been evaluated as an integrated whole. In extreme conditions, portions of the model worked at cross-purposes with other portions, causing the overall strategy to fail dramatically.

> *Comment*: Liakos violated Standard V(A). Members and candidates must understand the statistical significance of the results of the models they recommend and must be able to explain them to clients. Liakos did not take adequate care to ensure a thorough review of the whole model; its components were evaluated only individually. Because Carapalis clearly

intended to implement the strategy as a whole rather than as separate parts, Liakos should have tested how the components of the strategy interacted as well as how they performed individually.

Example 12 (Successful Due Diligence/Failed Investment):

Alton Newbury is an investment adviser to high-net-worth clients. A client with an aggressive risk profile in his investment policy statement asks about investing in the Top Shelf hedge fund. This fund, based in Calgary, Alberta, Canada, has reported 20% returns for the first three years. The fund prospectus states that its strategy involves long and short positions in the energy sector and extensive leverage. Based on his analysis of the fund's track record, the principals involved in managing the fund, the fees charged, and the fund's risk profile, Newbury recommends the fund to the client and secures a position in it. The next week, the fund announces that it has suffered a loss of 60% of its value and is suspending operations and redemptions until after a regulatory review. Newbury's client calls him in a panic and asks for an explanation.

> *Comment*: Newbury's actions were consistent with Standard V(A). Analysis of an investment that results in a reasonable basis for recommendation does not guarantee that the investment has no downside risk. Newbury should discuss the analysis process with the client while reminding him or her that past performance does not lead to guaranteed future gains and that losses in an aggressive investment portfolio should be expected.

Example 13 (Quantitative Model Diligence):

Barry Cannon is the lead quantitative analyst at CityCenter Hedge Fund. He is responsible for the development, maintenance, and enhancement of the proprietary models the fund uses to manage its investors' assets. Cannon reads several high-level mathematical publications and blogs to stay informed of current developments. One blog, run by Expert CFA, presents some intriguing research that may benefit one of CityCenter's current models. Cannon is under pressure from firm executives to improve the model's predictive abilities, and he incorporates the factors discussed in the online research. The updated output recommends several new investments to the fund's portfolio managers.

> *Comment*: Cannon has violated Standard V(A) by failing to have a reasonable basis for the new recommendations made to the portfolio managers. He needed to diligently research the effect of incorporating the new factors before offering the output recommendations. Cannon may use the blog for ideas, but it is his responsibility to determine the effect on the firm's proprietary models.
>
> See Standard VII(B) regarding the violation by "Expert CFA" in the use of the CFA designation.

Example 14 (Selecting a Service Provider):

Ellen Smith is a performance analyst at Artic Global Advisors, a firm that manages global equity mandates for institutional clients. She was asked by her supervisor to review five new performance attribution systems and recommend one that would more appropriately explain the firm's investment strategy to clients. On the list was a system she recalled learning about when visiting an exhibitor booth at a recent conference. The system is highly quantitative and something of a "black box" in how it calculates the attribution values. Smith recommended this option without researching the others because the sheer complexity of the process was sure to impress the clients.

Comment: Smith's actions do not demonstrate a sufficient level of diligence in reviewing this product to make a recommendation for selecting the service. Besides not reviewing or considering the other four potential systems, she did not determine whether the "black box" attribution process aligns with the investment practices of the firm, including its investments in different countries and currencies. Smith must review and understand the process of any software or system before recommending its use as the firm's attribution system.

Example 15 (Subadviser Selection):

Craig Jackson is working for Adams Partners, Inc., and has been assigned to select a hedge fund subadviser to improve the diversification of the firm's large fund-of-funds product. The allocation must be in place before the start of the next quarter. Jackson uses a consultant database to find a list of suitable firms that claim compliance with the GIPS standards. He calls more than 20 firms on the list to confirm their potential interest and to determine their most recent quarterly and annual total return values. Because of the short turnaround, Jackson recommends the firm with the greatest total return values for selection.

Comment: By considering only performance and GIPS compliance, Jackson has not conducted sufficient review of potential firms to satisfy the requirements of Standard V(A). A thorough investigation of the firms and their operations should be conducted to ensure that their addition would increase the diversity of clients' portfolios and that they are suitable for the fund-of-funds product.

Example 16 (Manager Selection):

Timothy Green works for Peach Asset Management, where he creates proprietary models that analyze data from the firm request for proposal questionnaires to identify managers for possible inclusion in the firm's fund-of-funds investment platform. Various criteria must be met to be accepted to the platform. Because of the number of respondents to the questionnaires, Green uses only the data submitted to make a recommendation for adding a new manager.

Comment: By failing to conduct any additional outside review of the information to verify what was submitted through the request for proposal, Green has likely not satisfied the requirements of Standard V(A). The amount of information requested from outside managers varies among firms. Although the requested information may be comprehensive, Green should ensure sufficient effort is undertaken to verify the submitted information before recommending a firm for inclusion. This requires that he go beyond the information provided by the manager on the request for proposal questionnaire and may include interviews with interested managers, reviews of regulatory filings, and discussions with the managers' custodian or auditor.

Example 17 (Technical Model Requirements):

Jérôme Dupont works for the credit research group of XYZ Asset Management, where he is in charge of developing and updating credit risk models. In order to perform accurately, his models need to be regularly updated with the latest market data.

Dupont does not interact with or manage money for any of the firm's clients. He is in contact with the firm's US corporate bond fund manager, John Smith, who has only very superficial knowledge of the model and who from time to time asks very basic questions regarding the output recommendations. Smith does not consult Dupont with respect to finalizing his clients' investment strategies.

Dupont's recently assigned objective is to develop a new emerging market corporate credit risk model. The firm is planning to expand into emerging credit, and the development of such a model is a critical step in this process. Because Smith seems to follow the model's recommendations without much concern for its quality as he develops his clients' investment strategies, Dupont decides to focus his time on the development of the new emerging market model and neglects to update the US model.

After several months without regular updates, Dupont's diagnostic statistics start to show alarming signs with respect to the quality of the US credit model. Instead of conducting the long and complicated data update, Dupont introduces new codes into his model with some limited new data as a quick "fix." He thinks this change will address the issue without needing to complete the full data update, so he continues working on the new emerging market model.

Several months following the quick "fix," another set of diagnostic statistics reveals nonsensical results and Dupont realizes that his earlier change contained an error. He quickly corrects the error and alerts Smith. Smith realizes that some of the prior trades he performed were due to erroneous model results. Smith rebalances the portfolio to remove the securities purchased on the basis of the questionable results without reporting the issue to anyone else.

> *Comment*: Smith violated standard V(A) because exercising "diligence, independence, and thoroughness in analyzing investments, making investment recommendations, and taking investment actions" means that members and candidates must understand the technical aspects of the products they provide to clients. Smith does not understand the model he is relying on to manage money. Members and candidates should also make reasonable enquiries into the source and accuracy of all data used in completing their investment analysis and recommendations.
>
> Dupont violated V(A) even if he does not trade securities or make investment decisions. Dupont's models give investment recommendations, and Dupont is accountable for the quality of those recommendations. Members and candidates should make reasonable efforts to test the output of preprogramed analytical tools they use. Such validation should occur before incorporating the tools into their decision-making process.
>
> See also Standard V(B)–Communication with Clients and Prospective Clients.

Standard V(B) Communication with Clients and Prospective Clients

Members and Candidates must:

1 Disclose to clients and prospective clients the basic format and general principles of the investment processes they use to analyze investments, select securities, and construct portfolios and must promptly disclose any changes that might materially affect those processes.

2 Disclose to clients and prospective clients significant limitations and risks associated with the investment process.
3 Use reasonable judgment in identifying which factors are important to their investment analyses, recommendations, or actions and include those factors in communications with clients and prospective clients.
4 Distinguish between fact and opinion in the presentation of investment analyses and recommendations.

Guidance

Highlights:

- *Informing Clients of the Investment Process*
- *Different Forms of Communication*
- *Identifying Risk and Limitations*
- *Report Presentation*
- *Distinction between Facts and Opinions in Reports*

Standard V(B) addresses member and candidate conduct with respect to communicating with clients. Developing and maintaining clear, frequent, and thorough communication practices is critical to providing high-quality financial services to clients. When clients understand the information communicated to them, they also can understand exactly how members and candidates are acting on their behalf, which gives clients the opportunity to make well-informed decisions about their investments. Such understanding can be accomplished only through clear communication.

Standard V(B) states that members and candidates should communicate in a recommendation the factors that were instrumental in making the investment recommendation. A critical part of this requirement is to distinguish clearly between opinions and facts. In preparing a research report, the member or candidate must present the basic characteristics of the security(ies) being analyzed, which will allow the reader to evaluate the report and incorporate information the reader deems relevant to his or her investment decision-making process.

Similarly, in preparing a recommendation about, for example, an asset allocation strategy, alternative investment vehicle, or structured investment product, the member or candidate should include factors that are relevant to the asset classes that are being discussed. Follow-up communication of significant changes in the risk characteristics of a security or asset strategy is required. Providing regular updates to any changes in the risk characteristics is recommended.

Informing Clients of the Investment Process

Members and candidates must adequately describe to clients and prospective clients the manner in which they conduct the investment decision-making process. Such disclosure should address factors that have positive and negative influences on the recommendations, including significant risks and limitations of the investment process used. The member or candidate must keep clients and other interested parties informed on an ongoing basis about changes to the investment process, especially newly identified significant risks and limitations. Only by thoroughly understanding the nature of the investment product or service can a client determine whether changes to that product or service could materially affect his or her investment objectives.

Understanding the basic characteristics of an investment is of great importance in judging the suitability of that investment on a standalone basis, but it is especially important in determining the impact each investment will have on the characteristics of a portfolio. Although the risk and return characteristics of a common stock might seem to be essentially the same for any investor when the stock is viewed in isolation, the effects of those characteristics greatly depend on the other investments held. For instance, if the particular stock will represent 90% of an individual's investments, the stock's importance in the portfolio is vastly different from what it would be to an investor with a highly diversified portfolio for whom the stock will represent only 2% of the holdings.

A firm's investment policy may include the use of outside advisers to manage various portions of clients' assets under management. Members and candidates should inform the clients about the specialization or diversification expertise provided by the external adviser(s). This information allows clients to understand the full mix of products and strategies being applied that may affect their investment objectives.

Different Forms of Communication

For purposes of Standard V(B), communication is not confined to a written report of the type traditionally generated by an analyst researching a security, company, or industry. A presentation of information can be made via any means of communication, including in-person recommendation or description, telephone conversation, media broadcast, or transmission by computer (e.g., on the internet).

Computer and mobile device communications have rapidly evolved over the past few years. Members and candidates using any social media service to communicate business information must be diligent in their efforts to avoid unintended problems because these services may not be available to all clients. When providing information to clients through new technologies, members and candidates should take reasonable steps to ensure that such delivery would treat all clients fairly and, if necessary, be considered publicly disseminated.

The nature of client communications is highly diverse—from one word ("buy" or "sell") to in-depth reports of more than 100 pages. A communication may contain a general recommendation about the market, asset allocations, or classes of investments (e.g., stocks, bonds, real estate) or may relate to a specific security. If recommendations are contained in capsule form (such as a recommended stock list), members and candidates should notify clients that additional information and analyses are available from the producer of the report.

Identifying Risks and Limitations

Members and candidates must outline to clients and prospective clients significant risks and limitations of the analysis contained in their investment products or recommendations. The type and nature of significant risks will depend on the investment process that members and candidates are following and on the personal circumstances of the client. In general, the use of leverage constitutes a significant risk and should be disclosed.

Members and candidates must adequately disclose the general market-related risks and the risks associated with the use of complex financial instruments that are deemed significant. Other types of risks that members and candidates may consider disclosing include, but are not limited to, counterparty risk, country risk, sector or industry risk, security-specific risk, and credit risk.

Investment securities and vehicles may have limiting factors that influence a client's or potential client's investment decision. Members and candidates must report to clients and prospective clients the existence of limitations significant to the decision-making process. Examples of such factors and attributes include, but are not limited to, investment liquidity and capacity. Liquidity is the ability to liquidate an investment on a timely basis at a reasonable cost. Capacity is the investment amount beyond which returns will be negatively affected by new investments.

The appropriateness of risk disclosure should be assessed on the basis of what was known at the time the investment action was taken (often called an *ex ante* basis). Members and candidates must disclose significant risks known to them at the time of the disclosure. Members and candidates cannot be expected to disclose risks they are unaware of at the time recommendations or investment actions are made. In assessing compliance with Standard V(B), it is important to establish knowledge of a purported significant risk or limitation. A one-time investment loss that occurs after the disclosure does not constitute a pertinent factor in assessing whether significant risks and limitations were properly disclosed. Having no knowledge of a risk or limitation that subsequently triggers a loss may reveal a deficiency in the diligence and reasonable basis of the research of the member or candidate but may not reveal a breach of Standard V(B).

Report Presentation

Once the analytical process has been completed, the member or candidate who prepares the report must include those elements that are important to the analysis and conclusions of the report so that the reader can follow and challenge the report's reasoning. A report writer who has done adequate investigation may emphasize certain areas, touch briefly on others, and omit certain aspects deemed unimportant. For instance, a report may dwell on a quarterly earnings release or new-product introduction and omit other matters as long as the analyst clearly stipulates the limits to the scope of the report.

Investment advice based on quantitative research and analysis must be supported by readily available reference material and should be applied in a manner consistent with previously applied methodology. If changes in methodology are made, they should be highlighted.

Distinction between Facts and Opinions in Reports

Standard V(B) requires that opinion be separated from fact. Violations often occur when reports fail to separate the past from the future by not indicating that earnings estimates, changes in the outlook for dividends, or future market price information are *opinions* subject to future circumstances.

In the case of complex quantitative analyses, members and candidates must clearly separate fact from statistical conjecture and should identify the known limitations of an analysis. Members and candidates may violate Standard V(B) by failing to identify the limits of statistically developed projections because such omission leaves readers unaware of the limits of the published projections.

Members and candidates should explicitly discuss with clients and prospective clients the assumptions used in the investment models and processes to generate the analysis. Caution should be used in promoting the perceived accuracy of any model or process to clients because the ultimate output is merely an estimate of future results and not a certainty.

Recommended Procedures for Compliance

Because the selection of relevant factors is an analytical skill, determination of whether a member or candidate has used reasonable judgment in excluding and including information in research reports depends heavily on case-by-case review rather than a specific checklist.

Members and candidates should encourage their firms to have a rigorous methodology for reviewing research that is created for publication and dissemination to clients.

To assist in the after-the-fact review of a report, the member or candidate must maintain records indicating the nature of the research and should, if asked, be able to supply additional information to the client (or any user of the report) covering factors not included in the report.

Application of the Standard

Example 1 (Sufficient Disclosure of Investment System):

Sarah Williamson, director of marketing for Country Technicians, Inc., is convinced that she has found the perfect formula for increasing Country Technicians' income and diversifying its product base. Williamson plans to build on Country Technicians' reputation as a leading money manager by marketing an exclusive and expensive investment advice letter to high-net-worth individuals. One hitch in the plan is the complexity of Country Technicians' investment system—a combination of technical trading rules (based on historical price and volume fluctuations) and portfolio construction rules designed to minimize risk. To simplify the newsletter, she decides to include only each week's top five "buy" and "sell" recommendations and to leave out details of the valuation models and the portfolio structuring scheme.

> *Comment*: Williamson's plans for the newsletter violate Standard V(B). Williamson need not describe the investment system in detail in order to implement the advice effectively, but she must inform clients of Country Technicians' basic process and logic. Without understanding the basis for a recommendation, clients cannot possibly understand its limitations or its inherent risks.

Example 2 (Providing Opinions as Facts):

Richard Dox is a mining analyst for East Bank Securities. He has just finished his report on Boisy Bay Minerals. Included in his report is his own assessment of the geological extent of mineral reserves likely to be found on the company's land. Dox completed this calculation on the basis of the core samples from the company's latest drilling. According to Dox's calculations, the company has more than 500,000 ounces of gold on the property. Dox concludes his research report as follows: "Based on the fact that the company has 500,000 ounces of gold to be mined, I recommend a strong BUY."

> *Comment*: If Dox issues the report as written, he will violate Standard V(B). His calculation of the total gold reserves for the property based on the company's recent sample drilling is a quantitative opinion, not a fact. Opinion must be distinguished from fact in research reports.

Example 3 (Proper Description of a Security):

Olivia Thomas, an analyst at Government Brokers, Inc., which is a brokerage firm specializing in government bond trading, has produced a report that describes an investment strategy designed to benefit from an expected decline in US interest rates. The firm's derivative products group has designed a structured product that will allow the firm's clients to benefit from this strategy. Thomas's report describing the strategy indicates that high returns are possible if various scenarios for declining interest rates are assumed. Citing the proprietary nature of the structured product underlying the strategy, the report does not describe in detail how the firm is able to offer such returns or the related risks in the scenarios, nor does the report address the likely returns of the strategy if, contrary to expectations, interest rates rise.

Comment: Thomas has violated Standard V(B) because her report fails to describe properly the basic characteristics of the actual and implied risks of the investment strategy, including how the structure was created and the degree to which leverage was embedded in the structure. The report should include a balanced discussion of how the strategy would perform in the case of rising as well as falling interest rates, preferably illustrating how the strategies might be expected to perform in the event of a reasonable variety of interest rate and credit risk–spread scenarios. If liquidity issues are relevant with regard to the valuation of either the derivatives or the underlying securities, provisions the firm has made to address those risks should also be disclosed.

Example 4 (Notification of Fund Mandate Change):

May & Associates is an aggressive growth manager that has represented itself since its inception as a specialist at investing in small-cap US stocks. One of May's selection criteria is a maximum capitalization of US$250 million for any given company. After a string of successful years of superior performance relative to its peers, May has expanded its client base significantly, to the point at which assets under management now exceed US$3 billion. For liquidity purposes, May's chief investment officer (CIO) decides to lift the maximum permissible market-cap ceiling to US$500 million and change the firm's sales and marketing literature accordingly to inform prospective clients and third-party consultants.

Comment: Although May's CIO is correct about informing potentially interested parties as to the change in investment process, he must also notify May's existing clients. Among the latter group might be a number of clients who not only retained May as a small-cap manager but also retained mid-cap and large-cap specialists in a multiple-manager approach. Such clients could regard May's change of criteria as a style change that distorts their overall asset allocations.

Example 5 (Notification of Fund Mandate Change):

Rather than lifting the ceiling for its universe from US$250 million to US$500 million, May & Associates extends its small-cap universe to include a number of non-US companies.

Comment: Standard V(B) requires that May's CIO advise May's clients of this change because the firm may have been retained by some clients specifically for its prowess at investing in US small-cap stocks. Other changes that require client notification are introducing derivatives to emulate a certain market sector or relaxing various other constraints, such as portfolio beta. In all such cases, members and candidates must disclose changes to all interested parties.

Example 6 (Notification of Changes to the Investment Process):

RJZ Capital Management is an active value-style equity manager that selects stocks by using a combination of four multifactor models. The firm has found favorable results when back testing the most recent 10 years of available market data in a new dividend discount model (DDM) designed by the firm. This model is based on projected inflation rates, earnings growth rates, and interest rates. The president of RJZ decides to replace its simple model that uses price to trailing 12-month earnings with the new DDM.

Comment: Because the introduction of a new and different valuation model represents a material change in the investment process, RJZ's president must communicate the change to the firm's clients. RJZ is moving away from a model based on hard data toward a new model that is at least partly dependent on the firm's forecasting skills. Clients would likely view such a model as a significant change rather than a mere refinement of RJZ's process.

Example 7 (Notification of Changes to the Investment Process):

RJZ Capital Management loses the chief architect of its multifactor valuation system. Without informing its clients, the president of RJZ decides to redirect the firm's talents and resources toward developing a product for passive equity management—a product that will emulate the performance of a major market index.

Comment: By failing to disclose to clients a substantial change to its investment process, the president of RJZ has violated Standard V(B).

Example 8 (Notification of Changes to the Investment Process):

At Fundamental Asset Management, Inc., the responsibility for selecting stocks for addition to the firm's "approved" list has just shifted from individual security analysts to a committee consisting of the research director and three senior portfolio managers. Eleanor Morales, a portfolio manager with Fundamental Asset Management, thinks this change is not important enough to communicate to her clients.

Comment: Morales must disclose the process change to all her clients. Some of Fundamental's clients might be concerned about the morale and motivation among the firm's best research analysts after such a change. Moreover, clients might challenge the stock-picking track record of the portfolio managers and might even want to monitor the situation closely.

Example 9 (Sufficient Disclosure of Investment System):

Amanda Chinn is the investment director for Diversified Asset Management, which manages the endowment of a charitable organization. Because of recent staff departures, Diversified has decided to limit its direct investment focus to large-cap securities and supplement the needs for small-cap and mid-cap management by hiring outside fund managers. In describing the planned strategy change to the charity, Chinn's update letter states, "As investment director, I will directly oversee the investment team managing the endowment's large-capitalization allocation. I will coordinate the selection and ongoing review of external managers responsible for allocations to other classes." The letter also describes the reasons for the change and the characteristics external managers must have to be considered.

Comment: Standard V(B) requires the disclosure of the investment process used to construct the portfolio of the fund. Changing the investment process from managing all classes of investments within the firm to the use of external managers is one example of information that needs to be communicated to clients. Chinn and her firm have embraced the principles of Standard V(B) by providing their client with relevant information. The charity can now make a reasonable decision about whether Diversified Asset Management remains the appropriate manager for its fund.

Example 10 (Notification of Changes to the Investment Process):

Michael Papis is the chief investment officer of his state's retirement fund. The fund has always used outside advisers for the real estate allocation, and this information is clearly presented in all fund communications. Thomas Nagle, a recognized sell-side research analyst and Papis's business school classmate, recently left the investment bank he worked for to start his own asset management firm, Accessible Real Estate. Nagle is trying to build his assets under management and contacts Papis about gaining some of the retirement fund's allocation. In the previous few years, the performance of the retirement fund's real estate investments was in line with the fund's benchmark but was not extraordinary. Papis decides to help out his old friend and also to seek better returns by moving the real estate allocation to Accessible. The only notice of the change in adviser appears in the next annual report in the listing of associated advisers.

Comment: Papis has violated Standard V(B). He attempted to hide the nature of his decision to change external managers by making only a limited disclosure. The plan recipients and the fund's trustees need to be aware when changes are made to ensure that operational procedures are being followed.

See also Standard IV(C)–Responsibilities of Supervisors, Standard V(A)–Diligence and Reasonable Basis, and Standard VI(A)–Disclosure of Conflicts.

Example 11 (Notification of Errors):

Jérôme Dupont works for the credit research group of XYZ Asset Management, where he is in charge of developing and updating credit risk models. In order to perform accurately, his models need to be regularly updated with the latest market data.

Dupont does not interact with or manage money for any of the firm's clients. He is in contact with the firm's US corporate bond fund manager, John Smith, who has only very superficial knowledge of the model and who from time to time asks very basic questions regarding the output recommendations. Smith does not consult Dupont with respect to finalizing his clients' investment strategies.

Dupont's recently assigned objective is to develop a new emerging market corporate credit risk model. The firm is planning to expand into emerging credit, and the development of such a model is a critical step in this process. Because Smith seems to follow the model's recommendations without much concern for its quality as he develops his clients' investment strategies, Dupont decides to focus his time on the development of the new emerging market model and neglects to update the US model.

After several months without regular updates, Dupont's diagnostic statistics start to show alarming signs with respect to the quality of the US credit model. Instead of conducting the long and complicated data update, Dupont introduces new codes into his model with some limited new data as a quick "fix." He thinks this change will address the issue without needing to complete the full data update, so he continues working on the new emerging market model.

Several months following the quick "fix," another set of diagnostic statistics reveals nonsensical results and Dupont realizes that his earlier change contained an error. He quickly corrects the error and alerts Smith. Smith realizes that some of the prior trades he performed were due to erroneous model results. Smith rebalances the portfolio to remove the securities purchased on the basis of the questionable results without reporting the issue to anyone else.

Comment: Smith violated V(B) by not disclosing a material error in the investment process. Clients should have been informed about the error and the corrective actions the firm was undertaking on their behalf.

See also Standard V(A)–Diligence and Reasonable Basis.

Example 12 (Notification of Risks and Limitations):

Quantitative analyst Yuri Yakovlev has developed an investment strategy that selects small-cap stocks on the basis of quantitative signals. Yakovlev's strategy typically identifies only a small number of stocks (10–20) that tend to be illiquid, but according to his backtests, the strategy generates significant risk-adjusted returns. The partners at Yakovlev's firm, QSC Capital, are impressed by these results. After a thorough examination of the strategy's risks, stress testing, historical back testing, and scenario analysis, QSC decides to seed the strategy with US$10 million of internal capital in order for Yakovlev to create a track record for the strategy.

After two years, the strategy has generated performance returns greater than the appropriate benchmark and the Sharpe ratio of the fund is close to 1.0. On the basis of these results, QSC decides to actively market the fund to large institutional investors. While creating the offering materials, Yakovlev informs the marketing team that the capacity of the strategy is limited. The extent of the limitation is difficult to ascertain with precision; it depends on market liquidity and other factors in his model that can evolve over time. Yakovlev indicates that given the current market conditions, investments in the fund beyond US$100 million of capital could become more difficult and negatively affect expected fund returns.

Alan Wellard, the manager of the marketing team, is a partner with 30 years of marketing experience and explains to Yakovlev that these are complex technical issues that will muddy the marketing message. According to Wellard, the offering material should focus solely on the great track record of the fund. Yakovlev does not object because the fund has only US$12 million of capital, very far from the US$100 million threshold.

> *Comment*: Yakovlev and Wellard have not appropriately disclosed a significant limitation associated with the investment product. Yakovlev believes this limitation, once reached, will materially affect the returns of the fund. Although the fund is currently far from the US$100 million mark, current and prospective investors must be made aware of this capacity issue. If significant limitations are complicated to grasp and clients do not have the technical background required to understand them, Yakovlev and Wellard should either educate the clients or ascertain whether the fund is suitable for each client.

Example 13 (Notification of Risks and Limitations):

Brickell Advisers offers investment advisory services mainly to South American clients. Julietta Ramon, a risk analyst at Brickell, describes to clients how the firm uses value at risk (VaR) analysis to track the risk of its strategies. Ramon assures clients that calculating a VaR at a 99% confidence level, using a 20-day holding period, and applying a methodology based on an *ex ante* Monte Carlo simulation is extremely effective. The firm has never had losses greater than those predicted by this VaR analysis.

> *Comment*: Ramon has not sufficiently communicated the risks associated with the investment process to satisfy the requirements of Standard V(B). The losses predicted by a VaR analysis depend greatly on the inputs used in the model. The size and probability of losses can differ significantly from what an individual model predicts. Ramon must disclose how the inputs were selected and the potential limitations and risks associated with the investment strategy.

Example 14 (Notification of Risks and Limitations):

Lily Smith attended an industry conference and noticed that John Baker, an investment manager with Baker Associates, attracted a great deal of attention from the conference participants. On the basis of her knowledge of Baker's reputation and the interest he received at the conference, Smith recommends adding Baker Associates to the approved manager platform. Her recommendation to the approval committee included the statement "John Baker is well respected in the industry, and his insights are consistently sought after by investors. Our clients are sure to benefit from investing with Baker Associates."

> *Comment*: Smith is not appropriately separating facts from opinions in her recommendation to include the manager within the platform. Her actions conflict with the requirements of Standard V(B). Smith is relying on her opinions about Baker's reputation and the fact that many attendees were talking with him at the conference. Smith should also review the requirements of Standard V(A) regarding reasonable basis to determine the level of review necessary to recommend Baker Associates.

Standard V(C) Record Retention

> Members and Candidates must develop and maintain appropriate records to support their investment analyses, recommendations, actions, and other investment-related communications with clients and prospective clients.

Guidance

Highlights:

- *New Media Records*
- *Records Are Property of the Firm*
- *Local Requirements*

Members and candidates must retain records that substantiate the scope of their research and reasons for their actions or conclusions. The retention requirement applies to decisions to buy or sell a security as well as reviews undertaken that do not lead to a change in position. Which records are required to support recommendations or investment actions depends on the role of the member or candidate in the investment decision-making process. Records may be maintained either in hard copy or electronic form.

Some examples of supporting documentation that assists the member or candidate in meeting the requirements for retention are as follows:

- personal notes from meetings with the covered company,
- press releases or presentations issued by the covered company,
- computer-based model outputs and analyses,
- computer-based model input parameters,
- risk analyses of securities' impacts on a portfolio,
- selection criteria for external advisers,

- notes from clients from meetings to review investment policy statements, and
- outside research reports.

New Media Records

The increased use of new and evolving technological formats (e.g., social media) for gathering and sharing information creates new challenges in maintaining the appropriate records and files. The nature or format of the information does not remove a member's or candidate's responsibility to maintain a record of information used in his or her analysis or communicated to clients.

Members and candidates should understand that although employers and local regulators are developing digital media retention policies, these policies may lag behind the advent of new communication channels. Such lag places greater responsibility on the individual for ensuring that all relevant information is retained. Examples of non-print media formats that should be retained include, but are not limited to,

- e-mails,
- text messages,
- blog posts, and
- Twitter posts.

Records Are Property of the Firm

As a general matter, records created as part of a member's or candidate's professional activity on behalf of his or her employer are the property of the firm. When a member or candidate leaves a firm to seek other employment, the member or candidate cannot take the property of the firm, including original forms or copies of supporting records of the member's or candidate's work, to the new employer without the express consent of the previous employer. The member or candidate cannot use historical recommendations or research reports created at the previous firm because the supporting documentation is unavailable. For future use, the member or candidate must re-create the supporting records at the new firm with information gathered through public sources or directly from the covered company and not from memory or sources obtained at the previous employer.

Local Requirements

Local regulators often impose requirements on members, candidates, and their firms related to record retention that must be followed. Firms may also implement policies detailing the applicable time frame for retaining research and client communication records. Fulfilling such regulatory and firm requirements satisfies the requirements of Standard V(C). In the absence of regulatory guidance or firm policies, CFA Institute recommends maintaining records for at least seven years.

Recommended Procedures for Compliance

The responsibility to maintain records that support investment action generally falls with the firm rather than individuals. Members and candidates must, however, archive research notes and other documents, either electronically or in hard copy, that support their current investment-related communications. Doing so will assist their firms in complying with requirements for preservation of internal or external records.

Application of the Standard

Example 1 (Record Retention and IPS Objectives and Recommendations):

One of Nikolas Lindstrom's clients is upset by the negative investment returns of his equity portfolio. The investment policy statement for the client requires that the portfolio manager follow a benchmark-oriented approach. The benchmark for the client includes a 35% investment allocation in the technology sector. The client acknowledges that this allocation was appropriate, but over the past three years, technology stocks have suffered severe losses. The client complains to the investment manager for allocating so much money to this sector.

> *Comment*: For Lindstrom, having appropriate records is important to show that over the past three years, the portion of technology stocks in the benchmark index was 35%, as called for in the IPS. Lindstrom should also have the client's IPS stating that the benchmark was appropriate for the client's investment objectives. He should also have records indicating that the investment has been explained appropriately to the client and that the IPS was updated on a regular basis. Taking these actions, Lindstrom would be in compliance with Standard V(C).

Example 2 (Record Retention and Research Process):

Malcolm Young is a research analyst who writes numerous reports rating companies in the luxury retail industry. His reports are based on a variety of sources, including interviews with company managers, manufacturers, and economists; on-site company visits; customer surveys; and secondary research from analysts covering related industries.

> *Comment*: Young must carefully document and keep copies of all the information that goes into his reports, including the secondary or third-party research of other analysts. Failure to maintain such files would violate Standard V(C).

Example 3 (Records as Firm, Not Employee, Property):

Martin Blank develops an analytical model while he is employed by Green Partners Investment Management, LLP (GPIM). While at the firm, he systematically documents the assumptions that make up the model as well as his reasoning behind the assumptions. As a result of the success of his model, Blank is hired to be the head of the research department of one of GPIM's competitors. Blank takes copies of the records supporting his model to his new firm.

> *Comment*: The records created by Blank supporting the research model he developed at GPIM are the records of GPIM. Taking the documents with him to his new employer without GPIM's permission violates Standard V(C). To use the model in the future, Blank must re-create the records supporting his model at the new firm.

STANDARD VI: CONFLICTS OF INTEREST

Standard VI(A) Disclosure of Conflicts

Members and Candidates must make full and fair disclosure of all matters that could reasonably be expected to impair their independence and objectivity or interfere with respective duties to their clients, prospective clients, and employer. Members and Candidates must ensure that such disclosures are prominent, are delivered in plain language, and communicate the relevant information effectively.

Guidance

Highlights:

- *Disclosure of Conflicts to Employers*
- *Disclosure to Clients*
- *Cross-Departmental Conflicts*
- *Conflicts with Stock Ownership*
- *Conflicts as a Director*

Best practice is to avoid actual conflicts or the appearance of conflicts of interest when possible. Conflicts of interest often arise in the investment profession. Conflicts can occur between the interests of clients, the interests of employers, and the member's or candidate's own personal interests. Common sources for conflict are compensation structures, especially incentive and bonus structures that provide immediate returns for members and candidates with little or no consideration of long-term value creation.

Identifying and managing these conflicts is a critical part of working in the investment industry and can take many forms. When conflicts cannot be reasonably avoided, clear and complete disclosure of their existence is necessary.

Standard VI(A) protects investors and employers by requiring members and candidates to fully disclose to clients, potential clients, and employers all actual and potential conflicts of interest. Once a member or candidate has made full disclosure, the member's or candidate's employer, clients, and prospective clients will have the information needed to evaluate the objectivity of the investment advice or action taken on their behalf.

To be effective, disclosures must be prominent and must be made in plain language and in a manner designed to effectively communicate the information. Members and candidates have the responsibility of determining how often, in what manner, and in what particular circumstances the disclosure of conflicts must be made. Best practices dictate updating disclosures when the nature of a conflict of interest changes materially—for example, if the nature of a conflict of interest worsens through the introduction of bonuses based on each quarter's profits as to opposed annual profits. In making and updating disclosures of conflicts of interest, members and candidates should err on the side of caution to ensure that conflicts are effectively communicated.

Disclosure of Conflicts to Employers

Disclosure of conflicts to employers may be appropriate in many instances. When reporting conflicts of interest to employers, members and candidates must give their employers enough information to assess the impact of the conflict. By complying with employer guidelines, members and candidates allow their employers to avoid potentially embarrassing and costly ethical or regulatory violations.

Reportable situations include conflicts that would interfere with rendering unbiased investment advice and conflicts that would cause a member or candidate to act not in the employer's best interest. The same circumstances that generate conflicts to be reported to clients and prospective clients also would dictate reporting to employers. Ownership of stocks analyzed or recommended, participation on outside boards, and financial or other pressures that could influence a decision are to be promptly reported to the employer so that their impact can be assessed and a decision on how to resolve the conflict can be made.

The mere appearance of a conflict of interest may create problems for members, candidates, and their employers. Therefore, many of the conflicts previously mentioned could be explicitly prohibited by an employer. For example, many employers restrict personal trading, outside board membership, and related activities to prevent situations that might not normally be considered problematic from a conflict-of-interest point of view but that could give the appearance of a conflict of interest. Members and candidates must comply with these restrictions. Members and candidates must take reasonable steps to avoid conflicts and, if they occur inadvertently, must report them promptly so that the employer and the member or candidate can resolve them as quickly and effectively as possible.

Standard VI(A) also deals with a member's or candidate's conflicts of interest that might be detrimental to the employer's business. Any potential conflict situation that could prevent clear judgment about or full commitment to the execution of a member's or candidate's duties to the employer should be reported to the member's or candidate's employer and promptly resolved.

Disclosure to Clients

Members and candidates must maintain their objectivity when rendering investment advice or taking investment action. Investment advice or actions may be perceived to be tainted in numerous situations. Can a member or candidate remain objective if, on behalf of the firm, the member or candidate obtains or assists in obtaining fees for services? Can a member or candidate give objective advice if he or she owns stock in the company that is the subject of an investment recommendation or if the member or candidate has a close personal relationship with the company managers? Requiring members and candidates to disclose all matters that reasonably could be expected to impair the member's or candidate's objectivity allows clients and prospective clients to judge motives and possible biases for themselves.

Often in the investment industry, a conflict, or the perception of a conflict, cannot be avoided. The most obvious conflicts of interest, which should always be disclosed, are relationships between an issuer and the member, the candidate, or his or her firm (such as a directorship or consultancy by a member; investment banking, underwriting, and financial relationships; broker/dealer market-making activities; and material beneficial ownership of stock). For the purposes of Standard VI(A), members and candidates beneficially own securities or other investments if they have a direct or indirect pecuniary interest in the securities, have the power to vote or direct the voting of the shares of the securities or investments, or have the power to dispose or direct the disposition of the security or investment.

A member or candidate must take reasonable steps to determine whether a conflict of interest exists and disclose to clients any known conflicts of the member's or candidate's firm. Disclosure of broker/dealer market-making activities alerts clients that a purchase or sale might be made from or to the firm's principal account and that the firm has a special interest in the price of the stock.

Additionally, disclosures should be made to clients regarding fee arrangements, subadvisory agreements, or other situations involving nonstandard fee structures. Equally important is the disclosure of arrangements in which the firm benefits directly from investment recommendations. An obvious conflict of interest is the rebate of a portion of the service fee some classes of mutual funds charge to investors. Members and candidates should ensure that their firms disclose such relationships so clients can fully understand the costs of their investments and the benefits received by their investment manager's employer.

Cross-Departmental Conflicts

Other circumstances can give rise to actual or potential conflicts of interest. For instance, a sell-side analyst working for a broker/dealer may be encouraged, not only by members of her or his own firm but by corporate issuers themselves, to write research reports about particular companies. The buy-side analyst is likely to be faced with similar conflicts as banks exercise their underwriting and security-dealing powers. The marketing division may ask an analyst to recommend the stock of a certain company in order to obtain business from that company.

The potential for conflicts of interest also exists with broker-sponsored limited partnerships formed to invest venture capital. Increasingly, members and candidates are expected not only to follow issues from these partnerships once they are offered to the public but also to promote the issues in the secondary market after public offerings. Members, candidates, and their firms should attempt to resolve situations presenting potential conflicts of interest or disclose them in accordance with the principles set forth in Standard VI(A).

Conflicts with Stock Ownership

The most prevalent conflict requiring disclosure under Standard VI(A) is a member's or candidate's ownership of stock in companies that he or she recommends to clients or that clients hold. Clearly, the easiest method for preventing a conflict is to prohibit members and candidates from owning any such securities, but this approach is overly burdensome and discriminates against members and candidates.

Therefore, sell-side members and candidates should disclose any materially beneficial ownership interest in a security or other investment that the member or candidate is recommending. Buy-side members and candidates should disclose their procedures for reporting requirements for personal transactions. Conflicts arising from personal investing are discussed more fully in the guidance for Standard VI(B).

Conflicts as a Director

Service as a director poses three basic conflicts of interest. First, a conflict may exist between the duties owed to clients and the duties owed to shareholders of the company. Second, investment personnel who serve as directors may receive the securities or options to purchase securities of the company as compensation for serving on the board, which could raise questions about trading actions that might increase the value of those securities. Third, board service creates the opportunity to receive material nonpublic information involving the company. Even though the information is confidential, the perception could be that information not available to the public is being communicated to a director's firm—whether a broker, investment adviser, or other type of organization. When members or candidates providing investment services also serve as directors, they should be isolated from those making investment decisions by the use of firewalls or similar restrictions.

Recommended Procedures for Compliance

Members or candidates should disclose special compensation arrangements with the employer that might conflict with client interests, such as bonuses based on short-term performance criteria, commissions, incentive fees, performance fees, and referral fees. If the member's or candidate's firm does not permit such disclosure, the member or candidate should document the request and may consider dissociating from the activity.

Members' and candidates' firms are encouraged to include information on compensation packages in firms' promotional literature. If a member or candidate manages a portfolio for which the fee is based on capital gains or capital appreciation (a performance fee), this information should be disclosed to clients. If a member, a candidate, or a member's or candidate's firm has outstanding agent options to buy stock as part of the compensation package for corporate financing activities, the amount and expiration date of these options should be disclosed as a footnote to any research report published by the member's or candidate's firm.

Application of the Standard

Example 1 (Conflict of Interest and Business Relationships):

Hunter Weiss is a research analyst with Farmington Company, a broker and investment banking firm. Farmington's merger and acquisition department has represented Vimco, a conglomerate, in all of Vimco's acquisitions for 20 years. From time to time, Farmington officers sit on the boards of directors of various Vimco subsidiaries. Weiss is writing a research report on Vimco.

> *Comment*: Weiss must disclose in his research report Farmington's special relationship with Vimco. Broker/dealer management of and participation in public offerings must be disclosed in research reports. Because the position of underwriter to a company entails a special past and potential future relationship with a company that is the subject of investment advice, it threatens the independence and objectivity of the report writer and must be disclosed.

Example 2 (Conflict of Interest and Business Stock Ownership):

The investment management firm of Dover & Roe sells a 25% interest in its partnership to a multinational bank holding company, First of New York. Immediately after the sale, Margaret Hobbs, president of Dover & Roe, changes her recommendation for First of New York's common stock from "sell" to "buy" and adds First of New York's commercial paper to Dover & Roe's approved list for purchase.

> *Comment*: Hobbs must disclose the new relationship with First of New York to all Dover & Roe clients. This relationship must also be disclosed to clients by the firm's portfolio managers when they make specific investment recommendations or take investment actions with respect to First of New York's securities.

Example 3 (Conflict of Interest and Personal Stock Ownership):

Carl Fargmon, a research analyst who follows firms producing office equipment, has been recommending purchase of Kincaid Printing because of its innovative new line of copiers. After his initial report on the company, Fargmon's wife inherits from a distant relative US$3 million of Kincaid stock. He has been asked to write a follow-up report on Kincaid.

> *Comment*: Fargmon must disclose his wife's ownership of the Kincaid stock to his employer and in his follow-up report. Best practice would be to avoid the conflict by asking his employer to assign another analyst to draft the follow-up report.

Example 4 (Conflict of Interest and Personal Stock Ownership):

Betty Roberts is speculating in penny stocks for her own account and purchases 100,000 shares of Drew Mining, Inc., for US$0.30 a share. She intends to sell these shares at the sign of any substantial upward price movement of the stock. A week later, her employer asks her to write a report on penny stocks in the mining industry to be published in two weeks. Even without owning the Drew stock, Roberts would recommend it in her report as a "buy." A surge in the price of the stock to the US$2 range is likely to result once the report is issued.

> *Comment*: Although this holding may not be material, Roberts must disclose it in the report and to her employer before writing the report because the gain for her will be substantial if the market responds strongly to her recommendation. The fact that she has only recently purchased the stock adds to the appearance that she is not entirely objective.

Example 5 (Conflict of Interest and Compensation Arrangements):

Samantha Snead, a portfolio manager for Thomas Investment Counsel, Inc., specializes in managing public retirement funds and defined benefit pension plan accounts, all of which have long-term investment objectives. A year ago, Snead's employer, in an attempt to motivate and retain key investment professionals, introduced a bonus compensation system that rewards portfolio managers on the basis of quarterly performance relative to their peers and to certain benchmark indexes. In an attempt to improve the short-term performance of her accounts, Snead changes her investment strategy and purchases several high-beta stocks for client portfolios. These purchases are seemingly contrary to the clients' investment policy statements. Following their purchase, an officer of Griffin Corporation, one of Snead's pension fund clients, asks why Griffin Corporation's portfolio seems to be dominated by high-beta stocks of companies that often appear among the most actively traded issues. No change in objective or strategy has been recommended by Snead during the year.

> *Comment*: Snead has violated Standard VI(A) by failing to inform her clients of the changes in her compensation arrangement with her employer, which created a conflict of interest between her compensation and her clients' IPSs. Firms may pay employees on the basis of performance, but pressure by Thomas Investment Counsel to achieve short-term performance goals is in basic conflict with the objectives of Snead's accounts.
>
> See also Standard III(C)–Suitability.

Example 6 (Conflict of Interest, Options, and Compensation Arrangements):

Wayland Securities works with small companies doing IPOs or secondary offerings. Typically, these deals are in the US$10 million to US$50 million range, and as a result, the corporate finance fees are quite small. To compensate for the small fees, Wayland Securities usually takes "agent options"—that is, rights (exercisable within a two-year time frame) to acquire up to an additional 10% of the current offering. Following an IPO performed by Wayland for Falk Resources, Ltd., Darcy Hunter, the head of corporate finance at Wayland, is concerned about receiving value for her Falk Resources options. The options are due to expire in one month, and the stock is not doing well. She contacts John Fitzpatrick in the research department of Wayland Securities, reminds him that he is eligible for 30% of these options, and indicates that now would be a good time to give some additional coverage to Falk Resources. Fitzpatrick agrees and immediately issues a favorable report.

> *Comment*: For Fitzpatrick to avoid being in violation of Standard VI(A), he must indicate in the report the volume and expiration date of agent options outstanding. Furthermore, because he is personally eligible for some of the options, Fitzpatrick must disclose the extent of this compensation. He also must be careful to not violate his duty of independence and objectivity under Standard I(B).

Example 7 (Conflict of Interest and Compensation Arrangements):

Gary Carter is a representative with Bengal International, a registered broker/dealer. Carter is approached by a stock promoter for Badger Company, who offers to pay Carter additional compensation for sales of Badger Company's stock to Carter's clients. Carter accepts the stock promoter's offer but does not disclose the arrangements to his clients or to his employer. Carter sells shares of the stock to his clients.

> *Comment*: Carter has violated Standard VI(A) by failing to disclose to clients that he is receiving additional compensation for recommending and selling Badger stock. Because he did not disclose the arrangement with Badger to his clients, the clients were unable to evaluate whether Carter's recommendations to buy Badger were affected by this arrangement. Carter's conduct also violated Standard VI(A) by failing to disclose to his employer monetary compensation received in addition to the compensation and benefits conferred by his employer. Carter was required by Standard VI(A) to disclose the arrangement with Badger to his employer so that his employer could evaluate whether the arrangement affected Carter's objectivity and loyalty.

Example 8 (Conflict of Interest and Directorship):

Carol Corky, a senior portfolio manager for Universal Management, recently became involved as a trustee with the Chelsea Foundation, a large not-for-profit foundation in her hometown. Universal is a small money manager (with assets under management of approximately US$100 million) that caters to individual investors. Chelsea has assets in excess of US$2 billion. Corky does not believe informing Universal of her involvement with Chelsea is necessary.

> *Comment*: By failing to inform Universal of her involvement with Chelsea, Corky violated Standard VI(A). Given the large size of the endowment at Chelsea, Corky's new role as a trustee can reasonably be expected to be time consuming, to the possible detriment of Corky's portfolio responsibilities with Universal. Also, as a trustee, Corky may become involved in the investment decisions at Chelsea. Therefore, Standard VI(A) obligates Corky to discuss becoming a trustee at Chelsea with her compliance officer

or supervisor at Universal before accepting the position, and she should have disclosed the degree to which she would be involved in investment decisions at Chelsea.

Example 9 (Conflict of Interest and Personal Trading):

Bruce Smith covers eastern European equities for Marlborough Investments, an investment management firm with a strong presence in emerging markets. While on a business trip to Russia, Smith learns that investing in Russian equities directly is difficult but that equity-linked notes that replicate the performance of underlying Russian equities can be purchased from a New York–based investment bank. Believing that his firm would not be interested in such a security, Smith purchases a note linked to a Russian telecommunications company for his own account without informing Marlborough. A month later, Smith decides that the firm should consider investing in Russian equities by way of the equity-linked notes. He prepares a write-up on the market that concludes with a recommendation to purchase several of the notes. One note he recommends is linked to the same Russian telecom company that Smith holds in his personal account.

Comment: Smith has violated Standard VI(A) by failing to disclose his purchase and ownership of the note linked to the Russian telecom company. Smith is required by the standard to disclose the investment opportunity to his employer and look to his company's policies on personal trading to determine whether it was proper for him to purchase the note for his own account. By purchasing the note, Smith may or may not have impaired his ability to make an unbiased and objective assessment of the appropriateness of the derivative instrument for his firm, but Smith's failure to disclose the purchase to his employer impaired his employer's ability to decide whether his ownership of the security is a conflict of interest that might affect Smith's future recommendations. Then, when he recommended the particular telecom notes to his firm, Smith compounded his problems by not disclosing that he owned the notes in his personal account—a clear conflict of interest.

Example 10 (Conflict of Interest and Requested Favors):

Michael Papis is the chief investment officer of his state's retirement fund. The fund has always used outside advisers for the real estate allocation, and this information is clearly presented in all fund communications. Thomas Nagle, a recognized sell-side research analyst and Papis's business school classmate, recently left the investment bank he worked for to start his own asset management firm, Accessible Real Estate. Nagle is trying to build his assets under management and contacts Papis about gaining some of the retirement fund's allocation. In the previous few years, the performance of the retirement fund's real estate investments was in line with the fund's benchmark but was not extraordinary. Papis decides to help out his old friend and also to seek better returns by moving the real estate allocation to Accessible. The only notice of the change in adviser appears in the next annual report in the listing of associated advisers.

Comment: Papis has violated Standard VI(A) by not disclosing to his employer his personal relationship with Nagle. Disclosure of his past history with Nagle would allow his firm to determine whether the conflict may have impaired Papis's independence in deciding to change managers.

See also Standard IV(C)–Responsibilities of Supervisors, Standard V(A)–Diligence and Reasonable Basis, and Standard V(B)–Communication with Clients and Prospective Clients.

Example 11 (Conflict of Interest and Business Relationships):

Bob Wade, trust manager for Central Midas Bank, was approached by Western Funds about promoting its family of funds, with special interest in the service-fee class. To entice Central to promote this class, Western Funds offered to pay the bank a service fee of 0.25%. Without disclosing the fee being offered to the bank, Wade asked one of the investment managers to review the Western Funds family of funds to determine whether they were suitable for clients of Central. The manager completed the normal due diligence review and determined that the funds were fairly valued in the market with fee structures on a par with their competitors. Wade decided to accept Western's offer and instructed the team of portfolio managers to exclusively promote these funds and the service-fee class to clients seeking to invest new funds or transfer from their current investments. So as to not influence the investment managers, Wade did not disclose the fee offer and allowed that income to flow directly to the bank.

> *Comment*: Wade is violating Standard VI(A) by not disclosing the portion of the service fee being paid to Central. Although the investment managers may not be influenced by the fee, neither they nor the client have the proper information about Wade's decision to exclusively market this fund family and class of investments. Central may come to rely on the new fee as a component of the firm's profitability and may be unwilling to offer other products in the future that could affect the fees received.
>
> See also Standard I(B)–Independence and Objectivity.

Example 12 (Disclosure of Conflicts to Employers):

Yehudit Dagan is a portfolio manager for Risk Management Bank (RMB), whose clients include retirement plans and corporations. RMB provides a defined contribution retirement plan for its employees that offers 20 large diversified mutual fund investment options, including a mutual fund managed by Dagan's RMB colleagues. After being employed for six months, Dagan became eligible to participate in the retirement plan, and she intends to allocate her retirement plan assets in six of the investment options, including the fund managed by her RMB colleagues. Dagan is concerned that joining the plan will lead to a potentially significant amount of paperwork for her (e.g., disclosure of her retirement account holdings and needing preclearance for her transactions), especially with her investing in the in-house fund.

> *Comment*: Standard VI(A) would not require Dagan to disclose her personal or retirement investments in large diversified mutual funds, unless specifically required by her employer. For practical reasons, the standard does not require Dagan to gain preclearance for ongoing payroll deduction contributions to retirement plan account investment options.
>
> Dagan should ensure that her firm does not have a specific policy regarding investment—whether personal or in the retirement account—for funds managed by the company's employees. These mutual funds may be subject to the company's disclosure, preclearance, and trading restriction procedures to identify possible conflicts prior to the execution of trades.

Standard VI(B) Priority of Transactions

Investment transactions for clients and employers must have priority over investment transactions in which a Member or Candidate is the beneficial owner.

Guidance

Highlights:

- *Avoiding Potential Conflicts*
- *Personal Trading Secondary to Trading for Clients*
- *Standards for Nonpublic Information*
- *Impact on All Accounts with Beneficial Ownership*

Standard VI(B) reinforces the responsibility of members and candidates to give the interests of their clients and employers priority over their personal financial interests. This standard is designed to prevent any potential conflict of interest or the appearance of a conflict of interest with respect to personal transactions. Client interests have priority. Client transactions must take precedence over transactions made on behalf of the member's or candidate's firm or personal transactions.

Avoiding Potential Conflicts

Conflicts between the client's interest and an investment professional's personal interest may occur. Although conflicts of interest exist, nothing is inherently unethical about individual managers, advisers, or mutual fund employees making money from personal investments as long as (1) the client is not disadvantaged by the trade, (2) the investment professional does not benefit personally from trades undertaken for clients, and (3) the investment professional complies with applicable regulatory requirements.

Some situations occur where a member or candidate may need to enter a personal transaction that runs counter to current recommendations or what the portfolio manager is doing for client portfolios. For example, a member or candidate may be required at some point to sell an asset to make a college tuition payment or a down payment on a home, to meet a margin call, or so on. The sale may be contrary to the long-term advice the member or candidate is currently providing to clients. In these situations, the same three criteria given in the preceding paragraph should be applied in the transaction so as to not violate Standard VI(B).

Personal Trading Secondary to Trading for Clients

Standard VI(B) states that transactions for clients and employers must have priority over transactions in securities or other investments for which a member or candidate is the beneficial owner. The objective of the standard is to prevent personal transactions from adversely affecting the interests of clients or employers. A member or candidate having the same investment positions or being co-invested with clients does not always create a conflict. Some clients in certain investment situations require members or candidates to have aligned interests. Personal investment positions or transactions of members or candidates or their firm should never, however, adversely affect client investments.

Standards for Nonpublic Information

Standard VI(B) covers the activities of members and candidates who have knowledge of pending transactions that may be made on behalf of their clients or employers, who have access to nonpublic information during the normal preparation of research recommendations, or who take investment actions. Members and candidates are prohibited from conveying nonpublic information to any person whose relationship to the member or candidate makes the member or candidate a beneficial owner of the person's securities. Members and candidates must not convey this information to any other person if the nonpublic information can be deemed material.

Impact on All Accounts with Beneficial Ownership

Members or candidates may undertake transactions in accounts for which they are a beneficial owner only after their clients and employers have had adequate opportunity to act on a recommendation. Personal transactions include those made for the member's or candidate's own account, for family (including spouse, children, and other immediate family members) accounts, and for accounts in which the member or candidate has a direct or indirect pecuniary interest, such as a trust or retirement account. Family accounts that are client accounts should be treated like any other firm account and should neither be given special treatment nor be disadvantaged because of the family relationship. If a member or candidate has a beneficial ownership in the account, however, the member or candidate may be subject to preclearance or reporting requirements of the employer or applicable law.

Recommended Procedures for Compliance

Policies and procedures designed to prevent potential conflicts of interest, and even the appearance of a conflict of interest, with respect to personal transactions are critical to establishing investor confidence in the securities industry. Therefore, members and candidates should urge their firms to establish such policies and procedures. Because investment firms vary greatly in assets under management, types of clients, number of employees, and so on, each firm should have policies regarding personal investing that are best suited to the firm. Members and candidates should then prominently disclose these policies to clients and prospective clients.

The specific provisions of each firm's standards will vary, but all firms should adopt certain basic procedures to address the conflict areas created by personal investing. These procedures include the following:

- *Limited participation in equity IPOs*: Some eagerly awaited IPOs rise significantly in value shortly after the issue is brought to market. Because the new issue may be highly attractive and sought after, the opportunity to participate in the IPO may be limited. Therefore, purchases of IPOs by investment personnel create conflicts of interest in two principal ways. First, participation in an IPO may have the appearance of taking away an attractive investment opportunity from clients for personal gain—a clear breach of the duty of loyalty to clients. Second, personal purchases in IPOs may have the appearance that the investment opportunity is being bestowed as an incentive to make future investment decisions for the benefit of the party providing the opportunity. Members and candidates can avoid these conflicts or appearances of conflicts of interest by not participating in IPOs.

  Reliable and systematic review procedures should be established to ensure that conflicts relating to IPOs are identified and appropriately dealt with by supervisors. Members and candidates should preclear their participation in IPOs, even in situations without any conflict of interest between a member's or candidate's participation in an IPO and the client's interests. Members and candidates should not benefit from the position that their clients occupy in the marketplace—through preferred trading, the allocation of limited offerings, or oversubscription.

- *Restrictions on private placements*: Strict limits should be placed on investment personnel acquiring securities in private placements, and appropriate supervisory and review procedures should be established to prevent noncompliance.

  Firms do not routinely use private placements for clients (e.g., venture capital deals) because of the high risk associated with them. Conflicts related to private placements are more significant to members and candidates who manage large pools of assets or act as plan sponsors because these managers may be offered special opportunities, such as private placements, as a reward or an enticement for continuing to do business with a particular broker.

  Participation in private placements raises conflict-of-interest issues that are similar to issues surrounding IPOs. Investment personnel should not be involved in transactions, including (but not limited to) private placements, that could be perceived as favors or gifts that seem designed to influence future judgment or to reward past business deals.

  Whether the venture eventually proves to be good or bad, managers have an immediate conflict concerning private placement opportunities. If and when the investments go public, participants in private placements have an incentive to recommend the investments to clients regardless of the suitability of the investments for their clients. Doing so increases the value of the participants' personal portfolios.

- *Establish blackout/restricted periods*: Investment personnel involved in the investment decision-making process should establish blackout periods prior to trades for clients so that managers cannot take advantage of their knowledge of client activity by "front-running" client trades (trading for one's personal account before trading for client accounts).

  Individual firms must decide who within the firm should be required to comply with the trading restrictions. At a minimum, all individuals who are involved in the investment decision-making process should be subject to the same restricted period. Each firm must determine specific requirements related to blackout and restricted periods that are most relevant to the firm while ensuring that the procedures are governed by the guiding principles set forth in the Code and Standards. Size of firm and type of securities purchased are relevant factors. For example, in a large firm, a blackout requirement is, in effect, a total trading ban because the firm is continually trading in most securities. In a small firm, the blackout period is more likely to prevent the investment manager from front-running.

- *Reporting requirements*: Supervisors should establish reporting procedures for investment personnel, including disclosure of personal holdings/beneficial ownerships, confirmations of trades to the firm and the employee, and preclearance procedures. Once trading restrictions are in place, they must be enforced. The best method for monitoring and enforcing procedures to eliminate conflicts of interest in personal trading is through reporting requirements, including the following:

  - **Disclosure of holdings in which the employee has a beneficial interest**. Disclosure by investment personnel to the firm should be made upon commencement of the employment relationship and at least annually thereafter. To address privacy considerations, disclosure of personal holdings should be handled in a confidential manner by the firm.

- **Providing duplicate confirmations of transactions**. Investment personnel should be required to direct their brokers to supply to firms duplicate copies or confirmations of all their personal securities transactions and copies of periodic statements for all securities accounts. The duplicate confirmation requirement has two purposes: (1) The requirement sends a message that there is independent verification, which reduces the likelihood of unethical behavior, and (2) it enables verification of the accounting of the flow of personal investments that cannot be determined from merely looking at holdings.
- **Preclearance procedures**. Investment personnel should examine all planned personal trades to identify possible conflicts prior to the execution of the trades. Preclearance procedures are designed to identify possible conflicts before a problem arises.

■ *Disclosure of policies*: Members and candidates should fully disclose to investors their firm's policies regarding personal investing. The information about employees' personal investment activities and policies will foster an atmosphere of full and complete disclosure and calm the public's legitimate concerns about the conflicts of interest posed by investment personnel's personal trading. The disclosure must provide helpful information to investors; it should not be simply boilerplate language, such as "investment personnel are subject to policies and procedures regarding their personal trading."

Application of the Standard

Example 1 (Personal Trading):

Research analyst Marlon Long does not recommend purchase of a common stock for his employer's account because he wants to purchase the stock personally and does not want to wait until the recommendation is approved and the stock is purchased by his employer.

> *Comment*: Long has violated Standard VI(B) by taking advantage of his knowledge of the stock's value before allowing his employer to benefit from that information.

Example 2 (Trading for Family Member Account):

Carol Baker, the portfolio manager of an aggressive growth mutual fund, maintains an account in her husband's name at several brokerage firms with which the fund and a number of Baker's other individual clients do a substantial amount of business. Whenever a hot issue becomes available, she instructs the brokers to buy it for her husband's account. Because such issues normally are scarce, Baker often acquires shares in hot issues but her clients are not able to participate in them.

> *Comment*: To avoid violating Standard VI(B), Baker must acquire shares for her mutual fund first and acquire them for her husband's account only after doing so, even though she might miss out on participating in new issues via her husband's account. She also must disclose the trading for her husband's account to her employer because this activity creates a conflict between her personal interests and her employer's interests.

Example 3 (Family Accounts as Equals):

Erin Toffler, a portfolio manager at Esposito Investments, manages the retirement account established with the firm by her parents. Whenever IPOs become available, she first allocates shares to all her other clients for whom the investment is appropriate; only then does she place any remaining portion in her parents' account, if the issue is appropriate for them. She has adopted this procedure so that no one can accuse her of favoring her parents.

> *Comment*: Toffler has violated Standard VI(B) by breaching her duty to her parents by treating them differently from her other accounts simply because of the family relationship. As fee-paying clients of Esposito Investments, Toffler's parents are entitled to the same treatment as any other client of the firm. If Toffler has beneficial ownership in the account, however, and Esposito Investments has preclearance and reporting requirements for personal transactions, she may have to preclear the trades and report the transactions to Esposito.

Example 4 (Personal Trading and Disclosure):

Gary Michaels is an entry-level employee who holds a low-paying job serving both the research department and the investment management department of an active investment management firm. He purchases a sports car and begins to wear expensive clothes after only a year of employment with the firm. The director of the investment management department, who has responsibility for monitoring the personal stock transactions of all employees, investigates and discovers that Michaels has made substantial investment gains by purchasing stocks just before they were put on the firm's recommended "buy" list. Michaels was regularly given the firm's quarterly personal transaction form but declined to complete it.

> *Comment*: Michaels violated Standard VI(B) by placing personal transactions ahead of client transactions. In addition, his supervisor violated Standard IV(C)–Responsibilities of Supervisors by permitting Michaels to continue to perform his assigned tasks without having signed the quarterly personal transaction form. Note also that if Michaels had communicated information about the firm's recommendations to a person who traded the security, that action would be a misappropriation of the information and a violation of Standard II(A)–Material Nonpublic Information.

Example 5 (Trading Prior to Report Dissemination):

A brokerage's insurance analyst, Denise Wilson, makes a closed-circuit TV report to her firm's branches around the country. During the broadcast, she includes negative comments about a major company in the insurance industry. The following day, Wilson's report is printed and distributed to the sales force and public customers. The report recommends that both short-term traders and intermediate investors take profits by selling that insurance company's stock. Seven minutes after the broadcast, however, Ellen Riley, head of the firm's trading department, had closed out a long "call" position in the stock. Shortly thereafter, Riley established a sizable "put" position in the stock. When asked about her activities, Riley claimed she took the actions to facilitate anticipated sales by institutional clients.

> *Comment*: Riley did not give customers an opportunity to buy or sell in the options market before the firm itself did. By taking action before the report was disseminated, Riley's firm may have depressed the price of the calls and increased the price of the puts. The firm could have avoided a conflict

of interest if it had waited to trade for its own account until its clients had an opportunity to receive and assimilate Wilson's recommendations. As it is, Riley's actions violated Standard VI(B).

Standard VI(C) Referral Fees

Members and Candidates must disclose to their employer, clients, and prospective clients, as appropriate, any compensation, consideration, or benefit received from or paid to others for the recommendation of products or services.

Guidance

Standard VI(C) states the responsibility of members and candidates to inform their employer, clients, and prospective clients of any benefit received for referrals of customers and clients. Such disclosures allow clients or employers to evaluate (1) any partiality shown in any recommendation of services and (2) the full cost of the services. Members and candidates must disclose when they pay a fee or provide compensation to others who have referred prospective clients to the member or candidate.

Appropriate disclosure means that members and candidates must advise the client or prospective client, before entry into any formal agreement for services, of any benefit given or received for the recommendation of any services provided by the member or candidate. In addition, the member or candidate must disclose the nature of the consideration or benefit—for example, flat fee or percentage basis, one-time or continuing benefit, based on performance, benefit in the form of provision of research or other noncash benefit—together with the estimated dollar value. Consideration includes all fees, whether paid in cash, in soft dollars, or in kind.

Recommended Procedures for Compliance

Members and candidates should encourage their employers to develop procedures related to referral fees. The firm may completely restrict such fees. If the firm does not adopt a strict prohibition of such fees, the procedures should indicate the appropriate steps for requesting approval.

Employers should have investment professionals provide to the clients notification of approved referral fee programs and provide the employer regular (at least quarterly) updates on the amount and nature of compensation received.

Application of the Standard

Example 1 (Disclosure of Referral Arrangements and Outside Parties):

Brady Securities, Inc., a broker/dealer, has established a referral arrangement with Lewis Brothers, Ltd., an investment counseling firm. In this arrangement, Brady Securities refers all prospective tax-exempt accounts, including pension, profit-sharing, and endowment accounts, to Lewis Brothers. In return, Lewis Brothers makes available to Brady Securities on a regular basis the security recommendations and reports of its research staff, which registered representatives of Brady Securities use in serving customers. In addition, Lewis Brothers conducts monthly economic and market reviews for Brady Securities personnel and directs all stock commission business generated by referral accounts to Brady Securities.

Willard White, a partner in Lewis Brothers, calculates that the incremental costs involved in functioning as the research department of Brady Securities are US$20,000 annually.

Referrals from Brady Securities last year resulted in fee income of US$200,000 for Lewis Brothers, and directing all stock trades through Brady Securities resulted in additional costs to Lewis Brothers' clients of US$10,000.

Diane Branch, the chief financial officer of Maxwell Inc., contacts White and says that she is seeking an investment manager for Maxwell's profit-sharing plan. She adds, "My friend Harold Hill at Brady Securities recommended your firm without qualification, and that's good enough for me. Do we have a deal?" White accepts the new account but does not disclose his firm's referral arrangement with Brady Securities.

> *Comment*: White has violated Standard VI(C) by failing to inform the prospective customer of the referral fee payable in services and commissions for an indefinite period to Brady Securities. Such disclosure could have caused Branch to reassess Hill's recommendation and make a more critical evaluation of Lewis Brothers' services.

Example 2 (Disclosure of Interdepartmental Referral Arrangements):

James Handley works for the trust department of Central Trust Bank. He receives compensation for each referral he makes to Central Trust's brokerage department and personal financial management department that results in a sale. He refers several of his clients to the personal financial management department but does not disclose the arrangement within Central Trust to his clients.

> *Comment*: Handley has violated Standard VI(C) by not disclosing the referral arrangement at Central Trust Bank to his clients. Standard VI(C) does not distinguish between referral payments paid by a third party for referring clients to the third party and internal payments paid within the firm to attract new business to a subsidiary. Members and candidates must disclose all such referral fees. Therefore, Handley is required to disclose, at the time of referral, any referral fee agreement in place among Central Trust Bank's departments. The disclosure should include the nature and the value of the benefit and should be made in writing.

Example 3 (Disclosure of Referral Arrangements and Informing Firm):

Katherine Roberts is a portfolio manager at Katama Investments, an advisory firm specializing in managing assets for high-net-worth individuals. Katama's trading desk uses a variety of brokerage houses to execute trades on behalf of its clients. Roberts asks the trading desk to direct a large portion of its commissions to Naushon, Inc., a small broker/dealer run by one of Roberts' business school classmates. Katama's traders have found that Naushon is not very competitive on pricing, and although Naushon generates some research for its trading clients, Katama's other analysts have found most of Naushon's research to be not especially useful. Nevertheless, the traders do as Roberts asks, and in return for receiving a large portion of Katama's business, Naushon recommends the investment services of Roberts and Katama to its wealthiest clients. This arrangement is not disclosed to either Katama or the clients referred by Naushon.

> *Comment*: Roberts is violating Standard VI(C) by failing to inform her employer of the referral arrangement.

Example 4 (Disclosure of Referral Arrangements and Outside Organizations):

Alex Burl is a portfolio manager at Helpful Investments, a local investment advisory firm. Burl is on the advisory board of his child's school, which is looking for ways to raise money to purchase new playground equipment for the school. Burl discusses a plan with his supervisor in which he will donate to the school a portion of his service fee from new clients referred by the parents of students at the school. Upon getting the approval from Helpful, Burl presents the idea to the school's advisory board and directors. The school agrees to announce the program at the next parent event and asks Burl to provide the appropriate written materials to be distributed. A week following the distribution of the flyers, Burl receives the first school-related referral. In establishing the client's investment policy statement, Burl clearly discusses the school's referral and outlines the plans for distributing the donation back to the school.

> *Comment*: Burl has not violated Standard VI(C) because he secured the permission of his employer, Helpful Investments, and the school prior to beginning the program and because he discussed the arrangement with the client at the time the investment policy statement was designed.

Example 5 (Disclosure of Referral Arrangements and Outside Parties):

The sponsor of a state employee pension is seeking to hire a firm to manage the pension plan's emerging market allocation. To assist in the review process, the sponsor has hired Thomas Arrow as a consultant to solicit proposals from various advisers. Arrow is contracted by the sponsor to represent its best interest in selecting the most appropriate new manager. The process runs smoothly, and Overseas Investments is selected as the new manager.

The following year, news breaks that Arrow is under investigation by the local regulator for accepting kickbacks from investment managers after they are awarded new pension allocations. Overseas Investments is included in the list of firms allegedly making these payments. Although the sponsor is happy with the performance of Overseas since it has been managing the pension plan's emerging market funds, the sponsor still decides to have an independent review of the proposals and the selection process to ensure that Overseas was the appropriate firm for its needs. This review confirms that, even though Arrow was being paid by both parties, the recommendation of Overseas appeared to be objective and appropriate.

> *Comment*: Arrow has violated Standard VI(C) because he did not disclose the fee being paid by Overseas. Withholding this information raises the question of a potential lack of objectivity in the recommendation of Overseas by Arrow; this aspect is in addition to questions about the legality of having firms pay to be considered for an allocation.
>
> Regulators and governmental agencies may adopt requirements concerning allowable consultant activities. Local regulations sometimes include having a consultant register with the regulatory agency's ethics board. Regulator policies may include a prohibition on acceptance of payments from investment managers receiving allocations and require regular reporting of contributions made to political organizations and candidates. Arrow would have to adhere to these requirements as well as the Code and Standards.

STANDARD VII: RESPONSIBILITIES AS A CFA INSTITUTE MEMBER OR CFA CANDIDATE

Standard VII(A) Conduct as Participants in CFA Institute Programs

Members and Candidates must not engage in any conduct that compromises the reputation or integrity of CFA Institute or the CFA designation or the integrity, validity, or security of CFA Institute programs.

Guidance

Highlights:

- *Confidential Program Information*
- *Additional CFA Program Restrictions*
- *Expressing an Opinion*

Standard VII(A) covers the conduct of CFA Institute members and candidates involved with the CFA Program and prohibits any conduct that undermines the public's confidence that the CFA charter represents a level of achievement based on merit and ethical conduct. There is an array of CFA Institute programs beyond the CFA Program that provide additional educational and credentialing opportunities, including the Certificate in Investment Performance Measurement (CIPM) Program and the CFA Institute Investment Foundations™ Program. The standard's function is to hold members and candidates to a high ethical criterion while they are participating in or involved with any CFA Institute program. Conduct covered includes but is not limited to

- giving or receiving assistance (cheating) on any CFA Institute examinations;
- violating the rules, regulations, and testing policies of CFA Institute programs;
- providing confidential program or exam information to candidates or the public;
- disregarding or attempting to circumvent security measures established for any CFA Institute examinations;
- improperly using an association with CFA Institute to further personal or professional goals; and
- misrepresenting information on the Professional Conduct Statement or in the CFA Institute Continuing Education Program.

Confidential Program Information

CFA Institute is vigilant about protecting the integrity of CFA Institute programs' content and examination processes. CFA Institute program rules, regulations, and policies prohibit candidates from disclosing confidential material gained during the exam process.

Examples of information that cannot be disclosed by candidates sitting for an exam include but are not limited to

- specific details of questions appearing on the exam and
- broad topical areas and formulas tested or not tested on the exam.

All aspects of the exam, including questions, broad topical areas, and formulas, tested or not tested, are considered confidential until such time as CFA Institute elects to release them publicly. This confidentiality requirement allows CFA Institute to maintain the integrity and rigor of exams for future candidates. Standard VII(A) does not prohibit candidates from discussing nonconfidential information or curriculum material with others or in study groups in preparation for the exam.

Candidates increasingly use online forums and new technology as part of their exam preparations. CFA Institute actively polices blogs, forums, and related social networking groups for information considered confidential. The organization works with both individual candidates and the sponsors of online or offline services to promptly remove any and all violations. As noted in the discussion of Standard I(A)–Knowledge of the Law, candidates, members, and the public are encouraged to report suspected violations to CFA Institute.

Additional CFA Program Restrictions

The CFA Program rules, regulations, and policies define additional allowed and disallowed actions concerning the exams. Violating any of the testing policies, such as the calculator policy, personal belongings policy, or the Candidate Pledge, constitutes a violation of Standard VII(A). Candidates will find all of these policies on the CFA Program portion of the CFA Institute website (www.cfainstitute.org). Exhibit 2 provides the Candidate Pledge, which highlights the respect candidates must have for the integrity, validity, and security of the CFA exam.

Members may participate as volunteers in various aspects of the CFA Program. Standard VII(A) prohibits members from disclosing and/or soliciting confidential material gained prior to or during the exam and grading processes with those outside the CFA exam development process.

Examples of information that cannot be shared by members involved in developing, administering, or grading the exams include but are not limited to

- questions appearing on the exam or under consideration,
- deliberation related to the exam process, and
- information related to the scoring of questions.

Members may also be asked to offer assistance with other CFA Institute programs, including but not limited to the CIPM and Investment Foundations programs. Members participating in any CFA Institute program should do so with the same level of integrity and confidentiality as is required of participation in the CFA Program.

Expressing an Opinion

Standard VII(A) does *not* cover expressing opinions regarding CFA Institute, the CFA Program, or other CFA Institute programs. Members and candidates are free to disagree and express their disagreement with CFA Institute on its policies, its procedures, or any advocacy positions taken by the organization. When expressing a personal opinion, a candidate is prohibited from disclosing content-specific information, including any actual exam question and the information as to subject matter covered or not covered in the exam.

Exhibit 2 Sample of CFA Program Testing Policies

Candidate Pledge	As a candidate in the CFA Program, I am obligated to follow Standard VII(A) of the CFA Institute Standards of Professional Conduct, which states that members and candidates must not engage in any conduct that compromises the reputation or integrity of CFA Institute or the CFA designation or the integrity, validity, or security of the CFA exam. ■ Prior to this exam, I have not given or received information regarding the content of this exam. During this exam, I will not give or receive any information regarding the content of this exam. ■ After this exam, I will not disclose **ANY** portion of this exam and I will not remove **ANY** exam materials from the testing room in original or copied form. I understand that all exam materials, including my answers, are the property of CFA Institute and will not be returned to me in any form. ■ I will follow **ALL** rules of the CFA Program as stated on the CFA Institute website and the back cover of the exam book. My violation of any rules of the CFA Program will result in CFA Institute voiding my exam results and may lead to suspension or termination of my candidacy in the CFA Program.

Application of the Standard

Example 1 (Sharing Exam Questions):

Travis Nero serves as a proctor for the administration of the CFA examination in his city. In the course of his service, he reviews a copy of the Level II exam on the evening prior to the exam's administration and provides information concerning the exam questions to two candidates who use it to prepare for the exam.

> *Comment*: Nero and the two candidates have violated Standard VII(A). By giving information about the exam questions to two candidates, Nero provided an unfair advantage to the two candidates and undermined the integrity and validity of the Level II exam as an accurate measure of the knowledge, skills, and abilities necessary to earn the right to use the CFA designation. By accepting the information, the candidates also compromised the integrity and validity of the Level II exam and undermined the ethical framework that is a key part of the designation.

Example 2 (Bringing Written Material into Exam Room):

Loren Sullivan is enrolled to take the Level II CFA examination. He has been having difficulty remembering a particular formula, so prior to entering the exam room, he writes the formula on the palm of his hand. During the afternoon section of the exam, a proctor notices Sullivan looking at the palm of his hand. She asks to see his hand and finds the formula.

> *Comment*: Because Sullivan wrote down information from the Candidate Body of Knowledge (CBOK) and took that written information into the exam room, his conduct compromised the validity of his exam performance and violated Standard VII(A). Sullivan's conduct was also in direct contradiction with the rules and regulations of the CFA Program, the Candidate Pledge, and the CFA Institute Code and Standards.

Example 3 (Writing after Exam Period End):

At the conclusion of the morning section of the Level I CFA examination, the proctors announce, "Stop writing now." John Davis has not completed the exam, so he continues to randomly fill in ovals on his answer sheet. A proctor approaches Davis's desk and reminds him that he should stop writing immediately. Davis, however, continues to complete the answer sheet. After the proctor asks him to stop writing two additional times, Davis finally puts down his pencil.

> *Comment*: By continuing to complete his exam after time was called, Davis has violated Standard VII(A). By continuing to write, Davis took an unfair advantage over other candidates, and his conduct compromised the validity of his exam performance. Additionally, by not heeding the proctor's repeated instructions, Davis violated the rules and regulations of the CFA Program.

Example 4 (Sharing Exam Content):

After completing Level II of the CFA exam, Annabelle Rossi posts on her blog about her experience. She posts the following: "Level II is complete! I think I did fairly well on the exam. It was really difficult, but fair. I think I did especially well on the derivatives questions. And there were tons of them! I think I counted 18! The ethics questions were really hard. I'm glad I spent so much time on the Code and Standards. I was surprised to see there were no questions at all about IPO allocations. I expected there to be a couple. Well, off to celebrate getting through it. See you tonight?"

> *Comment*: Rossi did not violate Standard VII(A) when she wrote about how difficult she found the exam or how well she thinks she may have done. By revealing portions of the CBOK covered on the exam and areas not covered, however, she did violate Standard VII(A) and the Candidate Pledge. Depending on the time frame in which the comments were posted, Rossi not only may have assisted future candidates but also may have provided an unfair advantage to candidates yet to sit for the same exam, thereby undermining the integrity and validity of the Level II exam.

Example 5 (Sharing Exam Content):

Level I candidate Etienne Gagne has been a frequent visitor to an internet forum designed specifically for CFA Program candidates. The week after completing the Level I examination, Gagne and several others begin a discussion thread on the forum about the most challenging questions and attempt to determine the correct answers.

> *Comment*: Gagne has violated Standard VII(A) by providing and soliciting confidential exam information, which compromises the integrity of the exam process and violates the Candidate Pledge. In trying to determine correct answers to specific questions, the group's discussion included question-specific details considered to be confidential to the CFA Program.

Example 6 (Sharing Exam Content):

CFA4Sure is a company that produces test-preparation materials for CFA Program candidates. Many candidates register for and use the company's products. The day after the CFA examination, CFA4Sure sends an e-mail to all its customers asking them to share with the company the hardest questions from the exam so that CFA4Sure can better prepare its customers for the next exam administration. Marisol Pena e-mails a summary of the questions she found most difficult on the exam.

> *Comment*: Pena has violated Standard VII(A) by disclosing a portion of the exam questions. The information provided is considered confidential until publicly released by CFA Institute. CFA4Sure is likely to use such feedback to refine its review materials for future candidates. Pena's sharing of the specific questions undermines the integrity of the exam while potentially making the exam easier for future candidates.
>
> If the CFA4Sure employees who participated in the solicitation of confidential CFA Program information are CFA Institute members or candidates, they also have violated Standard VII(A).

Example 7 (Discussion of Exam Grading Guidelines and Results):

Prior to participating in grading CFA examinations, Wesley Whitcomb is required to sign a CFA Institute Grader Agreement. As part of the Grader Agreement, Whitcomb agrees not to reveal or discuss the exam materials with anyone except CFA Institute staff or other graders. Several weeks after the conclusion of the CFA exam grading, Whitcomb tells several colleagues who are candidates in the CFA Program which question he graded. He also discusses the guideline answer and adds that few candidates scored well on the question.

> *Comment*: Whitcomb violated Standard VII(A) by breaking the Grader Agreement and disclosing information related to a specific question on the exam, which compromised the integrity of the exam process.

Example 8 (Compromising CFA Institute Integrity as a Volunteer):

Jose Ramirez is an investor-relations consultant for several small companies that are seeking greater exposure to investors. He is also the program chair for the CFA Institute society in the city where he works. Ramirez schedules only companies that are his clients to make presentations to the society and excludes other companies.

> *Comment*: Ramirez, by using his volunteer position at CFA Institute to benefit himself and his clients, compromises the reputation and integrity of CFA Institute and thus violates Standard VII(A).

Example 9 (Compromising CFA Institute Integrity as a Volunteer):

Marguerite Warrenski is a member of the CFA Institute GIPS Executive Committee, which oversees the creation, implementation, and revision of the GIPS standards. As a member of the Executive Committee, she has advance knowledge of confidential information regarding the GIPS standards, including any new or revised standards the committee is considering. She tells her clients that her Executive Committee membership will allow her to better assist her clients in keeping up with changes to the Standards and facilitating their compliance with the changes.

> *Comment*: Warrenski is using her association with the GIPS Executive Committee to promote her firm's services to clients and potential clients. In defining her volunteer position at CFA Institute as a strategic business advantage over competing firms and implying to clients that she would use confidential information to further their interests, Warrenski is compromising the reputation and integrity of CFA Institute and thus violating Standard VII(A). She may factually state her involvement with the Executive Committee but cannot infer any special advantage to her clients from such participation.

Standard VII(B) Reference to CFA Institute, the CFA Designation, and the CFA Program

When referring to CFA Institute, CFA Institute membership, the CFA designation, or candidacy in the CFA Program, Members and Candidates must not misrepresent or exaggerate the meaning or implications of membership in CFA Institute, holding the CFA designation, or candidacy in the CFA Program.

Guidance

Highlights:

- *CFA Institute Membership*
- *Using the CFA Designation*
- *Referring to Candidacy in the CFA Program*

Standard VII(B) is intended to prevent promotional efforts that make promises or guarantees that are tied to the CFA designation. Individuals must not exaggerate the meaning or implications of membership in CFA Institute, holding the CFA designation, or candidacy in the CFA Program.

Standard VII(B) is not intended to prohibit factual statements related to the positive benefit of earning the CFA designation. However, statements referring to CFA Institute, the CFA designation, or the CFA Program that overstate the competency of an individual or imply, either directly or indirectly, that superior performance can be expected from someone with the CFA designation are not allowed under the standard.

Statements that highlight or emphasize the commitment of CFA Institute members, CFA charterholders, and CFA candidates to ethical and professional conduct or mention the thoroughness and rigor of the CFA Program are appropriate. Members and candidates may make claims about the relative merits of CFA Institute, the CFA Program, or the Code and Standards as long as those statements are implicitly or explicitly stated as the opinion of the speaker. Statements that do not express opinions have to be supported by facts.

Standard VII(B) applies to any form of communication, including but not limited to communications made in electronic or written form (such as on firm letterhead, business cards, professional biographies, directory listings, printed advertising, firm brochures, or personal resumes) and oral statements made to the public, clients, or prospects.

CFA Institute Membership

The term "CFA Institute member" refers to "regular" and "affiliate" members of CFA Institute who have met the membership requirements as defined in the CFA Institute Bylaws. Once accepted as a CFA Institute member, the member must satisfy the following requirements to maintain his or her status:

- remit annually to CFA Institute a completed Professional Conduct Statement, which renews the commitment to abide by the requirements of the Code and Standards and the CFA Institute Professional Conduct Program, and
- pay applicable CFA Institute membership dues on an annual basis.

If a CFA Institute member fails to meet any of these requirements, the individual is no longer considered an active member. Until membership is reactivated, individuals must not present themselves to others as active members. They may state, however, that they were CFA Institute members in the past or refer to the years when their membership was active.

Using the CFA Designation

Those who have earned the right to use the Chartered Financial Analyst designation are encouraged to do so but only in a manner that does not misrepresent or exaggerate the meaning or implications of the designation. The use of the designation may be accompanied by an accurate explanation of the requirements that have been met to earn the right to use the designation.

"CFA charterholders" are those individuals who have earned the right to use the CFA designation granted by CFA Institute. These people have satisfied certain requirements, including completion of the CFA Program and required years of acceptable work experience. Once granted the right to use the designation, individuals must also satisfy the CFA Institute membership requirements (see above) to maintain their right to use the designation.

If a CFA charterholder fails to meet any of the membership requirements, he or she forfeits the right to use the CFA designation. Until membership is reactivated, individuals must not present themselves to others as CFA charterholders. They may state, however, that they were charterholders in the past.

Given the growing popularity of social media, where individuals may anonymously express their opinions, pseudonyms or online profile names created to hide a member's identity should not be tagged with the CFA designation.

Use of the CFA designation by a CFA charterholder is governed by the terms and conditions of the annual Professional Conduct Statement Agreement, entered into between CFA Institute and its membership prior to commencement of use of the CFA designation and reaffirmed annually.

Referring to Candidacy in the CFA Program

Candidates in the CFA Program may refer to their participation in the CFA Program, but such references must clearly state that an individual is a *candidate* in the CFA Program and must not imply that the candidate has achieved any type of partial designation. A person is a candidate in the CFA Program if

- the person's application for registration in the CFA Program has been accepted by CFA Institute, as evidenced by issuance of a notice of acceptance, and the person is enrolled to sit for a specified examination or
- the registered person has sat for a specified examination but exam results have not yet been received.

If an individual is registered for the CFA Program but declines to sit for an exam or otherwise does not meet the definition of a candidate as described in the CFA Institute Bylaws, then that individual is no longer considered an active candidate. Once the person is enrolled to sit for a future examination, his or her CFA Program candidacy resumes.

CFA Program candidates must never state or imply that they have a partial designation as a result of passing one or more levels or cite an expected completion date of any level of the CFA Program. Final award of the charter is subject to meeting the CFA Program requirements and approval by the CFA Institute Board of Governors.

If a candidate passes each level of the exam in consecutive years and wants to state that he or she did so, that is not a violation of Standard VII(B) because it is a statement of fact. If the candidate then goes on to claim or imply superior ability by obtaining the designation in only three years, however, he or she is in violation of Standard VII(B).

Exhibit 3 provides examples of proper and improper references to the CFA designation.

Exhibit 3 Proper and Improper References

Proper References	Improper References
"Completion of the CFA Program has enhanced my portfolio management skills."	"CFA charterholders achieve better performance results."
"John Smith passed all three CFA Program examinations in three consecutive years."	"John Smith is among the elite, having passed all three CFA examinations in three consecutive attempts."
"The CFA designation is globally recognized and attests to a charterholder's success in a rigorous and comprehensive study program in the field of investment management and research analysis."	"As a CFA charterholder, I am the most qualified to manage client investments."
"The credibility that the CFA designation affords and the skills the CFA Program cultivates are key assets for my future career development."	"As a CFA charterholder, Jane White provides the best value in trade execution."
"I enrolled in the CFA Program to obtain the highest set of credentials in the global investment management industry."	"Enrolling as a candidate in the CFA Program ensures one of becoming better at valuing debt securities."
"I passed Level I of the CFA Program."	"CFA, Level II"
"I am a 2010 Level III candidate in the CFA Program."	"CFA, Expected 2011"
"I passed all three levels of the CFA Program and may be eligible for the CFA charter upon completion of the required work experience."	"CFA, Expected 2011" "John Smith, Charter Pending"

Recommended Procedures for Compliance

Misuse of a member's CFA designation or CFA candidacy or improper reference to it is common by those in a member's or candidate's firm who do not possess knowledge of the requirements of Standard VII(B). As an appropriate step to reduce this risk, members and candidates should disseminate written information about Standard VII(B) and the accompanying guidance to their firm's legal, compliance, public relations, and marketing departments (see www.cfainstitute.org).

For materials that refer to employees' affiliation with CFA Institute, members and candidates should encourage their firms to create templates that are approved by a central authority (such as the compliance department) as being consistent with Standard VII(B). This practice promotes consistency and accuracy in the firm of references to CFA Institute membership, the CFA designation, and CFA candidacy.

Application of the Standard

Example 1 (Passing Exams in Consecutive Years):

An advertisement for AZ Investment Advisors states that all the firm's principals are CFA charterholders and all passed the three examinations on their first attempt. The advertisement prominently links this fact to the notion that AZ's mutual funds have achieved superior performance.

> *Comment*: AZ may state that all principals passed the three examinations on the first try as long as this statement is true, but it must not be linked to performance or imply superior ability. Implying that (1) CFA charterholders achieve better investment results and (2) those who pass the exams on the first try may be more successful than those who do not violates Standard VII(B).

Example 2 (Right to Use CFA Designation):

Five years after receiving his CFA charter, Louis Vasseur resigns his position as an investment analyst and spends the next two years traveling abroad. Because he is not actively engaged in the investment profession, he does not file a completed Professional Conduct Statement with CFA Institute and does not pay his CFA Institute membership dues. At the conclusion of his travels, Vasseur becomes a self-employed analyst accepting assignments as an independent contractor. Without reinstating his CFA Institute membership by filing his Professional Conduct Statement and paying his dues, he prints business cards that display "CFA" after his name.

> *Comment*: Vasseur has violated Standard VII(B) because his right to use the CFA designation was suspended when he failed to file his Professional Conduct Statement and stopped paying dues. Therefore, he no longer is able to state or imply that he is an active CFA charterholder. When Vasseur files his Professional Conduct Statement, resumes paying CFA Institute dues to activate his membership, and completes the CFA Institute reinstatement procedures, he will be eligible to use the CFA designation.

Example 3 ("Retired" CFA Institute Membership Status):

After a 25-year career, James Simpson retires from his firm. Because he is not actively engaged in the investment profession, he does not file a completed Professional Conduct Statement with CFA Institute and does not pay his CFA Institute membership dues. Simpson designs a plain business card (without a corporate logo) to hand out to friends with his new contact details, and he continues to put "CFA" after his name.

> *Comment*: Simpson has violated Standard VII(B). Because he failed to file his Professional Conduct Statement and ceased paying dues, his membership has been suspended and he has given up the right to use the CFA designation. CFA Institute has procedures, however, for reclassifying a member and charterholder as "retired" and reducing the annual dues. If he wants to obtain retired status, he needs to file the appropriate paperwork with CFA Institute. When Simpson receives his notification from CFA Institute that his membership has been reclassified as retired and he resumes paying reduced dues, his membership will be reactivated and his right to use the CFA designation will be reinstated.

Example 4 (Stating Facts about CFA Designation and Program):

Rhonda Reese has been a CFA charterholder since 2000. In a conversation with a friend who is considering enrolling in the CFA Program, she states that she has learned a great deal from the CFA Program and that many firms require their employees to be CFA charterholders. She would recommend the CFA Program to anyone pursuing a career in investment management.

> *Comment*: Reese's comments comply with Standard VII(B). Her statements refer to facts: The CFA Program enhanced her knowledge, and many firms require the CFA designation for their investment professionals.

Example 5 (Order of Professional and Academic Designations):

Tatiana Prittima has earned both her CFA designation and a PhD in finance. She would like to cite both her accomplishments on her business card but is unsure of the proper method for doing so.

> *Comment*: The order of designations cited on such items as resumes and business cards is a matter of personal preference. Prittima is free to cite the CFA designation either before or after citing her PhD. Multiple designations must be separated by a comma.

Example 6 (Use of Fictitious Name):

Barry Glass is the lead quantitative analyst at CityCenter Hedge Fund. Glass is responsible for the development, maintenance, and enhancement of the proprietary models the fund uses to manage its investors' assets. Glass reads several high-level mathematical publications and blogs to stay informed on current developments. One blog, run by Expert CFA, presents some intriguing research that may benefit one of CityCenter's current models. Glass is under pressure from firm executives to improve the model's predictive abilities, and he incorporates the factors discussed in the online research. The updated output recommends several new investments to the fund's portfolio managers.

> *Comment*: "Expert CFA" has violated Standard VII(B) by using the CFA designation inappropriately. As with any research report, authorship of online comments must include the charterholder's full name along with any reference to the CFA designation.
>
> See also Standard V(A), which Glass has violated for guidance on diligence and reasonable basis.

PRACTICE PROBLEMS

Unless otherwise stated in the question, all individuals in the following questions are CFA Institute members or candidates in the CFA Program and, therefore, are subject to the CFA Institute Code of Ethics and Standards of Professional Conduct.

1 Smith, a research analyst with a brokerage firm, decides to change his recommendation for the common stock of Green Company, Inc., from a "buy" to a "sell." He mails this change in investment advice to all the firm's clients on Wednesday. The day after the mailing, a client calls with a buy order for 500 shares of Green Company. In this circumstance, Smith should:

  A Accept the order.

  B Advise the customer of the change in recommendation before accepting the order.

  C Not accept the order because it is contrary to the firm's recommendation.

2 Which statement about a manager's use of client brokerage commissions violates the Code and Standards?

  A A client may direct a manager to use that client's brokerage commissions to purchase goods and services for that client.

  B Client brokerage commissions should be used to benefit the client and should be commensurate with the value of the brokerage and research services received.

  C Client brokerage commissions may be directed to pay for the investment manager's operating expenses.

3 Jamison is a junior research analyst with Howard & Howard, a brokerage and investment banking firm. Howard & Howard's mergers and acquisitions department has represented the Britland Company in all of its acquisitions for the past 20 years. Two of Howard & Howard's senior officers are directors of various Britland subsidiaries. Jamison has been asked to write a research report on Britland. What is the best course of action for her to follow?

  A Jamison may write the report but must refrain from expressing any opinions because of the special relationships between the two companies.

  B Jamison should not write the report because the two Howard & Howard officers serve as directors for subsidiaries of Britland.

  C Jamison may write the report if she discloses the special relationships with the company in the report.

4 Which of the following statements clearly *conflicts* with the recommended procedures for compliance presented in the CFA Institute *Standards of Practice Handbook*?

  A Firms should disclose to clients the personal investing policies and procedures established for their employees.

  B Prior approval must be obtained for the personal investment transactions of all employees.

  C For confidentiality reasons, personal transactions and holdings should not be reported to employers unless mandated by regulatory organizations.

5 Bronson provides investment advice to the board of trustees of a private university endowment fund. The trustees have provided Bronson with the fund's financial information, including planned expenditures. Bronson receives a

phone call on Friday afternoon from Murdock, a prominent alumnus, requesting that Bronson fax him comprehensive financial information about the fund. According to Murdock, he has a potential contributor but needs the information that day to close the deal and cannot contact any of the trustees. Based on the CFA Institute Standards, Bronson should:

A Send Murdock the information because disclosure would benefit the client.

B Not send Murdock the information to preserve confidentiality.

C Send Murdock the information, provided Bronson promptly notifies the trustees.

6 Willier is the research analyst responsible for following Company X. All the information he has accumulated and documented suggests that the outlook for the company's new products is poor, so the stock should be rated a weak "hold." During lunch, however, Willier overhears a financial analyst from another firm whom he respects offer opinions that conflict with Willier's forecasts and expectations. Upon returning to his office, Willier releases a strong "buy" recommendation to the public. Willier:

A Violated the Standards by failing to distinguish between facts and opinions in his recommendation.

B Violated the Standards because he did not have a reasonable and adequate basis for his recommendation.

C Was in full compliance with the Standards.

7 An investment management firm has been hired by ETV Corporation to work on an additional public offering for the company. The firm's brokerage unit now has a "sell" recommendation on ETV, but the head of the investment banking department has asked the head of the brokerage unit to change the recommendation from "sell" to "buy." According to the Standards, the head of the brokerage unit would be permitted to:

A Increase the recommendation by no more than one increment (in this case, to a "hold" recommendation).

B Place the company on a restricted list and give only factual information about the company.

C Assign a new analyst to decide if the stock deserves a higher rating.

8 Albert and Tye, who recently started their own investment advisory business, have registered to take the Level III CFA examination. Albert's business card reads, "Judy Albert, CFA Level II." Tye has not put anything about the CFA designation on his business card, but promotional material that he designed for the business describes the CFA requirements and indicates that Tye participates in the CFA Program and has completed Levels I and II. According to the Standards:

A Albert has violated the Standards, but Tye has not.

B Tye has violated the Standards, but Albert has not.

C Both Albert and Tye have violated the Standards.

9 Scott works for a regional brokerage firm. He estimates that Walkton Industries will increase its dividend by US$1.50 a share during the next year. He realizes that this increase is contingent on pending legislation that would, if enacted, give Walkton a substantial tax break. The US representative for Walkton's home district has told Scott that, although she is lobbying hard for the bill and prospects for its passage are favorable, concern of the US Congress over the federal deficit could cause the tax bill to be voted down. Walkton Industries has not made any statements about a change in dividend policy. Scott writes in his research report, "We expect Walkton's stock price to rise by at least US$8.00 a share by the end of the year because the dividend will increase by US$1.50 a share. Investors buying the stock at the current time should expect to realize a total return of at least 15% on the stock." According to the Standards:

A Scott violated the Standards because he used material inside information.

B Scott violated the Standards because he failed to separate opinion from fact.

C Scott violated the Standards by basing his research on uncertain predictions of future government action.

10 Which one of the following actions will help to ensure the fair treatment of brokerage firm clients when a new investment recommendation is made?

A Informing all people in the firm in advance that a recommendation is to be disseminated.

B Distributing recommendations to institutional clients prior to individual accounts.

C Minimizing the time between the decision and the dissemination of a recommendation.

11 The mosaic theory holds that an analyst:

A Violates the Code and Standards if the analyst fails to have knowledge of and comply with applicable laws.

B Can use material public information and nonmaterial nonpublic information in the analyst's analysis.

C Should use all available and relevant information in support of an investment recommendation.

12 Jurgen is a portfolio manager. One of her firm's clients has told Jurgen that he will compensate her beyond the compensation provided by her firm on the basis of the capital appreciation of his portfolio each year. Jurgen should:

A Turn down the additional compensation because it will result in conflicts with the interests of other clients' accounts.

B Turn down the additional compensation because it will create undue pressure on her to achieve strong short-term performance.

C Obtain permission from her employer prior to accepting the compensation arrangement.

13 One of the discretionary accounts managed by Farnsworth is the Jones Corporation employee profit-sharing plan. Jones, the company president, recently asked Farnsworth to vote the shares in the profit-sharing plan in favor of the slate of directors nominated by Jones Corporation and against the directors sponsored by a dissident stockholder group. Farnsworth does not want to lose this account because he directs all the account's trades to a brokerage firm that provides Farnsworth with useful information about tax-free investments. Although this information is not of value in managing the Jones Corporation account, it does help in managing several other accounts. The brokerage firm providing this information also offers the lowest commissions for trades and provides best execution. Farnsworth investigates the director issue, concludes that the management-nominated slate is better for the long-run performance of the company than the dissident group's slate, and votes accordingly. Farnsworth:

A Violated the Standards in voting the shares in the manner requested by Jones but not in directing trades to the brokerage firm.

B Did not violate the Standards in voting the shares in the manner requested by Jones or in directing trades to the brokerage firm.

C Violated the Standards in directing trades to the brokerage firm but not in voting the shares as requested by Jones.

14 Brown works for an investment counseling firm. Green, a new client of the firm, is meeting with Brown for the first time. Green used another counseling firm for financial advice for years, but she has switched her account to Brown's firm. After spending a few minutes getting acquainted, Brown explains to Green that she has discovered a highly undervalued stock that offers large potential gains. She recommends that Green purchase the stock. Brown has committed a violation of the Standards. What should she have done differently?

A Brown should have determined Green's needs, objectives, and tolerance for risk before making a recommendation of any type of security.

B Brown should have thoroughly explained the characteristics of the company to Green, including the characteristics of the industry in which the company operates.

C Brown should have explained her qualifications, including her education, training, and experience and the meaning of the CFA designation.

15 Grey recommends the purchase of a mutual fund that invests solely in long-term US Treasury bonds. He makes the following statements to his clients:

I. "The payment of the bonds is guaranteed by the US government; therefore, the default risk of the bonds is virtually zero."

II. "If you invest in the mutual fund, you will earn a 10% rate of return each year for the next several years based on historical performance of the market."

Did Grey's statements violate the CFA Institute Code and Standards?

A Neither statement violated the Code and Standards.

B Only statement I violated the Code and Standards.

C Only statement II violated the Code and Standards.

16 Anderb, a portfolio manager for XYZ Investment Management Company—a registered investment organization that advises investment firms and private accounts—was promoted to that position three years ago. Bates, her supervisor, is responsible for reviewing Anderb's portfolio account transactions and her required monthly reports of personal stock transactions. Anderb has been using Jonelli, a broker, almost exclusively for brokerage transactions for the portfolio account. For securities in which Jonelli's firm makes a market, Jonelli has been giving Anderb lower prices for personal purchases and higher prices for personal sales than Jonelli gives to Anderb's portfolio accounts and other investors. Anderb has been filing monthly reports with Bates only for those months in which she has no personal transactions, which is about every fourth month. Which of the following is *most likely* to be a violation of the Code and Standards?

A Anderb failed to disclose to her employer her personal transactions.

B Anderb owned the same securities as those of her clients.

C Bates allowed Anderb to use Jonelli as her broker for personal trades.

17 Which of the following is a correct statement of a member's or candidate's duty under the Code and Standards?

A In the absence of specific applicable law or other regulatory requirements, the Code and Standards govern the member's or candidate's actions.

B A member or candidate is required to comply only with applicable local laws, rules, regulations, or customs, even though the Code and Standards may impose a higher degree of responsibility or a higher duty on the member or candidate.

C A member or candidate who trades securities in a securities market where no applicable local laws or stock exchange rules regulate the use of material nonpublic information may take investment action based on material nonpublic information.

18 Ward is scheduled to visit the corporate headquarters of Evans Industries. Ward expects to use the information he obtains there to complete his research report on Evans stock. Ward learns that Evans plans to pay all of Ward's expenses for the trip, including costs of meals, hotel room, and air transportation. Which of the following actions would be the *best* course for Ward to take under the Code and Standards?

A Accept the expense-paid trip and write an objective report.

B Pay for all travel expenses, including costs of meals and incidental items.

C Accept the expense-paid trip but disclose the value of the services accepted in the report.

19 Which of the following statements is *correct* under the Code and Standards?

A CFA Institute members and candidates are prohibited from undertaking independent practice in competition with their employer.

B Written consent from the employer is necessary to permit independent practice that could result in compensation or other benefits in competition with a member's or candidate's employer.

C Members and candidates are prohibited from making arrangements or preparations to go into a competitive business before terminating their relationship with their employer.

20 Smith is a financial analyst with XYZ Brokerage Firm. She is preparing a purchase recommendation on JNI Corporation. Which of the following situations is *most likely* to represent a conflict of interest for Smith that would have to be disclosed?

A Smith frequently purchases items produced by JNI.

B XYZ holds for its own account a substantial common stock position in JNI.

C Smith's brother-in-law is a supplier to JNI.

21 Michelieu tells a prospective client, "I may not have a long-term track record yet, but I'm sure that you'll be very pleased with my recommendations and service. In the three years that I've been in the business, my equity-oriented clients have averaged a total return of more than 26% a year." The statement is true, but Michelieu only has a few clients, and one of his clients took a large position in a penny stock (against Michelieu's advice) and realized a huge gain. This large return caused the average of all of Michelieu's clients to exceed 26% a year. Without this one investment, the average gain would have been 8% a year. Has Michelieu violated the Standards?

A No, because Michelieu is not promising that he can earn a 26% return in the future.

B No, because the statement is a true and accurate description of Michelieu's track record.

C Yes, because the statement misrepresents Michelieu's track record.

22 An investment banking department of a brokerage firm often receives material nonpublic information that could have considerable value if used in advising the firm's brokerage clients. In order to conform to the Code and Standards, which one of the following is the best policy for the brokerage firm?

A Permanently prohibit both "buy" and "sell" recommendations of the stocks of clients of the investment banking department.

B Establish physical and informational barriers within the firm to prevent the exchange of information between the investment banking and brokerage operations.

C Monitor the exchange of information between the investment banking department and the brokerage operation.

23 Stewart has been hired by Goodner Industries, Inc., to manage its pension fund. Stewart's duty of loyalty, prudence, and care is owed to:

A The management of Goodner.

B The participants and beneficiaries of Goodner's pension plan.

C The shareholders of Goodner.

24 Which of the following statements is a stated purpose of disclosure in Standard VI(C)–Referral Fees?

A Disclosure will allow the client to request discounted service fees.

B Disclosure will help the client evaluate any possible partiality shown in the recommendation of services.

C Disclosure means advising a prospective client about the referral arrangement once a formal client relationship has been established.

25 Rose, a portfolio manager for a local investment advisory firm, is planning to sell a portion of his personal investment portfolio to cover the costs of his child's academic tuition. Rose wants to sell a portion of his holdings in Household Products, but his firm recently upgraded the stock to "strong buy." Which of the following describes Rose's options under the Code and Standards?

A Based on his firm's "buy" recommendation, Rose cannot sell the shares because he would be improperly prospering from the inflated recommendation.

B Rose is free to sell his personal holdings once his firm is properly informed of his intentions.

C Rose can sell his personal holdings but only when a client of the firm places an order to buy shares of Household.

26 A former hedge fund manager, Jackman, has decided to launch a new private wealth management firm. From his prior experiences, he believes the new firm needs to achieve US$1 million in assets under management in the first year. Jackman offers a $10,000 incentive to any adviser who joins his firm with the minimum of $200,000 in committed investments. Jackman places notice of the opening on several industry web portals and career search sites. Which of the following is *correct* according to the Code and Standards?

A A member or candidate is eligible for the new position and incentive if he or she can arrange for enough current clients to switch to the new firm and if the member or candidate discloses the incentive fee.

B A member or candidate may not accept employment with the new firm because Jackman's incentive offer violates the Code and Standards.

- **C** A member or candidate is not eligible for the new position unless he or she is currently unemployed because soliciting the clients of the member's or candidate's current employer is prohibited.

27 Carter works for Invest Today, a local asset management firm. A broker that provides Carter with proprietary research through client brokerage arrangements is offering a new trading service. The broker is offering low-fee, execution-only trades to complement its traditional full-service, execution-and-research trades. To entice Carter and other asset managers to send additional business its way, the broker will apply the commissions paid on the new service toward satisfying the brokerage commitment of the prior full-service arrangements. Carter has always been satisfied with the execution provided on the full-service trades, and the new low-fee trades are comparable to the fees of other brokers currently used for the accounts that prohibit soft dollar arrangements.

- **A** Carter can trade for his accounts that prohibit soft dollar arrangements under the new low-fee trading scheme.
- **B** Carter cannot use the new trading scheme because the commissions are prohibited by the soft dollar restrictions of the accounts.
- **C** Carter should trade only through the new low-fee scheme and should increase his trading volume to meet his required commission commitment.

28 Rule has worked as a portfolio manager for a large investment management firm for the past 10 years. Rule earned his CFA charter last year and has decided to open his own investment management firm. After leaving his current employer, Rule creates some marketing material for his new firm. He states in the material, "In earning the CFA charter, a highly regarded credential in the investment management industry, I further enhanced the portfolio management skills learned during my professional career. While completing the examination process in three consecutive years, I consistently received the highest possible scores on the topics of Ethics, Alternative Investments, and Portfolio Management." Has Rule violated Standard VII(B)–Reference to CFA Institute, the CFA Designation, and the CFA Program in his marketing material?

- **A** Rule violated Standard VII(B) in stating that he completed the exams in three consecutive years.
- **B** Rule violated Standard VII(B) in stating that he received the highest scores in the topics of Ethics, Alternative Investments, and Portfolio Management.
- **C** Rule did not violate Standard VII(B).

29 Stafford is a portfolio manager for a specialized real estate mutual fund. Her firm clearly describes in the fund's prospectus its soft dollar policies. Stafford decides that entering the CFA Program will enhance her investment decision-making skill and decides to use the fund's soft dollar account to pay the registration and exam fees for the CFA Program. Which of the following statements is *most likely* correct?

- **A** Stafford did not violate the Code and Standards because the prospectus informed investors of the fund's soft dollar policies.
- **B** Stafford violated the Code and Standards because improving her investment skills is not a reasonable use of the soft dollar account.
- **C** Stafford violated the Code and Standards because the CFA Program does not meet the definition of research allowed to be purchased with brokerage commissions.

30 Long has been asked to be the keynote speaker at an upcoming investment conference. The event is being hosted by one of the third-party investment managers currently used by his pension fund. The manager offers to cover all conference and travel costs for Long and make the conference registrations free for three additional members of his investment management team. To ensure that the conference obtains the best speakers, the host firm has arranged for an exclusive golf outing for the day following the conference on a local championship-caliber course. Which of the following is *least likely* to violate Standard I(B)?

A Long may accept only the offer to have his conference-related expenses paid by the host firm.

B Long may accept the offer to have his conference-related expenses paid and may attend the exclusive golf outing at the expense of the hosting firm.

C Long may accept the entire package of incentives offered to speak at this conference.

31 Andrews, a private wealth manager, is conducting interviews for a new research analyst for his firm. One of the candidates is Wright, an analyst with a local investment bank. During the interview, while Wright is describing his analytical skills, he mentions a current merger in which his firm is acting as the adviser. Andrews has heard rumors of a possible merger between the two companies, but no releases have been made by the companies concerned. Which of the following actions by Andrews is *least likely* a violation of the Code and Standards?

A Waiting until the next day before trading on the information to allow time for it to become public.

B Notifying all investment managers in his firm of the new information so none of their clients are disadvantaged.

C Placing the securities mentioned as part of the merger on the firm's restricted trading list.

32 Pietro, president of Local Bank, has hired the bank's market maker, Vogt, to seek a merger partner. Local is currently not listed on a stock exchange and has not reported that it is seeking strategic alternatives. Vogt has discussed the possibility of a merger with several firms, but they have all decided to wait until after the next period's financial data are available. The potential buyers believe the results will be worse than the results of prior periods and will allow them to pay less for Local Bank.

Pietro wants to increase the likelihood of structuring a merger deal quickly. Which of the following actions would *most likely* be a violation of the Code and Standards?

A Pietro could instruct Local Bank to issue a press release announcing that it has retained Vogt to find a merger partner.

B Pietro could place a buy order for 2,000 shares (or four times the average weekly volume) through Vogt for his personal account.

C After confirming with Local's chief financial officer, Pietro could instruct Local to issue a press release reaffirming the firm's prior announced earnings guidance for the full fiscal year.

33 ABC Investment Management acquires a new, very large account with two concentrated positions. The firm's current policy is to add new accounts for the purpose of performance calculation after the first full month of management. Cupp is responsible for calculating the firm's performance returns. Before the end of the initial month, Cupp notices that one of the significant holdings of the new accounts is acquired by another company, causing the value of the investment to double. Because of this holding, Cupp decides to account for the new portfolio as of the date of transfer, thereby allowing ABC Investment to reap the positive impact of that month's portfolio return.

A Cupp did not violate the Code and Standards because the GIPS standards allow composites to be updated on the date of large external cash flows.

B Cupp did not violate the Code and Standards because companies are allowed to determine when to incorporate new accounts into their composite calculation.

C Cupp violated the Code and Standards because the inclusion of the new account produces an inaccurate calculation of the monthly results according to the firm's stated policies.

34 Cannan has been working from home on weekends and occasionally saves correspondence with clients and completed work on her home computer. Because of worsening market conditions, Cannan is one of several employees released by her firm. While Cannan is looking for a new job, she uses the files she saved at home to request letters of recommendation from former clients. She also provides to prospective clients some of the reports as examples of her abilities.

A Cannan violated the Code and Standards because she did not receive permission from her former employer to keep or use the files after her employment ended.

B Cannan did not violate the Code and Standards because the files were created and saved on her own time and computer.

C Cannan violated the Code and Standards because she is prohibited from saving files on her home computer.

35 Quinn sat for the Level III CFA exam this past weekend. He updates his resume with the following statement: "In finishing the CFA Program, I improved my skills related to researching investments and managing portfolios. I will be eligible for the CFA charter upon completion of the required work experience."

A Quinn violated the Code and Standards by claiming he improved his skills through the CFA Program.

B Quinn violated the Code and Standards by incorrectly stating that he is eligible for the CFA charter.

C Quinn did not violate the Code and Standards with his resume update.

36 During a round of golf, Rodriguez, chief financial officer of Mega Retail, mentions to Hart, a local investment adviser and long-time personal friend, that Mega is having an exceptional sales quarter. Rodriguez expects the results to be almost 10% above the current estimates. The next day, Hart initiates the purchase of a large stake in the local exchange-traded retail fund for her personal account.

A Hart violated the Code and Standards by investing in the exchange-traded fund that included Mega Retail.

B Hart did not violate the Code and Standards because she did not invest directly in securities of Mega Retail.

C Rodriguez did not violate the Code and Standards because the comments made to Hart were not intended to solicit an investment in Mega Retail.

37 Park is very frustrated after taking her Level II exam. While she was studying for the exam, to supplement the curriculum provided, she ordered and used study material from a third-party provider. Park believes the additional material focused her attention on specific topic areas that were not tested while ignoring other areas. She posts the following statement on the provider's discussion board: "I am very dissatisfied with your firm's CFA Program Level II material. I found the exam extremely difficult and myself unprepared for specific questions after using your product. How could your service provide such limited

instructional resources on the analysis of inventories and taxes when the exam had multiple questions about them? I will not recommend your products to other candidates."

A Park violated the Code and Standards by purchasing third-party review material.

B Park violated the Code and Standards by providing her opinion on the difficulty of the exam.

C Park violated the Code and Standards by providing specific information on topics tested on the exam.

38 Paper was recently terminated as one of a team of five managers of an equity fund. The fund had two value-focused managers and terminated one of them to reduce costs. In a letter sent to prospective employers, Paper presents, with written permission of the firm, the performance history of the fund to demonstrate his past success.

A Paper did not violate the Code and Standards.

B Paper violated the Code and Standards by claiming the performance of the entire fund as his own.

C Paper violated the Code and Standards by including the historical results of his prior employer.

39 Townsend was recently appointed to the board of directors of a youth golf program that is the local chapter of a national not-for-profit organization. The program is beginning a new fund-raising campaign to expand the number of annual scholarships it provides. Townsend believes many of her clients make annual donations to charity. The next week in her regular newsletter to all clients, she includes a small section discussing the fund-raising campaign and her position on the organization's board.

A Townsend did not violate the Code and Standards.

B Townsend violated the Code and Standards by soliciting donations from her clients through the newsletter.

C Townsend violated the Code and Standards by not getting approval of the organization before soliciting her clients.

The following information relates to Questions 40–45

Anne Boswin, CFA, is a senior fixed-income analyst at Greenfield Financial Corporation. Boswin develops financial models for predicting changes in bond prices. On the premise that bonds of firms targeted for leveraged buyouts (LBOs) often decline in value, Boswin develops a model to predict which firms are likely to be subject to LBOs.

Boswin works closely with another analyst, Robert Acertado, CFA. Acertado uses Boswin's model frequently to identify potential LBO targets for further research. Using the model and his extensive research skills, Acertado makes timely investment recommendations and develops a strong track record.

Based on this record, Acertado receives an employment offer from the asset management division of Smith & Garner Investments, Inc., a diversified financial services firm. With Boswin's consent, Acertado downloads the model before leaving Greenfield.

At Smith & Garner, Acertado presents the idea of predicting LBO targets as a way to identify bonds that might decline in value and thus be good sell recommendations. After Acertado walks his boss through the model, the supervisor comments, "I like your idea and your model, Robert. I can see that we made the right decision in hiring you."

Because Smith & Garner has both an Investment Banking (IB) and Asset Management (AM) division, Acertado's supervisor reminds him that he should not attempt to contact or engage in conversation with anyone from the Investment Banking division. The supervisor also directs him to eat in the East end of the company cafeteria. "The West end is reserved for the IB folks, and you may laugh at this, but we actually put up a wall between the two ends. If anyone were to accuse us of not having a firewall, we could actually point to it!" Robert's supervisor also tells him, "There should be absolutely no conversation about divisional business while in the hall and elevator that serves as a common access to the cafeteria for both divisions. We are very strict about this."

The following week, Acertado is riding alone in the elevator when it stops on an IB floor. As the doors begin to slide open, Acertado hears a voice whispering, "I am so pleased that we were able to put the financing together for Country Industries. I was concerned because the leverage will go to 80%—higher than our typical deal." As soon as the doors open enough to reveal that the elevator is occupied, all conversation stops.

Late that afternoon, Acertado uses the LBO model to measure the probability of Country Industries receiving an LBO offer. According to the model, the probability is 62%—slightly more than the 60% Acertado generally requires before conducting additional research. It is late in the afternoon and Acertado has little time to research the matter fully before the end of the trading day. He checks his inputs to the model. In the interest of time, Acertado immediately recommends selling Country Industries' senior bonds held in any long-only accounts. He also recommends establishing positions in derivatives contracts that will benefit from a decline in the value of Country Industries' bonds.

The next morning, after the firm has established the derivatives positions he recommended, Acertado calls Boswin. Knowing that his former associate will be preparing Greenfield's monthly newsletter, he tells her, "I ran Country Industries through your model and I think it is likely that they will receive an LBO offer." Acertado explains some of the inputs he used in the model. At the conclusion of the conversation Boswin responds, "You may be right. Country Industries sounds like a possible LBO candidate, and thus, a sell rating on their senior bonds would be in order. If I'm lucky, I can finish researching the issue in time to include the recommendation in the upcoming newsletter. Thanks. It was good talking with you, Robert."

After the conversation with Acertado, Boswin quickly runs Country Industries through the model. Based on her inputs, the model calculates that the probability of an LBO is 40%—not enough, in Boswin's opinion, to justify further research. She wonders if there is a discrepancy between her inputs and Acertado's. Pressed for time, Boswin resumes her work on the upcoming newsletter rather than investigating the matter.

Acertado soon begins searching the internet for information on companies that the model predicts have more than a 60% probability of an LBO offer. He scours blogs and company websites looking for signs of a potential offer. He uses evidence of rumored offers in developing sell recommendations on various corporations' bonds.

40 When downloading the model from Greenfield Financial Corporation, does Acertado violate any CFA Institute Standards of Practice and Professional Conduct?

- **A** No.
- **B** Yes, because he does not have written permission from Boswin.
- **C** Yes, because he does not have permission from Greenfield Financial Corporation.

41 When using the model at Smith & Garner, Acertado is *least likely* to violate the Standard relating to:

A misrepresentation.

B loyalty to employer.

C material nonpublic information.

42 When making the recommendation regarding Country Industries, does Acertado violate any CFA Institute Standards?

A No.

B Yes, relating to diligence and reasonable basis.

C Yes, relating to material nonpublic information.

43 In his phone conversation with Boswin, Acertado *least likely* violates the CFA Institute Standard relating to:

A suitability.

B integrity of capital markets.

C preservation of confidentiality.

44 When analyzing the probability of an LBO of Country Industries, does Boswin violate any CFA Institute Standards?

A No.

B Yes, relating to independence and objectivity.

C Yes, relating to diligence and reasonable basis.

45 When searching blogs, does Acertado violate any CFA Institute Standards?

A No.

B Yes, because he misuses company resources.

C Yes, because he seeks inside information on the blogs.

The following information relates to Questions 46–50

Erik Brecksen, CFA, a portfolio manager at Apfelbaum Kapital, recently recruited Hans Grohl, a CFA candidate and recent MBA graduate from a top university with excellent quantitative analysis skills. Apfelbaum Kapital stresses "top-down" fundamental analysis and uses a team approach to investment management. The firm's investment professionals, all of whom are CFA charterholders or candidates, attend weekly investment committee meetings. At the meetings, analysts responsible for different industrial sectors present their research and recommendations. Following each presentation, the investment committee, consisting of senior portfolio managers, questions the analyst about the recommendation. If the majority of the committee agrees with the recommendation, the recommendation is approved and the stock is placed on a restricted list while the firm executes the necessary trades.

Apfelbaum considers its research proprietary. It is intended for the sole use of its investment professionals and is not distributed outside the firm. The names of all the investment personnel associated with the sector or investment class are listed on each research report regardless of their actual level of contribution to the report.

On Grohl's first day of work, Brecksen assigns him responsibility for a company that Brecksen covered previously. He provides Grohl with his past research including all of his files and reports. Brecksen instructs Grohl to report back when he has finished his research and is ready to submit his own research report on the company.

Grohl reads Brecksen's old reports before studying the financial statements of the company and its competitors. Taking advantage of his quantitative analysis skills, Grohl then conducts a detailed multi-factor analysis. Afterward, he produces a written buy recommendation using Brecksen's old research reports as a guide for format and submits a draft to Brecksen for review.

Brecksen reviews the work and indicates that he is not familiar with multi-factor analysis. He tells Grohl that he agrees with the buy recommendation, but instructs Grohl to omit the multi-factor analysis from the report. Grohl attempts to defend his research methodology, but is interrupted when Brecksen accepts a phone call. Grohl follows Brecksen's instructions and removes all mention of the multi-factor analysis from the final report. Brecksen presents the completed report at the weekly meeting with both his and Grohl's names listed on the document. After Brecksen's initial presentation, the committee turns to Grohl and asks about his research. Grohl takes the opportunity to mention the multi-factor analysis. Satisfied, the committee votes in favor of the recommendation and congratulates Grohl on his work.

Ottie Zardt, CFA, has worked as a real estate analyst for Apfelbaum for the past 18 months. A new independent rating service has determined that Zardt's recommendations have resulted in an excess return of 12% versus the industry's return of 2.7% for the past twelve months. After learning about the rating service, Zardt immediately updates the promotional material he is preparing for distribution at an upcoming industry conference. He includes a reference to the rating service and quotes its returns results and other information. Before distributing the material at the conference, he adds a footnote stating "Past performance is no guarantee of future success."

46 When preparing the initial draft for Brecksen's review, does Grohl violate any CFA Standards?

- **A** No.
- **B** Yes, because he used Brecksen's research reports without permission.
- **C** Yes, because he did not use reasonable judgment in identifying which factors were important to the analysis.

47 When instructing Grohl to eliminate the multi-factor analysis from the research report, does Brecksen violate any CFA Standards?

- **A** No.
- **B** Yes, relating to record retention.
- **C** Yes, relating to diligence and reasonable basis.

48 When removing the multi-factor analysis from his research report, does Grohl violate any CFA Standards?

- **A** No.
- **B** Yes, because he no longer has a reasonable basis for his recommendation.
- **C** Yes, because he is required to make full and fair disclosure of all relevant information.

49 When listing their names on the research report, do Brecksen and Grohl violate any CFA Standards?

- **A** No.
- **B** Yes, because Brecksen misrepresents his authorship.
- **C** Yes, because Grohl should dissociate from the report.

50 When distributing the material at the industry conference, does Zardt violate any CFA Standards?

A No.

B Yes, because Zardt does not verify the accuracy of the information.

C Yes, because analysts cannot claim performance or promote the accuracy of their recommendations.

The following information relates to Questions 51–56

Samuel Telline, CFA, is a portfolio manager at Aiklin Investments with discretionary authority over all of his accounts. One of his clients, Alan Caper, Chief Executive Officer (CEO) of Ellipse Manufacturing, invites Telline to lunch.

At the restaurant, the CEO reveals the reason for the lunch. "As you know Reinhold Partners has made an unsolicited cash offer for all outstanding shares of Ellipse Manufacturing. Reinhold has made it clear that I will not be CEO if they are successful. I can assure you that our shareholders will be better off in the long term if I'm in charge." Caper then shows Telline his projections for a new plan designed to boost both sales and operating margins.

"I know that your firm is the trustee for our firm's Employee Stock Ownership Plan (ESOP). I hope that the trustee will vote in the best interest of our shareholders—and that would be a vote against the takeover offer."

After looking through Caper's business plans, Telline says, "This plan looks good. I will recommend that the trustee vote against the offer."

Caper responds, "I remember my friend Karen Leighton telling me that the Leighton Family's Trust is managed by your firm. Perhaps the trustee could vote those shares against the acquisition as well. Karen Leighton is a close friend. I am sure that she would agree."

Telline responds, "The Family Trust is no longer managed by Aiklin." He adds, "I understand that the Trust is very conservatively managed. I doubt it that it would have holdings in Ellipse Manufacturing." Telline does not mention that although the Family Trust has changed investment managers, Karen Leighton remains an important client at Aiklin with significant personal holdings in Ellipse.

After lunch, Telline meets with Sydney Brown, CFA, trustee of the Ellipse ESOP. He shows her Caper's plan for improvements. "I think the plan is a good one and Caper is one of the firm's most profitable accounts. We don't want to lose him." Brown agrees to analyze the plan. After thoroughly analyzing both the plan and the takeover offer, Brown concludes that the takeover offer is best for the shareholders in the ESOP and votes the plan's shares in favor of the takeover offer.

A few months later the acquisition of Ellipse by Reinhold Partners is completed. Caper again meets Telline for lunch. "I received a generous severance package and I'm counting on you to manage my money well for me. While we are on the subject, I would like to be more aggressive with my portfolio. With my severance package, I can take additional risk." Telline and Caper discuss his current financial situation, risk tolerance, and financial objectives throughout lunch. Telline agrees to adjust Caper's investment policy statement (IPS) to reflect his greater appetite for risk and his increased wealth.

Back at the office, Telline realizes that with the severance package, Caper is now his wealthiest client. He also realizes that Caper's increased appetite for risk gives him a risk profile similar to that of another client. He pulls a copy of the other client's investment policy statement (IPS) and reviews it quickly before realizing that the two clients have very different tax situations. Telline quickly revises Caper's IPS to reflect the changes in his financial situation. He uses the other client's IPS as a reference when revising the section relating to Caper's risk tolerance. He then files the revised IPS in Caper's file.

The following week, an Aiklin analyst issues a buy recommendation on a small technology company with a promising software product. Telline reads the report carefully and concludes it would be suitable under Caper's new IPS. Telline places an order for 10,000 shares in Caper's account and then calls Caper to discuss the stock in more detail. Telline does not purchase the stock for any other clients. Although the one client has the same risk profile as Caper, that client does not have cash available in his account and Telline determines that selling existing holdings does not make sense.

In a subsequent telephone conversation, Caper expresses his lingering anger over the takeover. "You didn't do enough to persuade Aiklin's clients to vote against the takeover. Maybe I should look for an investment manager who is more loyal." Telline tries to calm Caper but is unsuccessful. In an attempt to change the topic of conversation, Telline states, "The firm was just notified of our allocation of a long-awaited IPO. Your account should receive a significant allocation. I would hate to see you lose out by moving your account." Caper seems mollified and concludes the phone call, "I look forward to a long-term relationship with you and your firm."

Aiklin distributes a copy of its firm policies regarding IPO allocations to all clients annually. According to the policy, Aiklin allocates IPO shares to each investment manager and each manager has responsibility for allocating shares to accounts for which the IPO is suitable. The statement also discloses that Aiklin offers different levels of service for different fees.

After carefully reviewing the proposed IPO and his client accounts, Telline determines that the IPO is suitable for 11 clients including Caper. Because the deal is oversubscribed, he receives only half of the shares he expected. Telline directs 50% of his allocation to Caper's account and divides the remaining 50% between the other ten accounts, each with a value equal to half of Caper's account.

51 When discussing the Leighton Family Trust, does Telline violate any CFA Institute Standards of Professional Conduct?

- **A** No.
- **B** Yes, relating to duties to clients.
- **C** Yes, relating to misrepresentation.

52 When deciding how to vote the ESOP shares, does Brown violate any CFA Institute Standards?

- **A** No.
- **B** Yes, relating to loyalty, prudence, and care.
- **C** Yes, relating to diligence and reasonable basis.

53 The Standard *least likely* to provide guidance for Telline when working with the clients' investment policy statements would be the Standard relating to:

- **A** suitability.
- **B** fair dealing.
- **C** loyalty, prudence, and care.

54 Does Telline violate any CFA Institute Standards when he places the buy order for shares in the technology company for Caper's account?

A No.

B Yes, relating to fair dealing.

C Yes, relating to diligence and reasonable basis.

55 Is Aiklin's policy with respect to IPO allocations consistent with required and recommended CFA Institute Standards?

A Yes.

B No, because the IPO policy disadvantages certain clients.

C No, because the different levels of service disadvantage certain clients.

56 Does Telline violate any CFA Institute Standards in his allocation of IPO shares to Caper's account?

A No.

B Yes, because the IPO is not suitable for Caper.

C Yes, because he does not treat all his clients fairly.

The following information relates to Questions 57–62

Adam Craw, CFA, is chief executive officer (CEO) of Crawfood, a European private equity firm specializing in food retailers. The retail food industry has been consolidating during the past two years as private equity funds have closed numerous deals and taken many companies private.

Crawfood recently hired Lillian Voser, a CFA Level II candidate, as a controller. On Voser's first day of work, the head of personnel informs her that by signing the employment contract, Voser agrees to comply with the company's code of ethics and compliance manual. She hands Voser copies of the code and compliance manual without further comment. Voser spends the next hour reading both documents. An excerpt from the compliance manual appears in Exhibit 1.

Exhibit 1 Crawfood Company Compliance Manual Excerpts

1 Employees must not accept gifts, benefits, compensation, or consideration that competes with, or might reasonably be expected to create a conflict of interest with their employer's interest unless they obtain written consent from all parties involved.

2 Officers have responsibility for ensuring that their direct reports—that is, employees whom they directly supervise—adhere to applicable laws, rules, and regulations.

3 Employees in possession of material nonpublic information should make reasonable efforts to achieve public dissemination of the information if such actions would not breach a duty.

4 Employees shall not trade or cause others to trade in securities of food retailers that may be potential takeover targets of their employer.

When she enters her new office that afternoon, Voser finds a large gift basket sent by her sister. The card reads "Congratulations on your new position." The basket is filled with expensive high-quality food items from Greenhornfood—a local small, publicly-traded food retailer, which produces many delicatessen products under its own brand name.

During the next two weeks, Voser meets with all of Crawfood's upper management, including the CEO. In his office, Craw praises Voser's efforts to complete the CFA program. "The program is demanding, but it is worthwhile." Craw then explains his investment strategy for choosing Crawfood's acquisition targets. He points to a large map on the wall with multi-colored pins marking Crawfood's previous takeovers. The map shows acquisitions in all the major cities of Germany with one exception—the home of Crawfood headquarters. Craw remarks, "We are currently in talks for another purchase. Confidentiality prohibits me from discussing it any further, but you will hear more about it soon."

Introduced to Greenhornfood by her sister, Voser quickly becomes a loyal customer. She considers it the best food retailer in the vicinity and she frequently purchases its products.

The following week, the local newspaper features an article about Greenhornfood and its young founders. The article describes the company's loyal and growing customer base as well as its poor quarterly financial results. Voser notes that the stock has steadily declined during the past twelve months. She concludes that the company has an inexperienced management team, but its popular product line and loyal customer base make the company a potential acquisition target. Voser calls her sister and recommends that she purchase Greenhornfood shares because "it would be an attractive acquisition for a larger company." Based on Voser's recommendation, her sister buys €3,000 worth of shares.

During the following two weeks the stock price of Greenhornfood continues to decline. Voser's sister is uncertain of what she should do with her position. She seeks Voser's advice. Voser recommends that her sister wait another few days before making her decision and promises to analyze the situation in the meantime.

While walking by Craw's office the following day, Voser sees a document with Greenhornfood's distinctive logo and overhears the company's name through an open office door. That evening, Voser tells her sister, "with the price decline, the stock is even more attractive." She recommends that her sister increase her position. Based on her recommendation her sister buys an additional €3,000 worth of Greenhornfood shares.

One month later, Crawfood publicly announces the acquisition of Greenhornfood Company at a 20% premium to the previous day's closing price. Following the announcement, Voser's sister boasts about Voser's excellent recommendation and timing to her broker.

Regulatory authorities initiate an investigation into suspicious trading in Greenhornfood shares and options preceding the formal announcement of the acquisition. Craw receives a letter from regulatory authorities stating that he is the subject of a formal investigation into his professional conduct surrounding the acquisition. He learns from the compliance officer that Voser is also under investigation. The compliance officer provides no details and out of respect for Voser's privacy, Craw makes no inquiries.

The situation remains unchanged and the matter is still pending with regulatory authorities several months later when Craw receives his annual Professional Conduct Statement (PCS) from CFA Institute. He reviews the text asking "In the last two years, have you been . . . the subject of . . . any investigation . . . in which your professional conduct, in either a direct or supervisory capacity, was at issue?"

57 Are Excerpts 2 and 3 of Crawfood's compliance procedures consistent with the CFA Institute Standards of Professional Conduct?

A Yes.

B No, because Excerpt 2 applies only to officers and their direct reports.

C No, because Excerpt 3 does not require employees to achieve public dissemination.

58 According to the CFA Institute Standards, must Voser obtain permission from her supervisor before accepting the Greenhornfood gift basket?

A No.

B Yes, because the value of the basket is higher than €50.

C Yes, because consent is required by the company's compliance procedures.

59 When making her initial recommendation to purchase Greenhornfood company shares, Voser *most likely* violates the Standard relating to:

A loyalty to employer.

B integrity of capital markets.

C diligence and reasonable basis.

60 When recommending the purchase of additional Greenhornfood company shares, Voser *least likely* violates the Standard relating to:

A loyalty to employer.

B integrity of capital markets.

C diligence and reasonable basis.

61 Does Craw violate any CFA Institute Standards?

A No.

B Yes, because he passes material nonpublic information to Voser.

C Yes, because he does not make reasonable efforts to prevent violations of applicable law.

62 According to the CFA Standards, Craw must disclose to CFA Institute the investigation into:

A his conduct.

B Voser's conduct.

C neither his conduct nor Voser's conduct.

SOLUTIONS

1 The correct answer is B. This question involves Standard III(B)–Fair Dealing. Smith disseminated a change in the stock recommendation to his clients but then received a request contrary to that recommendation from a client who probably had not yet received the recommendation. Prior to executing the order, Smith should take additional steps to ensure that the customer has received the change of recommendation. Answer A is incorrect because the client placed the order prior to receiving the recommendation and, therefore, does not have the benefit of Smith's most recent recommendation. Answer C is also incorrect; simply because the client request is contrary to the firm's recommendation does not mean a member can override a direct request by a client. After Smith contacts the client to ensure that the client has received the changed recommendation, if the client still wants to place a buy order for the shares, Smith is obligated to comply with the client's directive.

2 The correct answer is C. This question involves Standard III(A)–Loyalty, Prudence, and Care and the specific topic of soft dollars or soft commissions. Answer C is the correct choice because client brokerage commissions may not be directed to pay for the investment manager's operating expenses. Answer B describes how members and candidates should determine how to use brokerage commissions—that is, if the use is in the best interests of clients and is commensurate with the value of the services provided. Answer A describes a practice that is commonly referred to as "directed brokerage." Because brokerage is an asset of the client and is used to benefit the client, not the manager, such practice does not violate a duty of loyalty to the client. Members and candidates are obligated in all situations to disclose to clients their practices in the use of client brokerage commissions.

3 The correct answer is C. This question involves Standard VI(A)–Disclosure of Conflicts. The question establishes a conflict of interest in which an analyst, Jamison, is asked to write a research report on a company that is a client of the analyst's employer. In addition, two directors of the company are senior officers of Jamison's employer. Both facts establish that there are conflicts of interest that must be disclosed by Jamison in her research report. Answer B is incorrect because an analyst is not prevented from writing a report simply because of the special relationship the analyst's employer has with the company as long as that relationship is disclosed. Answer A is incorrect because whether or not Jamison expresses any opinions in the report is irrelevant to her duty to disclose a conflict of interest. Not expressing opinions does not relieve the analyst of the responsibility to disclose the special relationships between the two companies.

4 The correct answer is C. This question asks about compliance procedures relating to personal investments of members and candidates. The statement in answer C clearly conflicts with the recommended procedures in the *Standards of Practice Handbook*. Employers should compare personal transactions of employees with those of clients on a regular basis regardless of the existence of a requirement by any regulatory organization. Such comparisons ensure that employees' personal trades do not conflict with their duty to their clients, and the comparisons can be conducted in a confidential manner. The statement in answer A does not conflict with the procedures in the *Handbook*. Disclosure of such policies will give full information to clients regarding potential conflicts of interest on the part of those entrusted to manage their money. Answer B is incorrect because firms are encouraged to establish policies whereby employees clear their personal holdings and transactions with their employers.

5 The correct answer is B. This question relates to Standard III(A)–Loyalty, Prudence, and Care and Standard III(E)–Preservation of Confidentiality. In this case, the member manages funds of a private endowment. Clients, who are, in this case, the trustees of the fund, must place some trust in members and candidates. Bronson cannot disclose confidential financial information to anyone without the permission of the fund, regardless of whether the disclosure may benefit the fund. Therefore, answer A is incorrect. Answer C is incorrect because Bronson must notify the fund and obtain the fund's permission before publicizing the information.

6 The correct answer is B. This question relates to Standard V(A)–Diligence and Reasonable Basis. The opinion of another financial analyst is not an adequate basis for Willier's action in changing the recommendation. Answer C is thus incorrect. So is answer A because, although it is true that members and candidates must distinguish between facts and opinions in recommendations, the question does not illustrate a violation of that nature. If the opinion overheard by Willier had sparked him to conduct additional research and investigation that justified a change of opinion, then a changed recommendation would be appropriate.

7 The correct answer is B. This question relates to Standard I(B)–Independence and Objectivity. When asked to change a recommendation on a company stock to gain business for the firm, the head of the brokerage unit must refuse in order to maintain his independence and objectivity in making recommendations. He must not yield to pressure by the firm's investment banking department. To avoid the appearance of a conflict of interest, the firm should discontinue issuing recommendations about the company. Answer A is incorrect; changing the recommendation in any manner that is contrary to the analyst's opinion violates the duty to maintain independence and objectivity. Answer C is incorrect because merely assigning a new analyst to decide whether the stock deserves a higher rating will not address the conflict of interest.

8 The correct answer is A. Standard VII(B)–Reference to CFA Institute, the CFA Designation, and the CFA Program is the subject of this question. The reference on Albert's business card implies that there is a "CFA Level II" designation; Tye merely indicates in promotional material that he is participating in the CFA Program and has completed Levels I and II. Candidates may not imply that there is some sort of partial designation earned after passing a level of the CFA exam. Therefore, Albert has violated Standard VII(B). Candidates may communicate that they are participating in the CFA Program, however, and may state the levels that they have completed. Therefore, Tye has not violated Standard VII(B).

9 The correct answer is B. This question relates to Standard V(B)–Communication with Clients and Prospective Clients. Scott has issued a research report stating that he expects the price of Walkton Industries stock to rise by US$8 a share "because the dividend will increase" by US$1.50 per share. He has made this statement knowing that the dividend will increase only if Congress enacts certain legislation, an uncertain prospect. By stating that the dividend will increase, Scott failed to separate fact from opinion.

The information regarding passage of legislation is not material nonpublic information because it is conjecture, and the question does not state whether the US representative gave Scott her opinion on the passage of the legislation in confidence. She could have been offering this opinion to anyone who asked. Therefore, statement A is incorrect. It may be acceptable to base a recommendation, in part, on an expectation of future events, even though they may be uncertain. Therefore, answer C is incorrect.

10 The correct answer is C. This question, which relates to Standard III(B)–Fair Dealing, tests the knowledge of the procedures that will assist members and candidates in treating clients fairly when making investment recommendations. The step listed in C will help ensure the fair treatment of clients. Answer A may have negative effects on the fair treatment of clients. The more people who know about a pending change, the greater the chance that someone will inform some clients before the information's release. The firm should establish policies that limit the number of people who are aware in advance that a recommendation is to be disseminated. Answer B, distributing recommendations to institutional clients before distributing them to individual accounts, discriminates among clients on the basis of size and class of assets and is a violation of Standard III(B).

11 The correct answer is B. This question deals with Standard II(A)–Material Nonpublic Information. The mosaic theory states that an analyst may use material public information and nonmaterial nonpublic information in creating a larger picture than shown by any individual piece of information and the conclusions the analyst reaches become material only after the pieces are assembled. Answers A and C are accurate statements relating to the Code and Standards but do not describe the mosaic theory.

12 The correct answer is C. This question involves Standard IV(B)–Additional Compensation Arrangements. The arrangement described in the question—whereby Jurgen would be compensated beyond the compensation provided by her firm, on the basis of an account's performance—is not a violation of the Standards as long as Jurgen discloses the arrangement in writing to her employer and obtains permission from her employer prior to entering into the arrangement. Answers A and B are incorrect; although the private compensation arrangement could conflict with the interests of other clients and lead to short-term performance pressures, members and candidates may enter into such agreements as long as they have disclosed the arrangements to their employer and obtained permission for the arrangement from their employer.

13 The correct answer is B. This question relates to Standard III(A)–Loyalty, Prudence, and Care—specifically, a member's or candidate's responsibility for voting proxies and the use of client brokerage. According to the facts stated in the question, Farnsworth did not violate Standard III(A). Although the company president asked Farnsworth to vote the shares of the Jones Corporation profit-sharing plan a certain way, Farnsworth investigated the issue and concluded, independently, the best way to vote. Therefore, even though his decision coincided with the wishes of the company president, Farnsworth is not in violation of his responsibility to be loyal and to provide care to his clients. In this case, the participants and the beneficiaries of the profit-sharing plan are the clients, not the company's management. Had Farnsworth not investigated the issue or had he yielded to the president's wishes and voted for a slate of directors that he had determined was not in the best interest of the company, Farnsworth would have violated his responsibilities to the beneficiaries of the plan. In addition, because the brokerage firm provides the lowest commissions and best execution for securities transactions, Farnsworth has met his obligations to the client in using this brokerage firm. It does not matter that the brokerage firm also provides research information that is not useful for the account generating the commission because Farnsworth is not paying extra money of the client's for that information.

14 The correct answer is A. In this question, Brown is providing investment recommendations before making inquiries about the client's financial situation, investment experience, or investment objectives. Brown is thus violating

Standard III(C)–Suitability. Answers B and C provide examples of information members and candidates should discuss with their clients at the outset of the relationship, but these answers do not constitute a complete list of those factors. Answer A is the best answer.

15 The correct answer i*s* C. This question involves Standard I(C)–Misrepresentation. Statement I is a factual statement that discloses to clients and prospects accurate information about the terms of the investment instrument. Statement II, which guarantees a specific rate of return for a mutual fund, is an opinion stated as a fact and, therefore, violates Standard I(C). If statement II were rephrased to include a qualifying statement, such as "in my opinion, investors may earn . . . ," it would not be in violation of the Standards.

16 The correct answer is A. This question involves three of the Standards. Anderb, the portfolio manager, has been obtaining more favorable prices for her personal securities transactions than she gets for her clients, which is a breach of Standard III(A)–Loyalty, Prudence, and Care. In addition, she violated Standard I(D)–Misconduct by failing to adhere to company policy and by hiding her personal transactions from her firm. Anderb's supervisor, Bates, violated Standard IV(C)–Responsibilities of Supervisors; although the company had requirements for reporting personal trading, Bates failed to adequately enforce those requirements. Answer B does not represent a violation because Standard VI(B)–Priority of Transactions requires that personal trading in a security be conducted after the trading in that security of clients and the employer. The Code and Standards do not prohibit owning such investments, although firms may establish policies that limit the investment opportunities of members and candidates. Answer C does not represent a violation because the Code and Standards do not contain a prohibition against employees using the same broker for their personal accounts that they use for their client accounts. This arrangement should be disclosed to the employer so that the employer may determine whether a conflict of interest exists.

17 The correct answer is A because this question relates to Standard I(A)–Knowledge of the Law—specifically, global application of the Code and Standards. Members and candidates who practice in multiple jurisdictions may be subject to various securities laws and regulations. If applicable law is more strict than the requirements of the Code and Standards, members and candidates must adhere to applicable law; otherwise, members and candidates must adhere to the Code and Standards. Therefore, answer A is correct. Answer B is incorrect because members and candidates must adhere to the higher standard set by the Code and Standards if local applicable law is less strict. Answer C is incorrect because when no applicable law exists, members and candidates are required to adhere to the Code and Standards, and the Code and Standards prohibit the use of material nonpublic information.

18 The correct answer is B. The best course of action under Standard I(B)–Independence and Objectivity is to avoid a conflict of interest whenever possible. Therefore, for Ward to pay for all his expenses is the correct answer. Answer C details a course of action in which the conflict would be disclosed, but the solution is not as appropriate as avoiding the conflict of interest. Answer A would not be the best course because it would not remove the appearance of a conflict of interest; even though the report would not be affected by the reimbursement of expenses, it could appear to be.

19 The correct answer is B. Under Standard IV(A)–Loyalty, members and candidates may undertake independent practice that may result in compensation or other benefit in competition with their employer as long as they obtain consent from their employer. Answer C is not consistent with the Standards because the Standards allow members and candidates to make arrangements or preparations to go into competitive business as long as those arrangements do not interfere with their duty to their current employer. Answer A is not consistent with the Standards because the Standards do not include a complete prohibition against undertaking independent practice.

20 The correct answer is B. This question involves Standard VI(A)–Disclosure of Conflicts—specifically, the holdings of an analyst's employer in company stock. Answers A and C do not describe conflicts of interest that Smith would have to disclose. Answer A describes the use of a firm's products, which would not be a required disclosure. In answer C, the relationship between the analyst and the company through a relative is so tangential that it does not create a conflict of interest necessitating disclosure.

21 The correct answer is C. This question relates to Standard I(C)–Misrepresentation. Although Michelieu's statement about the total return of his clients' accounts on average may be technically true, it is misleading because the majority of the gain resulted from one client's large position taken against Michelieu's advice. Therefore, this statement misrepresents the investment performance the member is responsible for. He has not taken steps to present a fair, accurate, and complete presentation of performance. Answer B is thus incorrect. Answer A is incorrect because although Michelieu is not guaranteeing future results, his words are still a misrepresentation of his performance history.

22 The correct answer is B. The best policy to prevent violation of Standard II(A)–Material Nonpublic Information is the establishment of firewalls in a firm to prevent exchange of insider information. The physical and informational barrier of a firewall between the investment banking department and the brokerage operation prevents the investment banking department from providing information to analysts on the brokerage side who may be writing recommendations on a company stock. Prohibiting recommendations of the stock of companies that are clients of the investment banking department is an alternative, but answer A states that this prohibition would be permanent, which is not the best answer. Once an offering is complete and the material nonpublic information obtained by the investment banking department becomes public, resuming publishing recommendations on the stock is not a violation of the Code and Standards because the information of the investment banking department no longer gives the brokerage operation an advantage in writing the report. Answer C is incorrect because no exchange of information should be occurring between the investment banking department and the brokerage operation, so monitoring of such exchanges is not an effective compliance procedure for preventing the use of material nonpublic information.

23 The correct answer is B. Under Standard III(A)–Loyalty, Prudence, and Care, members and candidates who manage a company's pension fund owe these duties to the participants and beneficiaries of the pension plan, not the management of the company or the company's shareholders.

24 The correct answer is B. Answer B gives one of the two primary reasons listed in the *Handbook* for disclosing referral fees to clients under Standard VI(C)–Referral Fees. (The other is to allow clients and employers to evaluate the full cost of the services.) Answer A is incorrect because Standard VI(C) does not require members or candidates to discount their fees when they receive referral fees. Answer C is inconsistent with Standard VI(C) because disclosure of referral fees, to be effective, should be made to prospective clients before entering into a formal client relationship with them.

25 The correct answer is B. Standard VI(B)–Priority of Transactions does not limit transactions of company employees that differ from current recommendations as long as the sale does not disadvantage current clients. Thus, answer A is incorrect. Answer C is incorrect because the Standard does not require the matching of personal and client trades.

26 Answer C is correct. Standard IV(A)–Loyalty discusses activities permissible to members and candidates when they are leaving their current employer; soliciting clients is strictly prohibited. Thus, answer A is inconsistent with the Code and Standards even with the required disclosure. Answer B is incorrect because the offer does not directly violate the Code and Standards. There may be out-of-work members and candidates who can arrange the necessary commitments without violating the Code and Standards.

27 Answer A is correct. The question relates to Standard III(A)–Loyalty, Prudence, and Care. Carter believes the broker offers effective execution at a fee that is comparable with those of other brokers, so he is free to use the broker for all accounts. Answer B is incorrect because the accounts that prohibit soft dollar arrangements do not want to fund the purchase of research by Carter. The new trading scheme does not incur additional commissions from clients, so it would not go against the prohibitions. Answer C is incorrect because Carter should not incur unnecessary or excessive "churning" of the portfolios (excessive trading) for the purpose of meeting the brokerage commitments of soft dollar arrangements.

28 Answer B is correct according to Standard VII(B)–Reference to CFA Institute, the CFA Designation, and the CFA Program. CFA Program candidates do not receive their actual scores on the exam. Topic and subtopic results are grouped into three broad categories, and the exam is graded only as "pass" or "fail." Although a candidate may have achieved a topical score of "above 70%," she or he cannot factually state that she or he received the highest possible score because that information is not reported. Thus, answer C is incorrect. Answer A is incorrect as long as the member or candidate actually completed the exams consecutively. Standard VII(B) does not prohibit the communication of factual information about completing the CFA Program in three consecutive years.

29 Answer C is correct. According to Standard III(A)–Loyalty, Prudence, and Care, the CFA Program would be considered a personal or firm expense and should not be paid for with the fund's brokerage commissions. Soft dollar accounts should be used only to purchase research services that directly assist the investment manager in the investment decision-making process, not to assist the management of the firm or to further education. Thus, answer A is incorrect. Answer B is incorrect because the reasonableness of how the money is used is not an issue; the issue is that educational expense is not research.

30 Answer A is correct. Standard I(B)–Independence and Objectivity emphasizes the need for members and candidates to maintain their independence and objectivity. Best practices dictate that firms adopt a strict policy not to accept compensation for travel arrangements. At times, however, accepting paid travel would not compromise one's independence and objectivity. Answers B and C are incorrect because the added benefits—free conference admission for additional staff members and an exclusive golf retreat for the speaker—could be viewed as inducements related to the firm's working arrangements and not solely related to the speaking engagement. Should Long wish to bring other team members or participate in the golf outing, he or his firm should be responsible for the associated fees.

31 Answer C is correct. The guidance to Standard II(A)–Material Nonpublic Information recommends adding securities to the firm's restricted list when the firm has or may have material nonpublic information. By adding these securities to this list, Andrews would uphold this standard. Because waiting until the next day will not ensure that news of the merger is made public, answer A is incorrect. Negotiations may take much longer between the two companies, and the merger may never happen. Andrews must wait until the information is disseminated to the market before he trades on that information. Answer B is incorrect because Andrews should not disclose the information to other managers; no trading is allowed on material nonpublic information.

32 Answer B is correct. Through placing a personal purchase order that is significantly greater than the average volume, Pietro is violating Standard IIB–Market Manipulation. He is attempting to manipulate an increase in the share price and thus bring a buyer to the negotiating table. The news of a possible merger and confirmation of the firm's earnings guidance may also have positive effects on the price of Local Bank, but Pietro's actions in instructing the release of the information does not represent a violation through market manipulation. Announcements of this nature are common and practical to keep investors informed. Thus, answers A and C are incorrect.

33 Answer C is correct. Cupp violated Standard III(D)–Performance Presentations when he deviated from the firm's stated policies solely to capture the gain from the holding being acquired. Answer A is incorrect because the firm does not claim GIPS compliance and the GIPS standards require external cash flows to be treated in a consistent manner with the firm's documented policies. Answer B is incorrect because the firm does not state that it is updating its composite policies. If such a change were to occur, all cash flows for the month would have to be reviewed to ensure their consistent treatment under the new policy.

34 Answer A is correct. According to Standard V(C)–Record Retention, Cannan needed the permission of her employer to maintain the files at home after her employment ended. Without that permission, she should have deleted the files. All files created as part of a member's or candidate's professional activity are the property of the firm, even those created outside normal work hours. Thus, answer B is incorrect. Answer C is incorrect because the Code and Standards do not prohibit using one's personal computer to complete work for one's employer.

35 Answer B is correct. According to Standard VII(B)–Reference to CFA Institute, the CFA Designation, and the CFA Program, Quinn cannot claim to have finished the CFA Program or be eligible for the CFA charter until he officially learns that he has passed the Level III exam. Until the results for the most recent exam are released, those who sat for the exam should continue to refer to themselves as "candidates." Thus, answer C is incorrect. Answer A is incorrect because members and candidates may discuss areas of practice in which they believe the CFA Program improved their personal skills.

36 Answer A is correct. Hart's decision to invest in the retail fund appears directly correlated with Rodriguez's statement about the successful quarter of Mega Retail and thus violates Standard II(A)–Material Nonpublic Information. Rodriguez's information would be considered material because it would influence the share price of Mega Retail and probably influence the price of the entire exchange-traded retail fund. Thus, answer B is incorrect. Answer C is also incorrect because Rodriguez shared information that was both material and nonpublic. Company officers regularly have such knowledge about their firms, which is not a violation. The sharing of such information, however, even in a conversation between friends, does violate Standard II(A).

37 Answer C is correct. Standard VII(A)–Conduct as Members and Candidates in the CFA Program prohibits providing information to candidates or the public that is considered confidential to the CFA Program. In revealing that questions related to the analysis of inventories and analysis of taxes were on the exam, Park has violated this standard. Answer B is incorrect because the guidance for the standard explicitly acknowledges that members and candidates are allowed to offer their opinions about the CFA Program. Answer A is incorrect because candidates are not prohibited from using outside resources.

38 Answer B is correct. Paper has violated Standard III(D)–Performance Presentation by not disclosing that he was part of a team of managers that achieved the results shown. If he had also included the return of the portion he directly managed, he would not have violated the standard. Thus, answer A is incorrect. Answer C is incorrect because Paper received written permission from his prior employer to include the results.

39 Answer A is correct. Townsend has not provided any information about her clients to the leaders or managers of the golf program; thus, she has not violated Standard III(E)–Preservation of Confidentiality. Providing contact information about her clients for a direct-mail solicitation would have been a violation. Answer B is incorrect because the notice in the newsletter does not violate Standard III(E). Answer C is incorrect because the golf program's fund-raising campaign had already begun, so discussing the opportunity to donate was appropriate.

40 C is correct. Boswin, as an employee, developed the model on behalf of Greenfield. Therefore, Greenfield, not Boswin, is the owner of the model. Acertado violates Standard IV(A) Duties to Employers: Loyalty when he downloads the model without proper written permission from Greenfield Financial. Acertado is misappropriating employer assets.

41 C is correct. Acertado is least likely to violate Standard II(A) regarding Material Nonpublic Information when using the model at Smith and Garner. Acertado likely violated Standard IV(A), Loyalty, when he used the model. The Standard prohibits members who leave an employer from taking records or files—such as the model—without the written permission of the employer. Acertado also likely violated Standard I(C)–Misrepresentation when he failed to correct his supervisor's impression that the investment idea and the model were Acertado's creation.

42 C is correct. Acertado violates Standard II(A)–Material Nonpublic Information. He has a reasonable belief that the conversation that he overhears is from a reliable source and would have a material impact on security prices. According to the CFA Standards, he must not act, nor cause others to act on the information. Acertado does not violate the Standard relating to Diligence and Reasonable Basis because he bases the recommendation on a reliable model and checks his inputs prior to making the recommendation.

43 A is correct. Acertado least likely violates Standard III(C), which relates to suitability during his phone conversation with Boswin. According to the Standard, members in an advisory relationship with a client must determine an investment's suitability within the context of the client's portfolio. The Standard also requires that members make reasonable inquiries into a client or prospective client's investment experience; risk and return objectives; and financial constraints prior to making investment recommendations. Boswin is neither a client nor a prospective client, thus Acertado is not bound by the Standard of Suitability during their conversation. Acertado is, however, in jeopardy of violating other Standards—specifically those relating to Integrity of Capital Markets

and Preservation of Confidentiality by revealing material nonpublic information about a Smith & Garner client. According to Standard II(A), Acertado, who is in possession of material nonpublic information, must not act, nor cause others to act on the information. According to Standard III(E), members must keep information about current, former, and prospective clients confidential.

44 A is correct. Boswin uses her usual process in researching Country Industries. She is not in possession of material nonpublic information and she maintains her objectivity. Her use of the model provides a reasonable basis for the decision not to pursue additional research or make an investment recommendation regarding Country Industries.

45 A is correct. Blogs and company websites are in the public domain and thus do not constitute inside information. Acertado's use of blog sites to supplement his current research process is acceptable.

46 A is correct. Grohl exercised diligence, independence, and thoroughness in analyzing the company and its competitors. Brecksen provided his research reports for Grohl's use and using the reports as a guide was appropriate. Standard V(A) requires that members distinguish between fact and opinion in communicating investment recommendations to clients. The Standard does not apply to investment recommendations communicated to supervisors or internal investment committees.

47 A is correct. Brecksen does not consider the multi-factor analysis a critical component of the analysis or the resulting investment recommendation and thus, under Standards V(A) and (C), is not required to maintain a record of the analysis within the completed report.

Apfelbaum uses traditional "top-down" fundamental analysis in the investment process. The report followed the traditional format of previous reports on the same company. It contained a complete fundamental analysis and recommendation—indicating diligence and reasonable basis. The report also contained a multi-factor analysis—which is a quantitative analysis tool. If quantitative analysis were the basis of the investment recommendation, it would constitute a change in the general investment principles used by the firm. According to Standard V(B)–Communications with Clients and Prospective Clients, Brecksen and Grohl would be required to promptly disclose those changes to clients and prospective clients.

48 A is correct. Removing the multi-factor analysis from the research report does not constitute a violation. Grohl diligently prepared the internal document according to the firm's traditional format with a complete fundamental analysis and recommendation—indicating diligence and a reasonable basis for his recommendation. It would be wise for Grohl to retain records of the multi-factor analysis but he need not retain the analysis in the research report to comply with Standards V(A)–Diligence and Reasonable Basis or V(C)–Record Retention.

49 A is correct. According to Standard V(A)–Diligence and Reasonable Basis, research report conclusions or recommendations may represent the consensus of a group and not necessarily the views of the individual members listed. If the member believes that the consensus opinion has a reasonable basis, then he need not dissociate from the report.

50 B is correct. Zardt violated the Standard relating to Performance Presentation because he did not verify the accuracy of the return information before its distribution. According to Standard III(D), analysts may promote the success or accuracy of their recommendations, but they must make reasonable efforts to ensure that the information is fair, accurate, and complete. In addition to providing attribution, Zardt should take steps to ensure the accuracy of the data prior to distributing the material.

51 B is correct. Telline has a duty to preserve the confidentiality of current, former, and prospective clients. Telline violates Standard III(E)–Preservation of Confidentiality when he reveals information about the Leighton Family Trust.

52 A is correct. Brown conducts an independent and careful analysis of the plans' benefits for shareholders as well as the takeover offer. In doing so she puts the client's interests ahead of the firm's. Brown's actions are consistent with Standard III(A)–Loyalty, Prudence, and Care; Standard V(A)–Diligence and Reasonable Basis; and Standard III(B)–Fair Dealing.

53 B is correct. Telline is not likely to receive appropriate guidance on developing or revising investment policy statements from the Standard relating to Fair Dealing. Standard III(B) provides members with guidance on treating clients fairly when making investment recommendations, providing investment analysis, or taking investment action. Telline could obtain guidance from the Standards relating to Loyalty, Prudence, and Care and Suitability. Both Standard III(A) and (C) provide guidance for members in determining client objectives and the suitability of investments.

54 A is correct. Telline is careful to consider the investment's suitability for Caper's account. Telline's actions are consistent with CFA Institute Standards III(A)–Loyalty, Prudence, and Care and III(B)–Fair Dealing. Telline determines that the other client does not have the cash available in his account and selling existing holdings does not make sense.

55 B is correct. The firm violates Standard III(B)–Fair Dealing. Under Aiklin's policy, some clients for whom an IPO is suitable may not receive their pro-rata share of the issue. CFA Standards recommend that firms allocate IPOs on a pro-rata basis to clients, not to portfolio managers.

56 C is correct. Telline violates Standard III(B)–Fair Dealing by over-allocating shares to Caper. Telline carefully reviews both the proposed IPO and his client accounts to determine suitability. He fails to allocate the IPO shares on a pro-rata basis to all clients for whom the investment is suitable.

57 B is correct. Excerpt 2 is inconsistent with CFA Standards because it addresses only officers and only their direct reports, that is, employees whom they directly supervise. Standard IV (C) states that "any investment professionals who have employees subject to their control or influence" exercise supervisory responsibility. Excerpt 3 is consistent with CFA Standards. It is based on a quote from the *Standards of Practice Handbook* stating that "if a member or candidate determines that information is material, the member . . . should make reasonable efforts to achieve public dissemination." Members are not required to achieve public dissemination and those bound by a duty of loyalty or a duty to preserve confidentiality would refrain from doing so because it would breach their duty.

58 A is correct. According to Standard I(B)–Independence and Objectivity, members must use reasonable care and judgment to achieve and maintain independence and objectivity in their professional activities. Although it was sent to Voser's office, the gift basket is a private gift from Voser's sister and not likely to affect Voser's professional activities. According to Excerpt 4 of the Crawfood compliance manual and Standard IV(B)–Additional Compensation Arrangements, employees must obtain permission from their employer before accepting gifts, compensation, or other benefits that compete with, or might create a conflict of interest with, the employer's interests. The gift basket does not create a conflict or compete with the employer's interests.

59 A is correct. Voser most likely violated the Standard relating to loyalty to employer, Standard IV(A). While Voser used public information to develop the recommendation to purchase Greenhornfood shares, the company compliance guide states that she should not trade or cause others to trade in securities of companies that may be potential takeover targets. Voser's recommendation caused her sister to trade in Greenhornfood, violating the company's compliance policies, and possibly harming her employer in its attempt to acquire Greenhornfood.

By advising others to invest in a food retailer that she considered an attractive acquisition target, Voser deprived her employer of the advantage of her skills and abilities and may have caused harm to her employer. Voser could have recommended Greenhornfood to Craw rather than her sister as an acquisition target. Although the sister's trade in Greenhornfood was small, a large trade might have moved the stock price and caused harm to Crawfood in terms of additional cost.

60 C is correct. Voser least likely violated the Standard relating to diligence and reasonable basis. Voser initially applied the mosaic theory and had a reasonable basis for the trade as required by Standard V(A). Eventually, she came into possession of material nonpublic information (corporate logo on a document, overheard conversation). According to Standard II(A), once in possession of material nonpublic information, she is prohibited from acting or causing others to act. Voser also violated her duty of loyalty to her employer, Standard IV(A), by encouraging others to trade in Greenhornfood and possibly harming Crawfood's attempts to acquire the smaller company at an attractive price.

61 C is correct. Craw did not adequately fulfill his responsibilities as a supervisor. As stated in the *Standards of Practice Handbook*, members and candidates with supervisory responsibility also must understand what constitutes an adequate compliance system for their firms and make reasonable efforts to see that appropriate compliance procedures are established, documented, communicated to covered personnel, and followed. "Adequate" procedures are those designed to meet industry standards, regulatory requirements, the requirements of the Code and Standards, and the circumstances of the firm. Once compliance procedures are established, the supervisor must also make reasonable efforts to ensure that the procedures are monitored and enforced. According to Standard IV(C)–Responsibilities of Supervisors, adequate compliance procedures require that once a violation is discovered, Craw conduct a thorough investigation to determine the scope of wrongdoing.

62 A is correct. As stated on page ix of the *Standards of Practice Handbook,* "Members and candidates must self disclose on the annual Professional Conduct Statement all matters that question their professional conduct, such as involvement in civil litigation, a criminal investigation, or being the subject of a written complaint." Standard VII(A)–Conduct as Members and Candidates in the CFA Program prohibits conduct that compromises the reputation of the CFA designation including misrepresenting information on the Professional Conduct Statement. Members are encouraged but not required to report violations of others. At a minimum, Craw should remind Voser of her duty to report the investigation.

READING

3

Application of the Code and Standards: Level II

LEARNING OUTCOMES

Mastery	The candidate should be able to:
☐	**a.** evaluate practices, policies, and conduct relative to the CFA Institute Code of Ethics and Standards of Professional Conduct;
☐	**b.** explain how the practices, policies, and conduct do or do not violate the CFA Institute Code of Ethics and Standards of Professional Conduct.

INTRODUCTION

This reading presents cases to illustrate how the CFA Institute Code of Ethics and Standards of Professional Conduct (Code and Standards) can be applied in situations requiring professional and ethical judgement. Exhibit 1 presents a useful framework to help guide individuals in their ethical decision-making process and application of the Code and Standards. By identifying where the Code and Standards might be relevant and considering actions and consequences within this framework, individuals can make more ethically sound decisions.

Although the framework's components do not need to be addressed in the sequence shown, a review of the outcome should conclude the process. This review provides insights for improved decision making in the future.

Exhibit 1 A Framework for Ethical Decision Making

- Identify: Relevant facts, stakeholders and duties owed, ethical principles, conflicts of interest
- Consider: Situational influences, additional guidance, alternative actions
- Decide and act
- Reflect: Was the outcome as anticipated? Why or why not?

This reading presents a number of scenarios involving individuals in private and institutional asset management. The first three cases focus on identifying whether violations of the Code and Standards occurred, with discussion and rationale as to why or why not a violation may have taken place. The last two cases focus on identifying violations of the Code and Standards, taking necessary corrective actions, and developing a policy statement to help prevent future violations by a firm's employees. As you read through these cases, consider how applying the framework might have helped each individual in his or her decision making.

1 QUANTHOUSE

QuantHouse (QH) is a global investment firm that pioneered the use of quantitative techniques to implement investment strategies. Three years ago, the firm hired Daniel Singh, PhD, CFA, a well-known finance professor. Singh created the Artificial Trading Model (ATM), a comprehensive model that captures and processes a substantial amount of publicly available information (company financial data, news, and industry information) and then makes investment decisions largely without human interaction. ATM has three components: an Alpha Model, a Risk Model, and an Optimizer. The Alpha Model evaluates public companies based on their earnings and valuation. The Risk Model identifies stock-specific risk and common factor risks (industry specific, country specific, and stock fundamental risks). The Optimizer takes the output from the Alpha and Risk Models, balances them against one other, and recommends an optimal client portfolio based on the client's chosen benchmark.

Singh uses the ATM model exclusively for QH's institutional clients and does not mention it when talking with his high-net-worth individual clients. Singh and his team of programmers update the Alpha and Risk components of the ATM on a quarterly basis. On an annual basis, consistent with QH's guidelines for all computer-based models, Singh reviews the Optimizer component and conducts extensive scenario tests with the overall model. (Questions 1, 2)

Recently, some of QH's institutional clients have been voicing concerns about their portfolios' underperformance. In particular, they have expressed dissatisfaction with the overexposure to certain industries in their portfolios, an element that is partly controlled by the ATM's ability to manage risk. In response to these complaints, QH's director of research and Singh's supervisor, Charlotte Ringfield, CFA, asks Singh to review the model. After doing so, Singh finds that the Optimizer is incorrectly reading the Risk Model's assessment of common risk factors and, as a result, is not weighting them appropriately.

Singh then meets with his team of analysts who helped create the model to determine the source of the error. They find that some of the Risk Model components are sending information to the Optimizer in decimal form while other components are sending information in percentages. This improper scaling has resulted in the Optimizer giving inappropriate weights to some of the common risk factors. After discovering the source of the error, Singh and his team meet with Ringfield to present their findings. Singh advocates that the error be fixed as soon as possible, but Ringfield disagrees and tells him to correct the error when the Risk Model is updated at the end of the quarter. She also asks Singh to temporarily disable the common risk factors in the Risk Model until the model is updated. Ringfield then asks Singh and his team not to mention the error to others and reminds them that they signed confidentiality and nondisclosure agreements when they were hired. She goes on to say that she and the investment committee will handle all disclosures to clients and senior management once the model is updated at the end of the quarter. (Questions 3, 4)

Two weeks later, after a very turbulent period in the financial markets, more clients complain about their portfolios' underperformance. Ringfield tells the portfolio managers about the error. When clients inquire about their portfolio's performance, Ringfield attributes the performance to market volatility and the functioning of the model's common risk factors. (Question 5)

Disturbed by the behavior of his colleagues and superiors, who have not yet revealed the error to clients, Singh decides to leave QH. He interviews with QH's largest competitor, Algos-R-Us (ARU). During his interview with ARU's hiring committee, Singh shows them a proprietary model that he has been developing for the past two years in his spare time (nights and weekends). His new model, which he calls StockStar, is based on years of academic research at his university, and Singh considers it to be his life's work. In addition to back-testing the model, Singh has used StockStar to manage his personal portfolio and the portfolios of his family in order to generate actual performance results. Impressed by Singh and his model, the hiring committee not only offers him a job but also offers to pay him a special licensing fee for the use of his StockStar model. Singh accepts the offer and returns to QH to tender his resignation. (Question 6)

On his first day at ARU, Singh presents the following information to the marketing department about the StockStar model's performance. This information will be incorporated into a new marketing brochure that will be mailed to current and prospective clients.

StockStar Model Performance: Actual and Back-Tested Returns*

	Model's Performance (%)	Benchmark Return (%)	Excess Return (%)
Back-tested returns			
2012	10.5	6.5	4.0
2013	–2.5	–6.0	3.5
2014	20.3	25.0	–4.7
2015	0.5	–6.8	7.3
Actual returns			
2016	8.5	2.2	6.3
2017	12.0	–0.5	12.5

* The benchmark return is based on the S&P/ASX 200 index. Note that past performance does not guarantee future results.

In the brochure, Singh states: "This model has been used to manage real portfolios over the past two years and has outperformed its benchmark in both years. In back-tests, the model has outperformed its benchmark in three out of four years." (Question 7)

Case Questions

1 Does Singh violate the CFA Institute Code and Standards by using the ATM model exclusively for QH's institutional clients?

A No

B Yes, because the model may be suitable for some non-institutional clients

C Yes, because he must at least mention the model when talking to high-net worth individuals

A is correct. It is not a violation of Standard III(B): Duties to Clients–Fair Dealing to use different investment models when working with different types of clients. The ATM model may be suitable only for institutional clients because of the size of their portfolios, among other factors. Standard III(B) requires members and candidates to treat all clients fairly when disseminating investment recommendations; making material changes to prior investment recommendations; or when taking investment action with regard to general purchases, new issues, or secondary offerings. Each client has unique needs, investment criteria, and investment objectives, so not all investment opportunities are suitable for all clients.

B is incorrect. It is not a violation of the Code and Standards to use different investment models when working with different types of clients, because some models may be suitable only for specific clients. There is no information in the case to indicate that the model is suitable for clients other than institutional investors.

C is incorrect. There is no duty under the Code and Standards to disclose all available investment models when working with different types of clients, because some models may be suitable only for specific clients. There is no information in the case to indicate that the model is suitable for clients other than institutional investors.

2 Prior to the performance concerns voiced by QH's institutional clients, did Singh violate the CFA Institute Code and Standards in the updating of the ATM model components?

A Yes

B No, because he updates the model's Alpha and Risk components on a quarterly basis

C No, because he followed the firm's guidelines and annually reviews the Optimizer and conducts scenario testing on the overall model

A is correct. Singh violated Standard V(A): Investment Analysis, Recommendations, and Actions–Diligence and Reasonable Basis. Members and candidates must understand the statistical significance of the model results they recommend and must be able to explain these results to clients. Singh did not take adequate care to ensure a thorough review of the model was taking place with appropriate frequency. Although the Alpha and Risk components are updated quarterly, he reviews the Optimizer, which links the two prior components, and the overall model itself on only an annual basis. Singh should have tested each of the model's components, and their combined interactions, with the same quarterly frequency.

B is incorrect. Although Singh updates the Alpha and Risk components quarterly, he reviews the Optimizer, which links these two components, and the overall model itself on only an annual basis. He should review all components as well as the overall model with the same frequency.

C is incorrect. Singh should review all components as well as the overall model on the timeframe that is appropriate (quarterly here) and at a minimum conforms to the firm's guidelines.

3 According to the CFA Institute Code and Standards, what is the next action (from those below) that Singh should take following his conversation with Ringfield about the model error?

A Dissociate from the firm.

B Contact the firm's clients.

C Contact senior management.

C is correct. Singh should contact senior management before dissociating himself from QH or contacting QH clients. Upon discovery of the error, Singh should try to fix the model immediately. By not fixing the model immediately, Singh is harming QH clients. In not gaining Ringfield's approval to fix the error immediately, the next action Singh should take is to contact Ringfield's boss or senior management to make them aware of the situation.

According to Standard I(A): Professionalism–Knowledge of the Law, "If a member or candidate has reasonable grounds to believe that imminent or ongoing client or employer activities are illegal or unethical, the member or candidate must dissociate, or separate, from the activity. In extreme cases, dissociation may require a member or candidate to leave his or her employment. Members and candidates may take the following intermediate steps to dissociate from ethical violations of others when direct discussions with the person or persons committing the violation are unsuccessful. The first step should be to attempt to stop the behavior by bringing it to the attention of the employer through a supervisor or the firm's compliance department. If this attempt is unsuccessful, then members and candidates have a responsibility to step away and dissociate from the activity."

A is incorrect. Dissociating from the firm is the final step in the process. Standard I(A) establishes next steps as follows: "The first step should be to attempt to stop the behavior by bringing it to the attention of the employer through a supervisor or the firm's compliance department. If this attempt is unsuccessful, then members and candidates have a responsibility to step away and dissociate from the activity."

B is incorrect. Contacting the firm's clients directly is not a permitted intermediate step under Standard I(A).

4 Did Singh violate the CFA Institute Code and Standards by not immediately fixing the error in the ATM model?

- **A** Yes
- **B** No, because the error will be fixed next quarter
- **C** No, because Singh disabled the common risk factors in the Risk Model as ordered by Ringfield

A is correct. By not fixing the error in the ATM model immediately, Singh is violating Standard III(A): Duties to Clients–Loyalty, Prudence, and Care. Members and candidates have a duty of loyalty to their clients and must act with reasonable care and exercise prudent judgement. By not immediately fixing the error in the model, Singh is not acting for the benefit of clients, nor is he placing client interests before his employer's or his own interests.

B is incorrect. The error should have been fixed immediately. According to Standard III(A), "Investment actions must be carried out for the sole benefit of the client and in a manner the member or candidate believes, given the known facts and circumstances, to be in the best interest of the client." It was in the best interest of clients to fix the error in the model immediately.

C is incorrect. Disabling the common risk factors in the Risk Model did not address the underlying error. The error should have been fixed immediately regardless of Ringfield's order. According to Standard III(A), "investment actions must be carried out for the sole benefit of the client and in a manner the member or candidate believes, given the known facts and circumstances, to be in the best interest of the client." It was in the best interest of clients to fix the error in the model immediately.

5 Did Ringfield violate the CFA Institute Code and Standards when talking with clients about their portfolios' underperformance?

- **A** Yes

B No, because the market was turbulent

C No, because the model's common risk factors were to blame

A is correct. When talking with clients about their portfolios' underperformance, Ringfield was in violation of Standard I(C): Professionalism–Misrepresentation. Members and candidates must not knowingly make any misrepresentations relating to investment analysis, recommendations, actions, or other professional activities. By attributing the underperformance of client portfolios to market volatility, she is not telling them the real reason for the underperformance. In addition, the common risk factors have been disabled, so they are not functioning as intended for the model.

B is incorrect. The reason for the underperformance of client portfolios is the error in the model, not market turbulence. Ringfield violated Standard I(C) regarding misrepresentations relating to investment analysis, recommendations, actions, or other professional activities.

C is incorrect. The reason for the underperformance of client portfolios is the error in the model, not the model's common risk factors. Ringfield violated Standard I(C) regarding misrepresentations relating to investment analysis, recommendations, actions, or other professional activities.

6 Did Singh violate the CFA Institute Code and Standards with respect to his duties to his employer, QuantHouse, in developing his StockStar model?

A No

B Yes, because the model was developed while he was working at QH

C Yes, because he invests his personal and family portfolios using the model

A is correct. Singh is not in violation of the CFA Institute Code and Standards, Standard IV(A): Duties to Employers–Loyalty. In this case, Singh developed the StockStar model in his spare time (on nights and weekends) and used the model to manage only his personal and family portfolios. In addition, he has not been compensated for the model.

B is incorrect. It is not a violation of Standard IV(A): Duties to Employers–Loyalty to develop a model in his spare time.

C is incorrect. It is not a violation of Standard IV(A): Duties to Employers–Loyalty, to use his personal and family portfolios to test or invest in the model.

7 In the table that Singh provides to the marketing department, does he violate the CFA Institute Code and Standards?

A No, because he presented the performance information in the manner required by the CFA Institute Code and Standards

B Yes, because he should have included only the actual performance results of the model

C Yes, because he should have disclosed that he used his personal and family portfolios to generate actual results

A is correct. Singh has not violated Standard III(D): Duties to Clients–Performance Presentation related to performance presentation because he has presented both the actual and back-tested performance of the model and clearly distinguished between the two. He has also noted that past performance does not guarantee future results.

B is incorrect. Standard III(D): Duties to Clients–Performance Presentation encourages full disclosure of investment performance data. Both actual and simulated performance measures are allowed as long as they are clearly disclosed. Singh fully explained the performance results being reported, stating that results are simulated (back tested) when model results are used and indicating that the actual and back-tested results are gross of fees.

C is incorrect. Standard III(D): Duties to Clients–Performance Presentation does not prohibit showing past performance of funds managed as long as appropriate disclosures are made, including the person's role in generating that performance. Singh fully explained the performance results being reported, stating that results are simulated (back tested) when model results are used and indicating that the actual and back-tested results are gross of fees.

JR AND ASSOCIATES

2

Jacobs, Riccio, and Associates (JRA) is a global investment advisory firm that primarily provides high-net-worth individuals and their families with personalized wealth management solutions such as wealth planning, retirement planning, investment management, and trust and fiduciary services. In addition, the firm has a small number of institutional clients. JRA employs 25 investment advisers and portfolio managers.

Benjamin Jacobs, CFA, and Andrew Riccio, CFA, founded JRA 10 years ago. Prior to establishing the firm, Jacobs worked as a lawyer for Brightman Partners, a large and prestigious law firm that specializes in real estate, family law, and estate planning. Riccio worked as a Certified Public Accountant for Earnest & Olds (E&O), a multinational professional services firm that specializes in providing tax, consulting, and advisory services to corporations and individuals. Kathy Parker, CFA, joined the firm as the third senior partner two years after it was founded. Previously, she had worked for the Frontline Group, a broker/dealer. JRA acquires most of its clients through referral arrangements put in place by the three senior partners.

Jacobs has a fee-sharing arrangement with his former colleagues at Brightman Partners (BP) when they refer clients to JRA. The annual investment fee stated in JRA's marketing brochure is higher than the fee most of its clients pay because Jacobs offers a discount on the investment fee to clients who are referred by BP lawyers. This discount encourages the BP lawyers to market JRA's services to their clients. In return, JRA shares a portion of the client's' annual investment advisory fee with the referring lawyer. The lawyers at BP disclose this fee-sharing arrangement with the clients that they refer to JRA. JRA discloses all of this information in the JRA investment management agreement that individuals sign at the time they become clients. (Questions 1, 2)

Riccio offers a similar fee discount and sharing arrangement to accountants at his previous firm, Earnest & Olds (E&O), who refer their clients to JRA. Over time, however, Riccio has observed that many of JRA's clients are reluctant to tell their investment adviser about securities and real estate holdings that are managed at other firms. As a result, the adviser does not have a complete understanding of the client's overall financial position. To assist JRA advisers in developing more-realistic and accurate investment policy statements, the accountants at E&O provide a copy of their referred client's tax returns to the client's JRA adviser after they open an account at JRA. This step allows JRA advisers to "know their client" better and provides greater transparency into their client's financial condition. In return, JRA advisers provide their clients' quarterly account statements to their E&O accountants to help with their tax planning and year-end tax preparation. Client approval is not needed for this information sharing because clients sign confidentiality statements directly with their E&O accountants and JRA advisers, and because they often view their investment adviser and their accountant as a team. (Question 3)

Kathy Parker has a somewhat different referral arrangement in place with the Frontline Group. Frontline's brokerage unit refers all of its small institutional clients (pension plans, profit sharing plans, and endowments) that are looking for investment management to JRA. In return, all of the trading from these accounts continues to be executed through Frontline's broker/dealer. Because Frontline continues to provide "best price and best execution" to these clients, Parker believes no additional client disclosures are necessary because client trading is unaffected. (Question 4)

Since starting JRA, Jacobs and Riccio have developed a close relationship with Tim Carroll, an independent consultant they met at a networking event. Carroll is hired by pension funds to solicit and review proposals from investment advisers who wish to manage a portion of the pension fund's assets. Over the years, Carroll has been instrumental in JRA's success by referring several of his pension fund clients to the firm because of the firm's outstanding performance record and superior client service. To thank Carroll for all of his hard work on JRA's behalf (regardless of whether Carroll's pension fund clients actually hire JRA), Jacobs and Riccio each make sizable annual donations to Carroll's Children's Charity, a non-profit organization Carroll created to benefit orphans. Because these donations are made annually, they are not disclosed to the pension funds referred by Carroll who become JRA clients. (Question 5)

Recently, JRA hired Mufid Othan, an investment adviser and CFA charterholder who previously worked at JRA's largest competitor, Sack International. To attract Othan and his large "book of clients," JRA offered him $500 for each client he "brought over" from Sack. While at Sack, Othan was allowed to connect with all of his clients through his personal social media platforms. This not only enabled him to build an electronic database containing the names, addresses, phone numbers, and email addresses of all his clients but also helped him to provide superior client service by "following" his clients' personal and professional lives. When Othan tendered his resignation from Sack, he was immediately escorted out of the building. Othan spent the following weekend contacting all of his clients via social media to tell them about his resignation and to encourage them to join him at JRA. He did not disclose to them, however, that he was being paid $500 for each client he brought over from Sack. (Questions 6, 7)

A few weeks after beginning work at JRA, Othan hired Zane Ode, a recent college graduate, who recently found out she had passed Level III of the CFA Program examination. After hearing the good news about her success with Level III, Ode posted the following comments in a CFA candidate chatroom:

Comment 1	"I can't believe I passed the exam; the ethics questions were super hard."
Comment 2	"Wow, I scored above the Minimum Performance Score (MPS) on derivatives. I still don't know what answer was right for the two-part contango–backwardation question."
Comment 3	"The graders must have been quite lenient in grading my answers to the constructed response questions."

Ode now has three and a half years of experience in the investment industry. Nevertheless, Othan has already made a habit of introducing her to current and prospective clients as the firm's "newest CFA," and Ode has said nothing to correct him. (Questions 8, 9)

Case Questions

1 Does Jacobs violate the CFA Institute Code and Standards by offering his referral clients a lower investment advisory fee than the one quoted in JRA's marketing brochure?

A No

B Yes, because JRA is misrepresenting its fees

C Yes, because JRA is not dealing with its clients fairly

A is correct. Jacobs is not in violation of the CFA Institute Code and Standards. According to Standard III (B): Duties to Clients–Fair Dealing, members and candidates may provide more personal, specialized, or in-depth service to clients who are willing to pay for premium services through higher management fees or higher levels of brokerage. The term "fair" implies that the member or candidate must take care not to discriminate against any clients when disseminating investment recommendations or taking investment action.

B is incorrect. JRA is not misrepresenting its fees, because some of its clients are paying the fees that are disclosed in its marketing brochure. In addition, the advertised fees represent the highest fees that clients would pay.

C is incorrect. Standard III (B): Duties to Clients–Fair Dealing focuses on investment recommendations and taking investment action. The case provides no evidence that non-referred clients are being discriminated against or that referred clients are receiving preferential treatment, with respect to the dissemination of investment recommendations or the taking of investment action. Referred clients are simply receiving discounted fees.

2 Does Jacobs violate the CFA Institute Code and Standards in his disclosure of referral arrangements to his clients?

A Yes

B No, because the lawyers disclose to their clients the discount that JRA offers

C No, because the discount and the fee-sharing arrangement is disclosed to individuals at the time they sign the investment management agreement

A is correct. Jacobs is in violation of Standard VI (C): Conflicts of Interest–Referral Fees, which states, "Members and candidates must disclose to their employers, clients, and prospective clients, as appropriate, any compensation, consideration, or benefit received from or paid to others for the recommendation of products or services.... Appropriate disclosure means that members and candidates must advise the client or prospective client, before entry into any formal agreement for services, of any benefit given or received from the recommendation of any services provided by the member or candidate." In this case, the disclosure does not occur until the time the individual signs the investment management agreement, which is too late.

B is incorrect. The case facts state that BP lawyers disclose the fee-sharing arrangement to the clients they refer to JRA. The case facts do not state whether the lawyers disclose the discount offered by JRA. The behavior of the BP lawyers, however, is not covered by the Code and Standards. Disclosures, or lack thereof, by BP lawyers do nothing to mitigate JRA's duties and responsibilities.

C is incorrect. The discount is disclosed to JRA clients at the time they sign the investment management agreement. According to the Standard VI (C), disclosure must occur before the client enters into a formal agreement.

3 Do JRA advisers violate the CFA Institute Code and Standards by sharing client information with the accountants at E&O?

A Yes

B No, because the client views representatives from both firms as a team

C No, because the client has signed confidentiality agreements with both firms

A is correct. JRA advisers have violated the confidentiality of their clients by not obtaining client approval (written approval is recommended) in advance of sharing their information between each firm. According to Standard III(E): Duties to Clients–Preservation of Confidentiality, members and candidates are required to preserve the confidentiality of information communicated to them by their clients, prospective clients, and former clients. This standard is applicable when (1) the member or candidate receives information because of his or her special ability to conduct a portion of the client's business or personal affairs and (2) the member or candidate receives information that arises from or is relevant to that portion of the client's business that is the subject of the special or confidential relationship.

B is incorrect. Although the client has signed confidentiality agreements with both firms, the client has not signed an agreement allowing the sharing of information between the firms.

C is incorrect. Although the client may view representatives from both firms as a team, neither team has received client approval in advance of sharing the client's information. As a practical matter, if JRA advisers request information from prospective clients regarding other investment income and assets and the prospect denies existence of such assets, the adviser is under no obligation to perform additional due diligence to ascertain the existence of other assets.

4 Has Parker violated the CFA Institute Code and Standards in her referral arrangement with Frontline Group?

A Yes

B No, because Frontline Group continues to provide "best price" and "best execution"

C No, because nothing has changed—all client trades are still executed by Frontline

A is correct. By not disclosing the referral arrangement to clients who were referred to her by Frontline Group, Parker has violated Standard VI (C): Conflicts of Interest–Referral Fees, which states, "Members and candidates must disclose to their employers, clients, and prospective clients, as appropriate, any compensation, consideration, or benefit received from or paid to others for the recommendation of products or services.… Appropriate disclosure means that members and candidates must advise the client or prospective client, before entry into any formal agreement for services, of any benefit given or received from the recommendation of any services provided by the member or candidate." In this case, there is no evidence to suggest Parker disclosed her referral arrangement with Frontline Group to prospective clients. By not doing so, Parker violated Standard VI (C).

B is incorrect. Regardless of whether Frontline provides "best price" and "best execution" or whether the execution of client trades remains unchanged by Frontline, Parker must still disclose the referral arrangement to her clients.

C is incorrect. Parker must still disclose the referral arrangement to her clients, regardless of the fact that all client trades continue to be executed by Frontline.

5 Did Jacobs and Riccio violate the CFA Institute Code and Standards by making annual donations to Carroll's Children's Charity?

A No

B Yes, because these donations create a conflict of interest

C Yes, because these donations represent additional compensation to Carroll

B is correct. The donations made by Jacobs and Riccio give Carroll an incentive to refer potential clients to JRA and at the very least give the perception that Carroll's objectivity and independence have been compromised. Jacobs and Riccio are in violation of Standard I(B): Professionalism–Independence and Objectivity, which states, "Members and candidates must use reasonable care and judgement to achieve and maintain independence and objectivity in their professional activities. Members and candidates must not offer, solicit, or accept any gift, benefit, compensation, or consideration that reasonably could be expected to compromise their own or another's independence and objectivity."

A is incorrect. As already noted, donations made by Jacobs and Riccio give Carroll an incentive to refer potential clients to JRA. This at the very least gives the perception that Carroll's objectivity and independence have been compromised, and so Jacobs and Riccio are in violation of the Code and Standards, specifically Standard I(B): Professionalism–Independence and Objectivity.

C is incorrect. The donations were made to Carroll's charity and do not represent additional compensation to Carroll. Additional compensation is defined in Standard IV(B): Duties to Employers–Additional Compensation Arrangements as "gifts, benefits, or compensation, or consideration that competes with or might reasonably be expected to create a conflict of interest with their employer's interest." An additional compensation arrangement is one that creates a conflict of interest between the member or candidate and her employer.

6 Did Othan violate the CFA Institute Code and Standards by contacting his Sack International clients via social media after leaving Sack?

A No

B Yes, because he is using client confidential information

C Yes, because the client information he is using belongs to Sack

A is correct. Othan is not in violation of the CFA Institute Code and Standards. According to Standard IV(A): Duties to Employers–Loyalty, "Members and candidates should understand and abide by all applicable firm policies and regulations as to the acceptable use of social media platforms to interact with clients and prospective clients. This is especially important when a member or candidate is planning to leave an employer." In this case, Sack allowed Othan to use his personal social media platforms to connect with clients. In addition, he did not contact his former clients via social media to inform them about his departure until after he resigned from Sack.

B is incorrect. Contacting his clients via social media after leaving Sack, does not require Othan to use confidential client information.

C is incorrect. Othan used his personal social media platforms to connect with clients. These platforms are not the property of Sack.

7 Did Othan violate the CFA Institute Code and Standards by not disclosing to clients that he was receiving $500 for each client that he brought over to JRA from Sack?

A No

B Yes, because this is a referral fee

C Yes, because this is additional compensation

A is correct. Othan is not in violation of the CFA Institute Code and Standards. The $500 does not have to be disclosed to clients because it is not a referral fee or additional compensation, and it does not create a conflict of interest with his employer, clients, or prospective clients.

B is incorrect. According to Standard VI(C): Conflicts of Interest–Referral Fees, referral fees are "any compensation, consideration, or benefit received from or paid to others for the recommendation of products or services." The $500 Othan received from JRA for each client he brought over from Sack is not a referral fee because the $500 is being paid by the employer (JRA) to the employee (Othan) for services provided. This amount is compensation paid by the firm, not a fee charged to clients.

C is incorrect. Additional compensation is defined in Standard IV(B): Duties to Employers–Additional Compensation Arrangements as "gifts, benefits, or compensation, or consideration that competes with or might reasonably be expected to create a conflict of interest with their employer's interest." The $500 is not additional compensation, and there is no conflict with the employer's interests. Although disclosure of all bonus arrangements may add clarity, the Code and Standards do not require members and candidates to disclose how they are compensated.

8 Which of the comments Ode posted in the CFA candidate chatroom violated the CFA Institute Code and Standards?

A Comment 1

B Comment 2

C Comment 3

B is correct. Ode's comment 2 violated Standard VII(A): Responsibilities as a CFA Institute Member or CFA Candidate–Conduct as Participants in the CFA Institute Programs: "CFA Institute program rules, regulations, and policies prohibit candidates from disclosing confidential material gained during the exam process." Examples of information that cannot be disclosed by candidates sitting for an exam include but are not limited to the following:

- Specific detail of questions appearing on the exam (contango–backwardation).
- Broad topical areas and formulas tested or not tested on the exam (derivatives).

In this case, Ode disclosed specific details of questions appearing on the exam.

A is incorrect. In saying that the ethics questions were super hard, Ode did not disclose confidential information gained during the exam process.

C is incorrect. In saying that the graders must have been quite lenient in grading her answers to the constructed response questions, Ode did not disclose confidential information gained during the exam process.

9 Did Othan violate the CFA Institute Code and Standards in his description of Ode?

A Yes

B No, because Ode will be a CFA charterholder in another six months

C No, because Ode has successfully completed all three levels of the CFA Program

A is correct. Othan is in violation of the CFA Institute Code and Standards. Ode is not yet a CFA charterholder, and in referencing her as the firm's "newest CFA," Othan is misrepresenting Ode. Standard VII(B): Responsibilities as a CFA Institute Member or CFA Candidate–Reference to CFA Institute, the CFA Designation, and the CFA Program states that "'CFA Charterholders' are those individuals who have earned the right to use the CFA designation granted by CFA Institute. These people have satisfied certain requirements, including completion of the CFA Program, and required years of acceptable work experience." The recommended procedures for Ode's compliance with Standard VII(B) include educating others in the firm, including re-educating Othan, about her status.

B is incorrect. To be a CFA charterholder, Ode needs to have completed the required four years of work experience.

C is incorrect. The fact that she has completed all three levels of the CFA Program does not make Ode a CFA charterholder. To be a CFA charterholder, she must also have the required four years of work experience.

MAGADI ASSET MANAGEMENT

3

Magadi Asset Management (Magadi) is a global investment management firm based in Nairobi, Kenya. Magadi manages dedicated equity, fixed income, and real estate funds, as well as other alternative investment vehicles. The firm's clients include pension schemes, sovereign wealth funds, and high-net-worth individuals. Frederick Omondi, CFA, is Magadi's president and chief investment officer. Under Omondi, the CFA Code of Ethics and Standards of Professional of Conduct has been adopted as the firm's Code of Conduct for Magadi's employees.

Last year, Omondi established a proprietary trading desk at Magadi. The role of the proprietary traders is to actively trade African securities for the firm's benefit. Proprietary traders do not execute orders for Magadi's institutional or retail clients; these orders are handled by traders on the main trading desk. To increase cooperation among traders and encourage the sharing of best execution practices, both trading desks are located on the same floor at Magadi's headquarters. This proximity has allowed proprietary traders to hear customer order flow and also see customer order information on the computer screens of the main traders. To encourage collaboration between the two trading desks, Omondi offers bonuses to proprietary traders who provide trading ideas to the main traders for the benefit of their clientele.

To allay client concerns about potential front-running, Omondi has told clients that information concerning their orders and business affairs is kept confidential. He further explains the firm has instituted a firm-wide policy that expressly states the following: "Employees may not discuss the business affairs of any client with any other person, except on a strict need-to-know basis. Trade orders made by the proprietary traders that may be similar to client orders must be executed after the clients' orders have been fully executed by the main dealing desk traders." (Questions 1, 2)

Omondi's biggest business success this year was a large mandate from a sovereign wealth fund to invest in Magadi's managed funds. To secure the mandate win, Omondi hired, as a "sub-adviser" to the managed funds, a business development agent with contacts at the highest level within the government responsible for the sovereign wealth fund. Despite having very limited experience as a financial consultant, the agent had a number of close relationships with senior managers at the sovereign wealth fund because of his connections to the government officials responsible for the fund. The payments made by Omondi, through the sub-adviser, included a "deal fee" and other expenses that facilitated the governmental support of the sovereign wealth fund investment. Omondi did not require the agent to provide details regarding its activities or the specific expenses covered by the fee. The agent's expenses are charged to Omondi's managed funds. As a thank you for being awarded the mandate, Omondi made donations to the favorite charities of the sovereign wealth fund's top management, as he had promised during the due diligence process. (Question 3)

Three years ago, Magadi launched the Pan Africa Frontier Fund (PAFF), a non-listed equity unit trust with an investment mandate that prohibits the use of leverage. The mandate requires the following:

- 80% of the companies in the portfolio to be traded on at least one of the 17 securities exchanges operating within Africa;
- the portfolio be invested in a minimum of eight countries at all times;
- no more than 30% of the portfolio's value can be invested in any single country;
- no more than 10% of the portfolio's value can be invested in cash and cash equivalents; and
- no single security can account for more than 15% of the portfolio's value.

Since its launch, PAFF has significantly underperformed its peers and has had several quarters of negative returns. As a result, it ranks in the bottom performance quadrant relative to its peers.

Omondi recently hired Bukenya Kirabo, CFA, to take over management of PAFF. Kirabo was hired to improve PAFF's performance and move the fund to the top performance quadrant in rankings based on his extensive experience and knowledge of African equities, as well as his reputation as an astute investment manager. Kirabo has more than two decades of experience analyzing and investing in public companies across Africa. After graduating from a top local university, he moved to London, where he worked for a global asset management firm. Five years ago, Kirabo was transferred to the firm's regional office in Africa to manage one of the firm's local funds. During the last five years, Kirabo has generated average annual returns of 23%. Since returning to Africa, Kirabo has witnessed notable improvements in African securities markets, particularly in the area of settlement risk. Many local markets remain relatively illiquid, however, and most public companies in Africa are under-researched compared with other emerging markets. As a result, systematic risk is considerably higher in African markets than in other emerging markets.

Three months after being hired at Magadi, Kirabo meets with Omondi to review PAFF's most recent quarterly performance. During the meeting, he states, "PAFF's solid performance this quarter is a result of three changes I made:

> Change 1: Because of strong cash inflows into PAFF, I have increased the maximum level of cash and cash equivalents to 15% of the portfolio. Given the illiquid nature of many markets in which we are investing, I believe it is more prudent, and less risky, to take sufficient time to find attractive investment opportunities and build position holdings.
>
> Change 2: I have increased the portfolio's geographic diversification from 11 countries (stock exchanges) to 13. Securities traded on 13 different African stock exchanges (up from 11 previously) are now represented in the portfolio. This higher level of diversification has improved the portfolio's Sharpe and information ratios.
>
> Change 3: To increase accountability for PAFF's performance, I am now making all buy and sell decisions for PAFF. Previously, when the team of analysts was making the investment decisions, it was difficult to attribute an individual's contribution to fund performance."

Kirabo next meets with the marketing department to discuss PAFF's new sales campaign. During the meeting, he states, "Please include all of the mandate changes I have made in PAFF in the new brochure that will be distributed to prospective clients. You can also include the five-year investment performance I achieved while managing a fund at my previous employer. Please do not state where the performance was earned, however, because my previous employer is a direct competitor of Magadi. Finally, because the mandate changes are relatively trivial, there is no need to inform existing clients." (Questions 4, 5, 6)

PAFF currently owns 9% of the common stock of Mtume's, a mining company listed on the Botswana Stock Exchange. Kirabo has been reducing the fund's holdings in Mtume because of the company's declining revenues and profits. This morning, Kirabo speaks with Olivia Moroka, Mtume's chief financial officer. During their conversation, Moroka tells Kirabo, "You may want to stop selling your shares of Mtume, because our board of directors just received a very attractive all-cash offer of BWP500 million (Botswana pula) to purchase one of our mining subsidiaries. Although nothing is definite, the board will be meeting next week to vote on the offer."

After getting off the phone with Moroka, Kirabo calls the Magadi analyst who follows Mtume and tells her about his conversation. The analyst then incorporates the expected subsidiary sales price into her financial model of Mtume. The output from her revised model indicates that the sale proceeds will significantly enhance Mtume's credit standing and its ability to reinstitute shareholder cash distributions on an earlier-than-expected schedule and in larger-than-expected amounts. When the analyst tells Kirabo about her findings, Kirabo immediately calls the proprietary and main traders to tell them to start buying "any and all" shares of Mtume. He then calls Omondi and tells him about his conversation with Moroka. After Omondi gets off the phone with Kirabo, Ormondi calls his broker and purchases shares in Mtume for his personal account and the family accounts that he controls. (Questions 7, 8)

Case Questions

1 By allowing customer order information to be known to the traders on the proprietary desk, did traders on the main trading desk most likely violate the CFA Institute Code and Standards?

A Yes

B No, because this information was not shared outside of the firm

C No, because proprietary traders were not allowed to act on this information until after client orders were executed

A is correct. Traders on the main trading desk are in violation of Standard III(E): Duties to Clients–Preservation of Confidentiality. This standard requires members and candidates to preserve the confidentiality of information communicated to them by their clients, prospective clients, and former clients. The sharing of office space such that the proprietary traders can see the screens of the main traders is inappropriate because it allows confidential client information to be disclosed to individuals (proprietary traders) who did not need to know the information. To avoid sharing confidential information and violating firm policy, the main traders should have taken necessary action to ensure the client information was not advertently or inadvertently shared with the proprietary desk traders.

B is incorrect. Although the information was not shared externally, the main traders still allowed its disclosure to individuals who did not meet the "need to know" requirement and, in doing so, violated Standard III(E).

C is incorrect. Whether or not the proprietary traders acted on the information is irrelevant in this case. Traders on the main trading desk needed to take the necessary action to prevent the disclosure of confidential information and, in not doing so, they violated Standard III(E).

2 Did Omondi most likely violate the CFA Institute Code and Standards in supervising the employees in the two trading desks?

A Yes

B No, because he implemented a policy to prevent front-running

C No, because he encouraged collaboration between the two departments

A is correct. Omondi was in violation of Standard IV(C): Duties to Employers–Responsibilities of Supervisors. Members and candidates must promote actions by all employees under their supervision and authority to comply with applicable laws, rules, regulations, firm policies and the Code and Standards. Omondi failed to establish effective policies and procedures reasonably designed to prevent traders on the proprietary dealing desk from obtaining confidential customer information. Although the proprietary traders did not have direct access to the computer system used by the main traders to execute customer orders, by being co-located on the same floor, the proprietary traders could still view customer order information on the main traders' computer screens and hear them discuss customer orders. Omondi could have located one set of traders in a separate space or a different floor with security access restrictions. Omondi would also likely be in violation of Standard I(C): Professionalism–Misrepresentation, because his representation to customers was incorrect—that is, client information was made available to other employees outside of those operating on a "need-to-know" basis.

B is incorrect. Although Omondi announced a policy to mitigate front-running, the policy was ineffective and, as implemented, did not prevent or address the sharing of confidential client information (orders) to individuals who did not need to know this information (proprietary traders).

C is also incorrect. Encouraging collaboration between the two trading desks does not address the fact that the proprietary traders could see and hear confidential information about client orders from the main trading desk. Omondi failed to establish sufficient policies and procedures to ensure compliance with the Code and Standards as well as firm policy for the traders under his supervision.

3 Omondi most likely violated the CFA Institute Code and Standards when dealing with the sovereign wealth fund's top managers:

- **A** only by making charitable donations.
- **B** only by hiring a sub-adviser because of his high-level government contacts.
- **C** by both A and B.

C is correct. Omondi was in violation of Standard I(B): Professionalism–Independence and Objectivity. "When working to earn a new investment allocation, members and candidates should not offer gifts, contributions, or other compensation to influence the decision of the hiring representative. The offering of these items with the intent to impair another person's independence and objectivity would not comply with Standard I(B). Such prohibited actions may include offering donations to a charitable organization."

To better serve clients, investment professionals may delegate to third parties work that requires particular specialization, knowledge, or expertise. For instance, an investment adviser may hire sub-advisers to handle a particular strategy or investment style outside the scope of the adviser's ability or experience. A global adviser may hire a sub-adviser to manage an asset allocation invested in a particular country or region, and the payments to the sub-adviser would be legitimate investment expenses that could properly be passed on to investors in the fund. It is clear from the facts of this case, however, that Omondi is not hiring a true sub-adviser but instead paying locally connected officials to secure access for the sovereign wealth fund's investment. The "sub-adviser" has limited financial experience but is close to the government officials, and the "deal fees" are not supported by any documentation that details legitimate investment expenses. The "sub-advising expenses" charged by Omondi to the fund could, in all likelihood, be funding corrupt transactions and bribes through local intermediaries. This practice violates multiple standards, including I(A) Knowledge of the Law (because the conduct would violate any type of anti-bribery laws); I(C) Misrepresentation (improperly labeling the expenditures as investment fees); V(A) Diligence and Reasonable Basis (no reasonable and adequate basis for the "investment" action); and V(C) Record Retention (No appropriate records to support the action).

4 According to the CFA Institute Code and Standards, which of the changes in the PAFF Fund does Kirabo *not* have to disclose?

- **A** Change 1

B Change 2

C Change 3

B is correct. Change 2 is not required to be disclosed because, by increasing the country exposure to 13 nations, Kirabo is still within the 80% stated mandate. The investment process has not fundamentally changed. Changes 1 and 3 are modifications to the investment process that, according to Standard V(B): Investment Analysis, Recommendations, and Actions–Communication with Clients and Prospective Clients, must be disclosed. According to Standard V(B), members and candidates must disclose to clients and prospective clients the basic format and general principles of the investment processes they use to analyze investments, select securities, and construct portfolios, and they must promptly disclose any changes that might materially affect those processes.

A is incorrect. Change 1 is a change in the fund's mandate because the maximum amount of cash that the fund can hold has been increased to 15% from 10%.

C is incorrect. Change 3 is a change to the investment process, because all purchase and sell decisions are now being made by Kirabo instead of the team of analysts.

5 Does Kirabo most likely violate the CFA Institute Code and Standards by including his prior performance in the PAFF marketing brochure?

A No

B Yes, because the brochure should have stated the name of the firm where he earned prior performance

C Yes, because the marketing brochure should not show fund performance earned at a prior firm as part of his performance track record

B is correct. Kirabo was in violation of Standard III(D): Duties to Clients–Performance Presentation. Standard III(D) does not prohibit showing past performance of funds managed at a prior firm, as long as showing that record is accompanied by appropriate disclosures about where the performance took place and the person's specific role in achieving that performance. Kirabo does not disclose the name of the prior firm or that he alone managed the fund and was solely responsible for its performance. Consequently, he is in violation of Standard III(D): Performance Presentation. Kirabo would also be required to receive permission in writing from his previous employer to take his performance records with him when he left the firm, because the performance record is an asset of the firm, not of the individual employee. If he did not receive prior written permission, he would also be in violation of Standard IV(A): Duties to Employers–Loyalty, which requires members and candidates to protect their employers' interests, even when leaving the firm.

C is incorrect. There is no prohibition on including past investment performance under Standard III(D) so long as there are disclosures that clearly indicate it was earned at a previous entity and what the role the manager played in achieving that performance.

6 According to the CFA Institute Code and Standards, whom must Kirabo most likely inform of the material changes related to the PAFF?

A Current clients only

B Prospective clients only

C Current and prospective clients

C is correct. According to Standard V(B): Investment Analysis, Recommendations, and Actions–Communication with Clients and Prospective Clients, Kirabo must disclose to current and prospective clients both the fund mandate change and the change in the investment process.

7 Did Kirabo most likely violate the CFA Institute Code and Standards by purchasing additional shares of Mtume?

A Yes

B No, because the information that Kirabo learned from Moroka was not definite

C No, because his decision was based on the output from the analyst's revised model

A is correct. Kirabo was in violation of Standard II(A): Integrity of Capital Markets–Material Nonpublic Information. Members and candidates who possess material nonpublic information that could affect an investment's value must not act or cause others to act on the information. Information is "material" if its disclosure would probably affect the price of a security or if reasonable investors would want to know the information before making an investment decision. In addition to the substance and specificity of the information, the source or relative reliability of the information also determines materiality. In this case, factual information from a corporate insider regarding the purchase of a subsidiary is likely to be material. Although the offer is not definite or officially accepted by the board, its source, substance, and specificity are enough to make the information material. The output from the analyst's revised model was affected by the insider information.

B is incorrect. Information does not have to be definite to trigger the violation; it need only be considered both material and nonpublic. In this case, factual information from a corporate insider regarding the purchase of a subsidiary is both nonpublic and material. Trading on this information violates Standard II(A).

C is incorrect. The output from the analyst's revised model was affected by the insider information. Thus, his decision was based on material, nonpublic information, which violates Standard II(A): Material Nonpublic Information.

8 Did Omondi most likely violate the CFA Institute Code and Standards by purchasing shares for his personal and family accounts?

A Yes

B No, because the information is not definite

C No, because the board has not voted on the offer

A is correct. Omondi violated Standard II(A): Integrity of Capital Markets–Material Nonpublic Information. Members and candidates who possess material nonpublic information that could affect an investment's value must not act or cause others to act on the information. Information is "material" if its disclosure would probably affect the price of a security or if reasonable investors would want to know the information before making an investment decision. In addition to the substance and specificity of the information, the source or relative reliability of the information also determines materiality. In this case, factual information from a corporate insider regarding the purchase of a subsidiary is likely to be considered material. Although the offer is not definite or officially accepted by the board, its source, substance, and specificity are enough to make the information material. By purchasing shares informed by material nonpublic information, Omondi violated Standard II(A).

B is incorrect. Information does not have to be definite to trigger the violation; it need only be considered both material and nonpublic. In this case, factual information from a corporate insider regarding the purchase of a subsidiary is both nonpublic and material. Trading on this information violates Standard II(A).

C is incorrect. There is no requirement that the information must be about something that has actually occurred, such as the action having been taken. Under Standard II(A), "Information is considered material if its disclosure would probably have an impact on the price of a security or if reasonable investors would want to know the information before making an investment decision." Both statements are true here. Thus the information is material, and also nonpublic, so trading on this information violates Standard II(A): Material Nonpublic Information.

EDVARD STARK

4

Edvard Stark, CFA, is a private client adviser for Eyearne Bank, a small private bank.[1] In his role, Stark constructs and manages globally diversified fixed-income and equity portfolios for his clients based on the clients' respective investment objectives, risk tolerance, and time horizon. As part of his service, Stark periodically reviews client assets held outside the bank and makes recommendations for those assets. Clients have often followed Stark's advice. In providing this service, Stark has been able to cultivate stronger relationships and build his client assets under management at the bank.

Stark has been following developments in digital currencies, also known as cryptocurrencies, for some time. When the national securities regulator announced, some months ago, its decision to regulate cryptocurrencies as securities and began issuing guidance on cryptocurrency best practices, Stark concluded it was time to consider digital currencies for himself and his clients. Intrigued by the rapid appreciation in value many cryptocurrencies have exhibited, he believes cryptocurrencies may offer clients the potential for higher returns as well as diversification benefits.

Stark spends two weekends researching the top cryptocurrencies. All are digital currencies created to facilitate different types of secure transactions over the internet. He learns that cryptocurrencies are "held" in online wallets set up by individual account holders and that individuals may earn additional cryptocurrency tokens by helping administer the cryptocurrency network through an activity called "mining." Stark has read that it is difficult for later entrants to a cryptocurrency network to make money through mining because competitive pressure tends to raise the required level of capital investment over time, so he decides to focus his efforts and research on the newer cryptocurrencies.

After considering several of the newer cryptocurrencies, Stark decides the best opportunity is with a digital currency called Meerine. To limit his risk of being wrong on the cryptocurrency's potential, Stark decides to give a buy recommendation to only a few of his smallest clients. He recommends a 1% position in Meerine to these clients. Each of these clients establishes an online wallet to hold his cryptocurrency tokens and buys the recommended position in Meerine.

As Stark monitors Meerine's price over the next month, he learns more about its trading patterns and its acceptance in the marketplace. Although Meerine's price exhibits significant volatility, Stark feels optimistic about its potential. From his research, he knows there may also be an opportunity in mining Meerine's currency. Mining would involve using his own computing resources to help process Meerine's digital transactions, but in return, he could earn additional Meerine tokens for his Meerine account.

1 **Edvard Stark:** David B. Stevens, CIMC, CFA. *Ethics Cases*. © 2017 CFA Institute. All rights reserved. Consistent with the Eleventh Edition of the *Standards of Practice Handbook*.

To learn how to do this, Stark attends a local cryptocurrency conference and numerous workshops on mining. Stark believes mining Meerine's currency will give him a better understanding of cryptocurrencies and the technology supporting Meerine; this understanding, in turn, will help him make better cryptocurrency investment recommendations for his clients.

After mining Meerine's currency by running the mining software as a background process on his home computer for several months, Stark believes he is competent in his understanding of cryptocurrencies and their underlying technology. Mining has also provided him with a way to augment his salary from Eyearne Bank by adding Meerine tokens to his digital account. During this time, Meerine's price has continued to rise strongly. Stark decides to recommend a 3% Meerine position for all clients.

In his client review meetings, Stark highlights Meerine's cryptocurrency as an exciting opportunity. He illustrates the low correlation of cryptocurrencies with traditional assets and shows the strong performance of Meerine since his initial 1% buy recommendation. He shares with clients that he is mining the currency for Meerine and discusses his new 3% buy recommendation with each client. His clients, knowing little about cryptocurrencies, have few questions and no objections. Stark is pleased and feels his recommendation has been well received.

Because Meerine is a newer cryptocurrency, its daily trading volume is low and it will take his clients several days to establish their positions. As a miner, Stark receives a steady flow of Meerine tokens into his digital wallet from his mining activities. He offers his larger clients the opportunity to buy Meerine tokens directly from him so that they do not miss out on any potential appreciation of Meerine while trying to establish their positions.

Identify violations or possible violations of the Code and Standards by Stark. For each identified violation, state what actions Stark should have taken and make a short policy statement a firm could use to guide employees to help prevent similar violations in the future.

This case highlights ethical challenges individuals may face during their careers as markets evolve and innovative financial products are introduced. The violations or potential violations of the Code and Standards in this case relate to a member's duties to clients; duties to employer; duties regarding investment analysis, recommendations, and actions; and duty to disclose conflicts of interest to the employer and clients.

Duties to Clients

Standard III(B): Fair Dealing states that members and candidates must treat all clients fairly when taking investment action with regard to general purchases, new issues, or secondary offerings. Stark's offer to directly fill orders for his largest clients without making the same offer to all his clients is a breach of Standard III(B).

Standard III(B) does not state that all clients must be treated "equally." Members and candidates may differentiate their services to clients, but different levels of service must not disadvantage or negatively affect clients. When making investments in new offerings, however, members and candidates should distribute the issues to all customers for whom the investments are appropriate in a fair and equitable manner.

In this instance, Stark has clearly violated Standard III(B). Stark's offer to fill allocations from his Meerine account for only his largest clients puts his other clients at an economic disadvantage. Stark has a duty to all his clients to provide fair and impartial access to Meerine tokens.

Actions Required

Because Stark knows there is a limited market for Meerine tokens, he should either (1) offer each of his clients the opportunity to buy Meerine tokens directly from him, collect their orders, and then allocate his available tokens to each client in proportion to their planned investment or (2) not offer to sell any of his tokens to his clients.

Policy Statement for a Firm

"All client accounts participating in a new issue or security with limited liquidity will be executed as a block trade and shall receive the same execution price. All trade allocations to client accounts shall be made on a pro rata basis prior to or immediately following part or all of a block trade."

Standard III(C): Suitability obligates members and candidates who are in an investment advisory relationship with clients to consider carefully the needs, circumstances, and objectives of the clients when determining the appropriateness and suitability of a given investment. In judging the suitability of a potential investment, the member or candidate should review many aspects of the client's knowledge, experience related to investing, and financial situation. These aspects include, but are not limited to, the risk profile of the investment as compared with the constraints of the client, the impact of the investment on the diversity of the portfolio, and whether the client has the means or net worth to assume the associated risk. Although the national securities regulator is now regulating cryptocurrencies, they are still more suitable for speculation than as an investment, given that no clear consensus exists for determining future expected value for cryptocurrencies.

Although Stark has considered the potential risk reduction benefits from diversification, his recommendation that all his clients buy a 3% position in Meerine without specific regard to suitability regarding client circumstances or whether this investment is consistent with each client's written objectives, mandates, or constraints is a clear violation of Standard III(C). An additional violation of this standard is Stark's decision to initially recommend the Meerine investment only for his smallest accounts. Rather than being determined by his clients' investment objectives, including risk tolerance, his decision is driven by the desire to limit his personal and Eyearne Bank's risk of being wrong in his recommendation.

Actions Required

Although Stark is clearly excited about the possible benefits of cryptocurrencies, he needs to properly assess each client's circumstances and determine on the basis of her risk tolerance, goals, and objectives whether the client should invest in Meerine and, if so, what the appropriate level of exposure is for that client.

Policy Statement for a Firm

"When making any investment recommendations to clients, investment advisers must carefully consider the impact the proposed change will have on portfolio diversification, how the investment's risk parameters align with the client's assessed risk tolerance, and whether the proposed investment fits within the overall investment strategy, taking into account the client's time horizon, return objectives, and constraints, as well as the type and nature of the client."

Duties to Employers

Standard IV(B): Additional Compensation Arrangements requires members and candidates to obtain permission from their employer before accepting compensation or other benefits from third parties for any services that might create a conflict with their employer's interest.

Stark has begun mining Meerine for additional cryptocurrency compensation. Doing so creates a conflict of interest with Eyearne Bank, because mining Meerine's cryptocurrency involves activities that compete with Eyearne's services. Mining involves verifying transactions that occur outside of traditional banking channels. As part of normal operations, banks facilitate transactions through credit cards and checking accounts. In mining, Stark is supporting a service that is competitive with the bank, which creates a conflict. Earning outside compensation is not itself a violation of the Code and Standards, but Stark should disclose it to his employer for the consideration of conflicts.

Actions Required

Stark needs to disclose to his supervisor or the compliance department at Eyearne Bank his intention to mine Meerine and the potential earnings expected from this activity. He will need to receive written consent from Eyearne before beginning any mining activity.

Policy Statement for a Firm

"Employees must disclose any external employment or compensation arrangement to the firm and receive express written permission before undertaking any such arrangement. Failure to comply is a violation of company policy and is subject to disciplinary procedures up to and including termination."

Investment Analysis, Recommendations, and Actions

Under Standard V(A): Diligence and Reasonable Basis, members and candidates must exercise diligence, independence, and thoroughness in making investment recommendations. Although Stark had done some research before recommending that clients buy cryptocurrency, he was still in the learning process when he made the buy recommendation to his smallest clients; therefore, he is in violation of Standard V(A).

Standard V(A) does not require perfect knowledge but does require diligence and thoroughness from members and candidates in gathering as much information and knowledge as possible to inform their professional judgement before making an investment recommendation in order to have a reasonable and adequate basis for making the recommendation.

Actions Required

Stark should develop a written evaluation of cryptocurrencies and Meerine in particular, detailing the background information and decision framework that support his investment recommendation for cryptocurrencies and Meerine. Stark's report should consider risks as well as benefits.

Policy Statement for a Firm

"Purchases or recommendations to purchase are limited to securities on the 'Approved for Investment Purchase List' (Approved List). Securities can be added to the Approved List after review and approval by the Investment Committee. A written research report detailing risks and opportunities is required for evaluation by the Investment Committee. The report should also note whether the security is considered speculative or non-speculative."

Conflicts of Interest

Under Standard VI(A): Disclosure of Conflicts, members and candidates must make full and fair disclosure of all matters that could reasonably be expected to impair their independence and objectivity. Members and candidates must maintain their objectivity when rendering investment advice. Requiring members and candidates to disclose all matters that reasonably could be expected to impair the member's or candidate's objectivity allows clients to judge an adviser's motives and possible biases for themselves.

Stark's mining of Meerine and his recommendation that clients invest in Meerine is a conflict because he is advocating that his clients buy an investment with limited liquidity in which he has a personal holding. His clients' purchases would likely cause Meerine's price to rise, thereby directly benefiting Stark's position. His lack of full disclosure is a violation of Standard VI(A). Furthermore, his decision to sell some of his own cryptocurrency directly to his clients is a conflict that needs to be disclosed to all his clients who are considering his recommendation to buy Meerine, as well as to his employer, Eyearne Bank. Although he reveals his mining activity in client meetings held after his recommendation of a 3% position in Meerine for all clients, clients should be given an alternative cryptocurrency to invest in to avoid the direct conflict. Also, the information he shares in the client meetings does not fully disclose his conflicted position, because his clients have limited knowledge of cryptocurrencies and may not understand the conflict of Stark's mining activities and his investment recommendation to buy Meerine.

Because Stark's clients have a limited knowledge of cryptocurrencies, his duty to disclose his conflict of interest is of paramount importance so that his clients can fully evaluate his recommendation.

Actions Required

Stark should clearly disclose to his clients and Eyearne Bank his conflict of interest in mining Meerine and recommending Meerine for purchase to his clients. Because cryptocurrencies are relatively unfamiliar to most of his clients, he will need to make sure his clients fully understand his conflict. Before recommending Meerine to clients, Stark should also determine a suitable alternative cryptocurrency from those he researched for those clients who are uncomfortable with the conflict of interest.

Policy Statement for a Firm

"Employees shall not use their position, directly or indirectly, for private gain or financial benefit, to advance personal interests, or to obtain favors or benefits for themselves, their families, or any other person. Effective conflict management requires all employees to identify and disclose to the company's Compliance Officer all actual or potential conflicts of interest as they become aware of them. Because it is impossible to describe every conflict of interest, all employees are required to exercise sound judgment, seek advice when appropriate, escalate concerns, obtain review of certain activities as required by this policy and other applicable business and jurisdiction-specific policies and procedures, and adhere to the highest ethical standards."

5 SUBATH AGARWAY

Subath Agarway, CFA, has recently joined CrowdWisdom as vice president, and he is in charge of due diligence.[2] Agarway is the ninth employee of CrowdWisdom, a young venture capital company that matches investors with startup companies in need of capital. His position at the online company is a newly created one. As head of the due diligence function, Agarway's role is to identify suitable companies for CrowdWisdom to offer to potential investors. Agarway is the only CFA charterholder on the team, which includes two co-founders, Craig Miller and Stephane Etienne. Both Miller and Etienne have substantial experience and strong networks from working at other industry startup companies.

Since its startup four years ago, CrowdWisdom has grown rapidly, funding 50 startup companies with almost $10 million from investors through its online matching platform. CrowdWisdom's business model markets to a wide range of startup companies seeking public capital. Startups in need of funds submit a listing application to CrowdWisdom. Application approval by CrowdWisdom's due diligence function allows companies to list on the platform for a fee, thereby becoming visible to platform investors as possible investments. Investors on the CrowdWisdom platform include both sophisticated and unsophisticated investors. Owing to a successful business model, Agarway is receiving an unprecedented number of applications from startups wishing to list on the company's platform.

The company's business plan calls for aggressive growth to maintain market share and secure CrowdWisdom's next round of funding. The founders' mandate is to list 100 companies on the CrowdWisdom platform in the next 18 months. In the longer term, the founders hope to do an initial public offering of CrowdWisdom's stock.

CrowdWisdom's early success has resulted in part from Miller's and Etienne's work in attracting platform investors who are willing to capitalize young startup companies. Leveraging their collective network, the founders created a large database of potential platform investors shortly after CrowdWisdom was created. As investors began investing on the platform, the founders pioneered an "Investor Club" whose members were the most active in providing capital through the CrowdWisdom platform. Investor Club members receive access to market intelligence research in addition to the research on CrowdWisdom listed companies that Agarway prepares and posts on the website.

To keep the database growing, Miller asks Agarway to consider companies whose customers appear to be a strong fit from a potential future investor standpoint. Agarway has experience marketing equity investments to customers of platform companies and knows that many companies have successfully raised funds by soliciting their own customers to become investors. Agarway also knows that CrowdWisdom's policies must comply with rules governing marketing over the internet, which include opt-in/opt-out preferences, age of person(s) marketed to, and required disclosures.

During the next two months, Agarway reviews the presentation materials for more than 100 companies that want to list on the CrowdWisdom platform. Agarway uses a process of due diligence he developed over several years, most recently as head of research for his previous employer, FunderWise, a lesser-known crowdfunding platform.

Agarway's due diligence process consists of a two-step process he developed through trial and error at FunderWise. First, he reviews materials provided by companies to screen out those with a potential market for their product or service of less than $1 billion and those with perceived product or service viability concerns. Together, these criteria typically screen out 75% of applicant companies. Second,

2 **Subath Agarway:** Cynthia Harrington, CFA. *Ethics Cases*. © 2017 CFA Institute. All rights reserved. Consistent with the Eleventh Edition of the *Standards of Practice Handbook*.

Agarway investigates the remaining companies by closely reviewing audited financial statements and interviewing company executives and customers. He is confident in his process and has personally invested in several FunderWise listed companies using this approach.

After considerable time and effort investigating the companies that made it past the first screen, Agarway's additional research leads him to reject almost all the remaining companies. The rejected companies appear to have issues with improper revenue recognition, questionable user claims, and regulatory litigation.

Of the few remaining applications, Agarway believes one of the most promising is that of a company called Deko, an information technology startup. Deko has impressive founders, attractive prospects, and a unique product. Additionally, Deko seems to have an enviable customer base that CrowdWisdom could approach for future investor funding activity. Deko is unique in that most of its users are preteens and teenagers who love Deko's software, which allows them to create digital collections of their possessions and then share these collections with their friends online. The company's strategy is to market its crowdfunded shares through email communications to the young users. The email contains an announcement on the company's crowdfunding offer and states the offer is available to adults over the age of 18. Companies with loyal users who often bring in friends and family have proven to be among the more successful at equity crowdfunding campaigns.

During this time, Agarway is asked to take on additional responsibilities. His days and evenings include speaking with founders of listed companies, answering investor questions, and working with attorneys to finalize listing transactions for new companies.

Several months later, after the founders present at two global startup conferences, Agarway's stack of applications for review grows to 300 companies. To meet CrowdWisdom's aggressive growth goals, Miller and Etienne suggest to Agarway that he target an application acceptance rate of 10%. They suggest Agarway research at least half of the applying companies in his second-stage process to meet the 10% acceptance rate.

Agarway patiently explains his process and his challenge in finding time to review applicants. In response, the founders suggest he find ways to reduce the time spent on each application. Miller and Etienne also recommend the acceptance of two companies whose founders Miller and Etienne met at the recent conferences.

Several activities in the case are or could be in violation of the Code and Standards. Identify violations or possible violations, state what actions Agarway and the firm should take to correct the violations, and make a short policy statement a firm could use to guide employees to help prevent similar violations from occurring in the future.

This case highlights challenges individuals may face during their careers when working for younger firms whose core business may not be traditional financial services or investment management, or when working for firms where they may be the only CFA charterholder or one of just a few CFA charterholders employed.

Professionalism

Standard I(A): Knowledge of the Law requires candidates and members to understand the applicable laws and regulations of the countries and jurisdictions where they engage in professional activities. Agarway should review the global rules governing online marketing to Deko's teen and preteen customers. The company strategy of offering equity to users' parents (or other adults in the household) through communications to its teen and preteen user base may put CrowdWisdom and Agarway at risk because it is illegal in many countries to collect information on such individuals over the internet without first obtaining parental permission. Unless Agarway can confirm that Deko is in compliance with this requirement, the use of CrowdWisdom's platform to solicit preteens could be against the law. Because Deko has cleared Agarway's due diligence process, if Deko were to be added to CrowdWisdom's platform, CrowdWisdom and Deko could be at risk for prosecution.

Actions Required

Although Agarway does not need to know the laws in every jurisdiction, he does need to stay informed about relevant legal limitations. Agarway and CrowdWisdom should establish a procedure whereby employees are regularly informed about changes in applicable laws and regulations. CrowdWisdom should also have legal counsel available to review planned additions to the platform to ensure that the company's strategy is not in conflict with relevant law.

Policy Statement for a Firm

"When determining whether a company should be included on the platform, careful consideration must be taken to determine whether the company's business strategy violates laws related to marketing and solicitation, particularly if the strategy targets minors or vulnerable adults (those with physical or mental disabilities). If the strategy targets minors or vulnerable adults, legal counsel will be consulted before listing the company on the platform."

Standard I(B): Independence and Objectivity obligates members and candidates to use reasonable care and judgment to achieve and maintain independence and objectivity in their professional activities. Pressure by CrowdWisdom's founders to modify Agarway's due diligence process to increase the number of listing approvals and shorten the review time frame is likely to put Agarway's independence and objectivity at risk.

Actions Required

Agarway and CrowdWisdom's senior leaders need to create and document a company-approved due diligence process, which will likely blend Agarway's past work with input from CrowdWisdom's founders. As a CFA charterholder, Agarway will need to be comfortable that the process has a reasonable basis and can be applied objectively.

Policy Statement for a Firm

"The selection of companies for inclusion on the platform will comply with the due diligence process approved by the firm's Board as detailed in the Selection Due Diligence Memorandum approved on 25 January 20XX."

Conflicts of Interest

Standard VI(A): Disclosure of Conflicts requires members and candidates to make full and fair disclosure of all matters that could reasonably be expected to impair their independence and objectivity. Identifying and managing conflicts is a reality of working in the investment industry, where conflicts are often present. When a conflict cannot be reasonably avoided, clear and complete disclosure of its existence is necessary. Some possible conflicts of interest exist in this scenario: CrowdWisdom's Investor Club selective access to additional market intelligence research and Agarway's personal investment in several companies that could be competitors of firms he is evaluating for the platform or future additions to the CrowdWisdom platform. Conflicts of interest may be inevitable and must be disclosed in a timely manner so that all parties involved can understand the circumstances and potential effects.

Actions Required

CrowdWisdom's Investor Club, which provides select investors with preferential access to additional market intelligence research, needs to be disclosed so that all investors can understand and evaluate the circumstances, the possible impact, and the potential disadvantage they may be placed at relative to Investor Club members.

In his personal portfolio, Agarway has invested in companies that could be competitors of firms he is reviewing in his due diligence work. His personal investments need to be disclosed to both his supervisor and CrowdWisdom's compliance officer. Additionally, if these firms also list on the CrowdWisdom platform, Agarway's personal investments would need to be disclosed to CrowdWisdom's users so that they can evaluate the independence and objectivity of each company's inclusion on the platform.

Policy Statement for a Firm

"Employees of the firm must disclose all personal investment holdings to the company's Compliance Officer, and that disclosure must be updated quarterly for public stocks and when invested for private holdings. All employee investments in companies that raise funds through the firm must be approved in advance by the Compliance Officer—or in the case of the founders, by the Board—before the transactions' closing and must be communicated to the firm's clients."

Quantitative Methods

STUDY SESSION

Study Session 2	Quantitative Methods (1)
Study Session 3	Quantitative Methods (2)

TOPIC LEVEL LEARNING OUTCOME

The candidate should be able to explain regression, time series analysis, and simulation and their uses in investment decision-making. The candidate should also be able to interpret the results of a regression, time-series analysis, or simulation.

Quantitative methods such as correlation, regression, time series, and simulation provide the means to identify and assess the relationships that exist between variables. Measuring the direction and strength of these relationships, with some level of confidence, can provide valuable insights for many investment-related activities.

QUANTITATIVE METHODS

STUDY SESSION

2

Quantitative Methods (1)

This study session provides coverage on how linear regression and time-series analysis are used as tools in financial analysis for identifying relationships among variables. The session begins by examining linear regression with a single (independent) variable to explain or predict the value of another (dependent) variable. Multiple regression, using more than one independent variable to explain or predict a dependent variable, is explored next. Time-series analysis, in which the dependent variable's past values are included as independent variables, concludes the session.

READING ASSIGNMENTS

Reading 4	Introduction to Linear Regression by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA
Reading 5	Multiple Regression by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA
Reading 6	Time-Series Analysis by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA

READING

4

Introduction to Linear Regression

by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA

Richard A. DeFusco, PhD, CFA, is at the University of Nebraska-Lincoln (USA). Dennis W. McLeavey, DBA, CFA, is at the University of Rhode Island (USA). Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). David E. Runkle, PhD, CFA, is at Trilogy Global Advisors (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	**a.** distinguish between the dependent and independent variables in a linear regression;
☐	**b.** explain the assumptions underlying linear regression and interpret regression coefficients;
☐	**c.** calculate and interpret the standard error of estimate, the coefficient of determination, and a confidence interval for a regression coefficient;
☐	**d.** formulate a null and alternative hypothesis about a population value of a regression coefficient and determine the appropriate test statistic and whether the null hypothesis is rejected at a given level of significance;
☐	**e.** calculate the predicted value for the dependent variable, given an estimated regression model and a value for the independent variable;
☐	**f.** calculate and interpret a confidence interval for the predicted value of the dependent variable;
☐	**g.** describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret the *F*-statistic;
☐	**h.** describe limitations of regression analysis.

1 INTRODUCTION

As a financial analyst, you will often need to examine whether a variable is able to explain another variable. For example, you might want to know whether the spread between a company's return on invested capital and its cost of capital helps to explain the company's value in the marketplace. Regression analysis is a tool for examining this issue. This reading[1] introduces basic concepts in regression analysis, a powerful technique for examining the ability of one or more variables (independent variables) to explain or predict another variable (the dependent variable).

The reading is organized as follows. Section 2 describes linear regression with one independent variable. Section 3 explains the assumptions of the linear regression model. Sections 4 and 5 respectively explain the standard error of estimate and the coefficient of determination. Section 6 addresses testing hypotheses concerning the population values of the intercept and slope coefficient of a regression model. Section 7 describes the uses of analysis of variance (ANOVA) in a regression. Section 8 explains prediction intervals and section 9 describes limitations of regression analysis. Section 10 summarizes the reading.

2 LINEAR REGRESSION

Linear regression with one independent variable, sometimes called simple linear regression, models the relationship between two variables as a straight line. When the linear relationship between the two variables is significant, linear regression provides a simple model for forecasting the value of one variable, known as the dependent variable, given the value of the second variable, known as the independent variable. The following sections explain linear regression in more detail.

2.1 Linear Regression with One Independent Variable

As a financial analyst, you will often want to understand the relationship between financial or economic variables, or to predict the value of one variable using information about the value of another variable. For example, you may want to know the impact of changes in the 10-year Treasury bond yield on the earnings yield of the S&P 500 (the earnings yield is the reciprocal of the price-to-earnings ratio). If the relationship between those two variables is linear, you can use linear regression to summarize it.

Linear regression allows us to use one variable to make predictions about another, test hypotheses about the relation between two variables, and quantify the strength of the relationship between the two variables. The remainder of this reading focuses on linear regression with a single independent variable. In the next reading, we will examine regression with more than one independent variable.

Regression analysis begins with the dependent variable (denoted Y), the variable that you are seeking to explain. The independent variable (denoted X) is the variable you are using to explain changes in the dependent variable. For example, you might try to explain small-stock returns (the dependent variable) based on returns to the S&P 500 (the independent variable). Or you might try to explain inflation (the dependent variable) as a function of growth in a country's money supply (the independent variable).

1 Examples in this reading were updated in 2018 by Professor Sanjiv Sabherwal of the University of Texas, Arlington.

Linear regression assumes a linear relationship between the dependent and the independent variables. The following regression equation describes that relation:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i,\ i = 1, \ldots, n \tag{1}$$

This equation states that the **dependent variable**, Y, is equal to the intercept, b_0, plus a slope coefficient, b_1, times the **independent variable**, X, plus an **error term**, ε. The error term represents the portion of the dependent variable that cannot be explained by the independent variable. We refer to the intercept b_0 and the slope coefficient b_1 as the **regression coefficients**.

Regression analysis uses two principal types of data: cross-sectional and time series. Cross-sectional data involve many observations on X and Y for the same time period. Those observations could come from different companies, asset classes, investment funds, people, countries, or other entities, depending on the regression model. For example, a cross-sectional model might use data from many companies to test whether predicted earnings-per-share growth explains differences in price-to-earnings ratios (P/Es) during a specific time period. The word "explain" is frequently used in describing regression relationships. One estimate of a company's P/E that does not depend on any other variable is the average P/E. If a regression of a P/E on an independent variable tends to give more accurate estimates of P/E than just assuming that the company's P/E equals the average P/E, we say that the independent variable helps *explain* P/Es because using that independent variable improves our estimates. Finally, note that if we use cross-sectional observations in a regression, we usually denote the observations as $i = 1, 2, \ldots, n$.

Time-series data use many observations from different time periods for the same company, asset class, investment fund, person, country, or other entity, depending on the regression model. For example, a time-series model might use monthly data from many years to test whether US inflation rates determine US short-term interest rates.[2] If we use time-series data in a regression, we usually denote the observations as $t = 1, 2, \ldots, T$.[3]

Exactly how does linear regression estimate b_0 and b_1? Linear regression, also known as linear least squares, computes a line that best fits the observations; it chooses values for the intercept, b_0, and slope, b_1, that minimize the sum of the squared vertical distances between the observations and the regression line. Linear regression chooses the **estimated parameters** or **fitted parameters** $\hat{b}_0$ and $\hat{b}_1$ in Equation 1 to minimize[4]

$$\sum_{i=1}^{n}\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2 \tag{2}$$

In this equation, the term $\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2$ means (dependent variable – predicted value of dependent variable)^2. Using this method to estimate the values of $\hat{b}_0$ and $\hat{b}_1$, we can fit a line through the observations on X and Y that best explains the value that Y takes for any particular value of X.[5]

2 A mix of time-series and cross-sectional data, also known as panel data, is now frequently used in financial analysis. The analysis of panel data is an advanced topic that Greene (2018) discusses in detail.

3 In this reading, we primarily use the notation $i = 1, 2, \ldots, n$ even for time series to prevent confusion that would be caused by switching back and forth between different notations.

4 Hats over the symbols for coefficients indicate estimated values.

5 For a discussion of the precise statistical sense in which the estimates of b_0 and b_1 are optimal, see Greene (2018).

Note that we never observe the population parameter values b_0 and b_1 in a regression model. Instead, we observe only $\hat{b}_0$ and $\hat{b}_1$, which are estimates of the population parameter values. Thus predictions must be based on the parameters' estimated values, and testing is based on estimated values in relation to the hypothesized population values.

Suppose that we want to estimate the regression relation between the annual rate of inflation (the dependent variable) and annual rate of money supply growth (the independent variable) for six industrialized countries. Table 1 shows the average annual growth rate in the money supply and the average annual inflation rate from 1980 to 2016 for the six countries (n = 6).

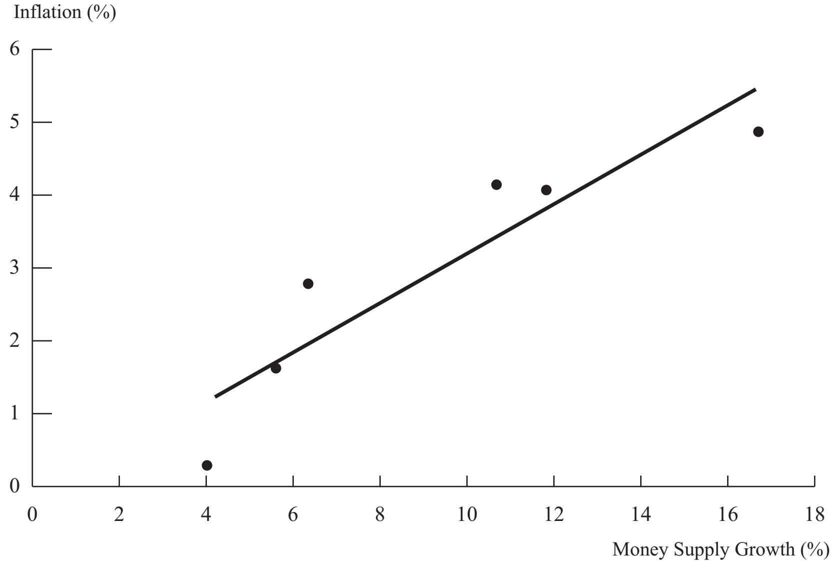
Table 1 Annual Money Supply Growth Rate and Inflation Rate by Country, 1980–2016

Country	Money Supply Growth Rate (%)	Inflation Rate (%)
Australia	10.68	4.14
Japan	4.00	0.32
South Korea	16.65	4.86
Switzerland	5.54	1.67
United Kingdom	11.81	4.10
United States	6.28	2.77
Average	9.16	2.98

Source: The World Bank.

Figure 1 gives a visual example of how linear regression works. The figure shows a scatter plot constructed by using the data for each country in Table 1 to mark a point on a graph, and the linear regression that results from estimating the following equation: Long-term rate of inflation = b_0 + b_1 (Long-term rate of money supply growth) + ε.

Figure 1 Fitted Regression Line Explaining the Inflation Rate Using Growth in the Money Supply by Country, 1980–2016



Source: The World Bank.

The vertical distance from each of the six data points to the fitted regression line is the regression residual, which is the difference between the actual value of the dependent variable and the predicted value of the dependent variable made by the regression equation. Linear regression chooses the estimated coefficients $\hat{b}_0$ and $\hat{b}_1$ in Equation 1 such that the sum of the squared vertical distances is minimized. The estimated regression equation is Long-term inflation = –0.0008 + 0.3339 (Long-term money supply growth).[6]

According to this regression equation, if the long-term money supply growth is 0 for any country, the long-term rate of inflation in that country is predicted to be –0.08 percent. For every 1-percentage-point increase in the long-term rate of money supply growth for a country (say, from 3 percent to 4 percent), the long-term inflation rate is predicted to increase by 0.3339 percentage points. In a regression such as this one, which contains one independent variable, the slope coefficient equals $\text{Cov}(Y,X)/\text{Var}(X)$. Table 2 shows how to compute the slope coefficient from the data in Table 1.[7] The individual observations on countries' average annual money supply growth from 1980–2016 are denoted X_i, and individual observations on countries' annual average inflation rate from 1980–2016 are denoted Y_i. The next two columns show the calculations for the inputs to the slope coefficient: the sample covariance and the sample variance of X. The last column is included because we need the sample standard deviation of Y to compute the coefficient of determination later in Section 5.

6 We entered the monthly rates as decimals. Also, we used rounded numbers in the formulas discussed later to estimate the regression equation.

7 We have not used full precision in the table's calculations. We used the average value of the money supply growth rate of 0.5496/6 = 0.0916, rounded to four decimal places, in the cross-product and squared deviation calculations, and similarly, we used the mean inflation rate as rounded to 0.0298 in those calculations. We computed standard deviation as the square root of variance rounded to six decimal places, as shown in the table. Had we used full precision in all calculations, some of the table's entries would be slightly different but would not materially affect our conclusions.

Table 2 Sample Covariance and Sample Variances: Annual Money Supply Growth Rate and Inflation Rate by Country, 1980–2016

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Squared Deviations $(X_i - \bar{X})^2$	Squared Deviations $(Y_i - \bar{Y})^2$
Australia	0.1068	0.0414	0.0001763	0.0002310	0.0001346
Japan	0.0400	0.0032	0.0013726	0.0026626	0.0007076
South Korea	0.1665	0.0486	0.0014081	0.0056100	0.0003534
Switzerland	0.0554	0.0167	0.0004742	0.0013104	0.0001716
United Kingdom	0.1181	0.0410	0.0002968	0.0007023	0.0001254
United States	0.0628	0.0277	0.0000605	0.0008294	0.0000044
Sum	0.5496	0.1786	0.0037885	0.0113457	0.0014970
Average	0.0916	0.0298			
Covariance			**0.0007577**		
Variance				**0.0022691**	0.0002994
Standard deviation				0.0476356	0.0173033

Notes:

1 Divide the cross-product sum by $n - 1$ (with $n = 6$) to obtain the covariance of X and Y.

2 Divide the squared deviations sums by $n - 1$ (with $n = 6$) to obtain the variances of X and Y.

Source: The World Bank.

$\text{Cov}(Y,X) = 0.0007577$

$\text{Var}(X) = 0.0022691$

$\text{Cov}(Y,X)/\text{Var}(X) = 0.0007577/0.0022691$

$\hat{b}_1 = 0.3339$

In a linear regression, the fitted regression line passes through the point corresponding to the means of the dependent and the independent variables. As shown in Table 2 (excerpted below), from 1980 to 2016, the mean long-term growth rate of the money supply for these six countries was 9.16 percent, whereas the mean long-term inflation rate was 2.98 percent.

Table 2 (excerpted)

	Money Supply Growth Rate	Inflation Rate
Average	9.16%	2.98%

Because the point (9.16, 2.98) lies on the regression line $\hat{b}_0 = \overline{Y} - \hat{b}_1\overline{X}$, we can solve for the intercept using this point as follows:

$$\hat{b}_0 = 0.0298 - 0.3339(0.0916) = -0.0008$$

We are showing how to solve the linear regression equation step by step to make the source of the numbers clear. Typically, an analyst will use the data analysis function on a spreadsheet or a statistical package to perform linear regression analysis. Later, we will discuss how to use regression residuals to quantify the uncertainty in a regression model.

3 ASSUMPTIONS OF THE LINEAR REGRESSION MODEL

We have discussed how to interpret the coefficients in a linear regression model. Now we turn to the statistical assumptions underlying this model. Suppose that we have n observations on both the dependent variable, Y, and the independent variable, X, and we want to estimate Equation 1:

$$Y_i = b_0 + b_1X_i + \varepsilon_i,\ i = 1, \ldots, n$$

To be able to draw valid conclusions from a linear regression model with a single independent variable, we need to make the following six assumptions, known as the classic normal linear regression model assumptions:

1 The relationship between the dependent variable, Y, and the independent variable, X is linear in the parameters b_0 and b_1. This requirement means that b_0 and b_1 are raised to the first power only and that neither b_0 nor b_1 is multiplied or divided by another regression parameter (as in b_0/b_1, for example). The requirement does not exclude X from being raised to a power other than 1.

2 The independent variable, X, is not random.[8]

3 The expected value of the error term is 0: $E(\varepsilon) = 0$.

4 The variance of the error term is the same for all observations: $E\left(\varepsilon_i^2\right) = \sigma_\varepsilon^2$, i = 1, …, n.

5 The error term, ε, is uncorrelated across observations. Consequently, $E(\varepsilon_i\varepsilon_j) = 0$ for all i not equal to j.[9]

6 The error term, ε, is normally distributed.[10]

Now we can take a closer look at each of these assumptions.

8 Although we assume that the independent variable in the regression model is not random, that assumption is clearly often not true. For example, it is unrealistic to assume that the monthly returns to the S&P 500 are not random. If the independent variable is random, then is the regression model incorrect? Fortunately, no. Econometricians have shown that even if the independent variable is random, we can still rely on the results of regression models given the crucial assumption that the error term is uncorrelated with the independent variable. The mathematics underlying this reliability demonstration, however, are quite difficult. See, for example, Greene (2018).

9 $\text{Var}(\varepsilon_i) = E[\varepsilon_i - E(\varepsilon_i)]^2 = E(\varepsilon_i - 0)^2 = E(\varepsilon_i)^2$. $\text{Cov}(\varepsilon_i, \varepsilon_j) = E\{[\varepsilon_i - E(\varepsilon_i)][\varepsilon_j - E(\varepsilon_j)]\} = E[(\varepsilon_i - 0)\ (\varepsilon_j - 0)] = E(\varepsilon_i\varepsilon_j) = 0$.

10 If the regression errors are not normally distributed, we can still use regression analysis. Econometricians who dispense with the normality assumption use chi-square tests of hypotheses rather than F-tests. This difference usually does not affect whether the test will result in a particular null hypothesis being rejected.

Assumption 1 is critical for a valid linear regression. If the relationship between the independent and dependent variables is nonlinear in the parameters, then estimating that relation with a linear regression model will produce invalid results. For example, $Y_i = b_0 e^{b_1 X_i} + \varepsilon_i$ is nonlinear in b_1, so we could not apply the linear regression model to it.[11]

Even if the dependent variable is nonlinear, linear regression can be used as long as the regression is linear in the parameters. So, for example, linear regression can be used to estimate the equation $Y_i = b_0 + b_1 X_i^2 + \varepsilon_i$.

Assumptions 2 and 3 ensure that linear regression produces the correct estimates of b_0 and b_1.

Assumptions 4, 5, and 6 let us use the linear regression model to determine the distribution of the estimated parameters $\hat{b}_0$ and $\hat{b}_1$ and thus test whether those coefficients have a particular value.

- Assumption 4, that the variance of the error term is the same for all observations, is also known as the homoskedasticity assumption. The reading on multiple regression discusses how to test for and correct violations of this assumption.
- Assumption 5, that the errors are uncorrelated across observations, is also necessary for correctly estimating the variances of the estimated parameters $\hat{b}_0$ and $\hat{b}_1$. The reading on multiple regression discusses violations of this assumption.
- Assumption 6, that the error term is normally distributed, allows us to easily test a particular hypothesis about a linear regression model.[12]

EXAMPLE 1

Evaluating Economic Forecasts

If economic forecasts were completely accurate, every prediction of a change in an economic variable in a quarter would exactly match the actual change that occurs in that quarter. Even though forecasts may be inaccurate, we hope at least that they are unbiased—that is, that the expected value of the forecast error is zero. An unbiased forecast can be expressed as E(Actual change − Predicted change) = 0. In fact, most evaluations of forecast accuracy test whether forecasts are unbiased.[13]

In the euro area, the Survey of Professional Forecasters (SPF) gathers professional forecasters' predictions about many economic variables.[14] Since 1999, SPF has gathered predictions on the euro area inflation rate using the change in the Harmonised Index of Consumer Prices (HICP) for the prices of consumer goods and services acquired by households to measure inflation. Figure 2 shows a scatter plot of the mean forecast made in the first quarter of a year for the percentage change in HICP during that year and the actual percentage change

11 For more information on nonlinearity in the parameters, see Gujarati and Porter (2011).

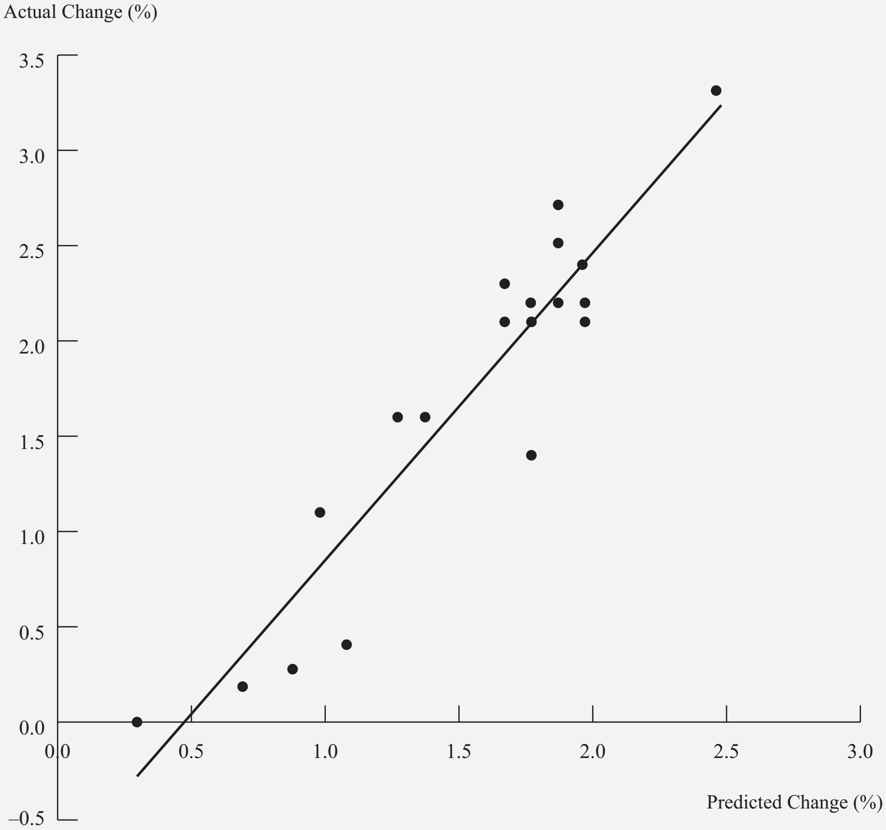
12 For large sample sizes, we may be able to drop the assumption of normality by appeal to the central limit theorem; see Greene (2018). Asymptotic theory shows that, in many cases, the test statistics produced by standard regression programs are valid even if the error term is not normally distributed. Non-normality of some financial time series can be quite severe. With severe non-normality, even with a relatively large number of observations, invoking asymptotic theory to justify using test statistics from linear regression models may be inappropriate.

13 See, for example, Keane and Runkle (1990).

14 The euro area survey is conducted by the European Central Bank (ECB). A survey of professional forecasters is also conducted by the Federal Reserve Bank of Philadelphia for the United States.

in HICP, from 1999 through 2017, along with the fitted regression line for the equation Actual percentage change = b_0 + b_1 (Predicted percentage change) + ε. If the forecasts are unbiased, the intercept, b_0, should be 0 and the slope, b_1, should be 1. We should also find E(Actual change − Predicted change) = 0. If forecasts are actually unbiased, as long as b_0 = 0 and b_1 = 1, the error term [Actual change − b_0 − b_1(Predicted change)] will have an expected value of 0, as required by Assumption 3 of the linear regression model. With unbiased forecasts, any other values of b_0 and b_1 would yield an error term with an expected value different from 0.

Figure 2 Actual Change in Euro Area HICP versus Predicted Change



Source: European Central Bank.

If b_0 = 0 and b_1 = 1, our best guess of actual change in HICP would be 0 if professional forecasters' predictions of change in HICP were 0. For every 1-percentage-point increase in the prediction of change by the professional forecasters, the regression model would predict a 1-percentage-point increase in actual change.

The fitted regression line in Figure 2 comes from the equation Actual change = −0.7679 + 1.5922(Predicted change). It seems that the estimated values of b_0 and b_1 are not particularly close to the values b_0 = 0 and b_1 = 1 that are consistent with unbiased forecasts. Later in this reading, we discuss how to test the hypotheses that b_0 = 0 and b_1= 1.

4 THE STANDARD ERROR OF ESTIMATE

The linear regression model sometimes describes the relationship between two variables quite well, but sometimes it does not. We must be able to distinguish between these two cases to use regression analysis effectively. Therefore, in this section and the next, we discuss statistics that measure how well a given linear regression model captures the relationship between the dependent and independent variables.

Figure 2, for example, shows what appears to be a strong relation between predicted inflation and actual inflation. If we knew professional forecasters' predictions for inflation in a particular quarter, we would be reasonably certain that we could use this regression model to forecast actual inflation relatively accurately.

In other cases, however, the relation between the dependent and independent variables is not strong. Figure 3 shows a scatter plot of the monthly returns of the Standard & Poor's 500 Index and the monthly inflation rate in the United States from January 1990 through December 2017, along with the fitted regression line for the equation: Returns to S&P 500 = b_0 + b_1 (Rate of Inflation) + ε. In this figure, the actual observations are generally much farther from the fitted regression line than in Figure 2. Using the estimated regression equation to predict monthly stock returns assuming a particular level of inflation might result in an inaccurate forecast.

As noted, the regression relation in Figure 3 appears less precise than that in Figure 2. The standard error of estimate (sometimes called the standard error of the regression) measures this uncertainty. This statistic is very much like the standard deviation for a single variable, except that it measures the standard deviation of $\hat{\varepsilon}_i$, the residual term in the regression.

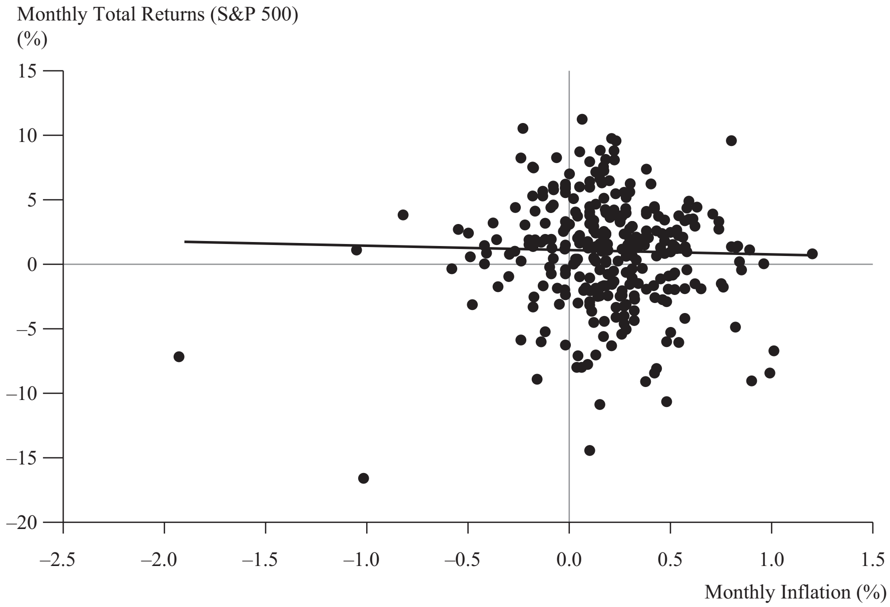
The formula for the standard error of estimate (SEE) for a linear regression model with one independent variable is

$$\text{SEE} = \left(\frac{\sum_{i=1}^{n} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i \right)^2}{n-2} \right)^{1/2} = \left(\frac{\sum_{i=1}^{n} \left(\hat{\varepsilon}_i \right)^2}{n-2} \right)^{1/2} \tag{3}$$

In the numerator of this equation, we are computing the difference between the dependent variable's actual value for each observation and its predicted value $\left(\hat{b}_0 + \hat{b}_1 X_i\right)$ for each observation. The difference between the actual and predicted values of the dependent variable is the regression residual, $\hat{\varepsilon}_i$, which is often referred to as the error term.

Equation 3 looks very much like the formula for computing a standard deviation, except that $n - 2$ appears in the denominator instead of $n - 1$. We use $n - 2$ because the sample includes n observations and the linear regression model estimates two parameters ($\hat{b}_0$ and $\hat{b}_1$); the difference between the number of observations and the number of parameters is $n - 2$. This difference is also called the degrees of freedom; it is the denominator needed to ensure that the estimated standard error of estimate is unbiased.

Figure 3 Fitted Regression Line Explaining Stock Returns by Inflation during 1990–2017



Sources: Bureau of Labor Statistics and finance.yahoo.com as of 23 June 2018.

EXAMPLE 2

Computing the Standard Error of Estimate

Recall that the estimated regression equation for the inflation and money supply growth data shown in Figure 1 was $Y_i = -0.0008 + 0.3341X_i$. Table 3 uses this estimated equation to compute the data needed for the standard error of estimate.

Table 3 Computing the Standard Error of Estimate

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Predicted Inflation Rate $\hat{Y}_i$	Regression Residual $Y_i - \hat{Y}_i$	Squared Residual $(Y_i - \hat{Y}_i)^2$
Australia	0.1068	0.0414	0.0349	0.0065	0.000042
Japan	0.0400	0.0032	0.0126	–0.0094	0.000088
South Korea	0.1665	0.0486	0.0548	–0.0062	0.000038
Switzerland	0.0554	0.0167	0.0177	–0.0010	0.000001
United Kingdom	0.1181	0.0410	0.0387	0.0023	0.000005

(continued)

Table 3 (Continued)

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Predicted Inflation Rate $\hat{Y}_i$	Regression Residual $Y_i - \hat{Y}_i$	Squared Residual $\left(Y_i - \hat{Y}_i\right)^2$
United States	0.0628	0.0277	0.0202	0.0075	0.000056
Sum					0.000230

Source: The World Bank.

The first and second columns of numbers in Table 3 show the long-term money supply growth rates, X_i, and long-term inflation rates, Y_i, for the six countries. The third column of numbers shows the predicted value of the dependent variable from the fitted regression equation for each observation. For the United States, for example, the predicted value of long-term inflation is –0.0008 + 0.3339(0.0628) = 0.0202 or 2.02 percent. The next-to-last column contains the regression residual, which is the difference between the actual value of the dependent variable, Y_i, and the predicted value of the dependent variable, $\left(\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i\right)$. So, for the United States, the residual is 0.0277 – 0.0202 = 0.0075 or 0.75 percent. The last column contains the squared regression residual. The sum of the squared residuals is 0.000230. Applying the formula for the standard error of estimate, we obtain

$\left(\dfrac{0.000230}{6-2}\right)^{1/2} = 0.007583$ Thus the standard error of estimate is about 0.76 percent.

Later, we will combine this estimate with estimates of the uncertainty about the parameters in this regression to determine confidence intervals for predicting inflation rates from money supply growth. We will see that smaller standard errors result in more accurate predictions.

5 THE COEFFICIENT OF DETERMINATION

Although the standard error of estimate gives some indication of how certain we can be about a particular prediction of Y using the regression equation, it still does not tell us how well the independent variable explains variation in the dependent variable. The coefficient of determination does exactly this: It measures the fraction of the total variation in the dependent variable that is explained by the independent variable.

We can compute the coefficient of determination in two ways. The simpler method, which can be used in a linear regression with one independent variable, is to square the correlation coefficient between the dependent and independent variables. The historical or sample correlation between two variables can be computed by dividing the sample covariance between those two variables by the product of sample standard deviations of the two variables. For example, the correlation coefficient between the long-term rate of money growth and the long-term rate of inflation between 1980 and 2016 for six industrialized countries is computed using numbers reported in Table 2 as 0.000758/(0.047634 × 0.017292) = 0.9203. Thus, the coefficient of determination in the regression shown in Figure 1 is $(0.9203)^2 = 0.8470$. So, in this regression, the long-term rate of money supply growth explains approximately 84.7 percent of the variation in the long-term rate of inflation across the countries between 1980 and 2016. (Relatedly, note that the square root of the coefficient of determination in a one-independent-variable linear regression, after attaching the sign of the estimated slope coefficient, gives the correlation coefficient between the dependent and independent variables.)

The problem with this method is that it cannot be used when we have more than one independent variable.[15] Therefore, we need an alternative method of computing the coefficient of determination for multiple independent variables. We now present the logic behind that alternative.

If we did not know the regression relationship, our best guess for the value of any particular observation of the dependent variable would simply be $\bar{Y}$, the mean of the dependent variable. One measure of accuracy in predicting Y_i based on $\bar{Y}$ is the sample variance of Y_i, $\sum_{i=1}^{n}\frac{(Y_i - \bar{Y})^2}{n-1}$. An alternative to using $\bar{Y}$ to predict a particular observation Y_i is using the regression relationship to make that prediction. In that case, our predicted value would be $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$. If the regression relationship works well, the error in predicting Y_i using $\hat{Y}_i$ should be much smaller than the error in predicting Y_i using $\bar{Y}$. If we call $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ the total variation of Y and $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ the unexplained variation from the regression, then we can measure the explained variation from the regression using the following equation:

$$\text{Total variation} = \text{Unexplained variation} + \text{Explained variation} \tag{4}$$

The coefficient of determination is the fraction of the total variation that is explained by the regression. This gives us the relationship

$$\begin{aligned} R^2 &= \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} \\ &= 1 - \frac{\text{Unexplained variation}}{\text{Total variation}} \end{aligned} \tag{5}$$

Note that total variation equals explained variation plus unexplained variation, as shown in Equation 4. Most regression programs report the coefficient of determination as R^2.[16]

EXAMPLE 3

Inflation Rate and Growth in the Money Supply

Using the data in Table 3, we can see that the unexplained variation from the regression, which is the sum of the squared residuals, equals 0.000230. Table 4 shows the computation of total variation in the dependent variable, the long-term rate of inflation.

15 We will discuss such models in the reading on multiple regression.

16 As we illustrate in the tables of regression output later in this reading, regression programs also report multiple R, which is the correlation between the actual values and the forecast values of Y. The coefficient of determination is the square of multiple R.

Table 4 Computing Total Variation

Country	Money Supply Growth Rate X_i	Inflation Rate Y_i	Deviation from Mean $Y_i - \bar{Y}$	Squared Deviation $\left(Y_i - \bar{Y}\right)^2$
Australia	0.1068	0.0414	0.0116	0.000135
Japan	0.0400	0.0032	−0.0266	0.000708
South Korea	0.1665	0.0486	0.0188	0.000353
Switzerland	0.0554	0.0167	−0.0131	0.000172
United Kingdom	0.1181	0.0410	0.0112	0.000125
United States	0.0628	0.0277	−0.0021	0.000004
	Average:	0.0298	Sum:	0.001497

Source: International Monetary Fund.

The average inflation rate for this period is 2.98 percent. The next-to-last column shows the amount each country's long-term inflation rate deviates from that average; the last column shows the square of that deviation. The sum of those squared deviations is the total variation in Y for the sample (0.001497), shown in Table 4.

Compute the coefficient of determination for the regression.

Solution:

The coefficient of determination for the regression is

$$\frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = \frac{0.001497 - 0.000230}{0.001497} = 0.8464$$

Note that this method gives the same result that we obtained earlier.[17] We will use this method again in the reading on multiple regression; when we have more than one independent variable, this method is the only way to compute the coefficient of determination.

6 HYPOTHESIS TESTING

In this section, we address testing hypotheses concerning the population values of the intercept or slope coefficient of a regression model. This topic is critical in practice. For example, we may want to check a stock's valuation using the capital asset pricing model; we hypothesize that the stock has a market-average beta or level of systematic risk. Or we may want to test the hypothesis that economists' forecasts of the inflation rate are unbiased (not overestimates or underestimates, on average). In each case, does the evidence support the hypothesis? Questions such as these can be addressed with hypothesis tests within a regression model. Such tests are often t-tests of the value of the intercept or slope coefficient(s). To understand the concepts involved in this test, it is useful to first review a simple, equivalent approach based on confidence intervals.

17 The slight difference is due to rounding.

We can perform a hypothesis test using the confidence interval approach if we know three things: 1) the estimated parameter value, $\hat{b}_0$ or $\hat{b}_1$, 2) the hypothesized value of the parameter, b_0 or b_1, and 3) a confidence interval around the estimated parameter. A confidence interval is an interval of values that we believe includes the true parameter value, b_1, with a given degree of confidence. To compute a confidence interval, we must select the significance level for the test and know the standard error of the estimated coefficient.

Suppose we regress a stock's returns on a stock market index's returns and find that the slope coefficient ($\hat{b}_1$) is 1.5 with a standard error ($s_{\hat{b}_1}$) of 0.200. Assume we used 62 monthly observations in our regression analysis. The hypothesized value of the parameter (b_1) is 1.0, the market average slope coefficient. The estimated and the population slope coefficients are often called beta, because the population coefficient is often represented by the Greek symbol beta (β) rather than the b_1 we use in this reading. Our null hypothesis is that $b_1 = 1.0$ and $\hat{b}_1$ is the estimate for b_1. We will use a 95 percent confidence interval for our test, or we could say that the test has a significance level of 0.05.

Our confidence interval will span the range $\hat{b}_1 - t_c s_{\hat{b}_1}$ to $\hat{b}_1 + t_c s_{\hat{b}_1}$ or

$$\hat{b}_1 \pm t_c s_{\hat{b}_1} \tag{6}$$

where t_c is the critical t value.[18] The critical value for the test depends on the number of degrees of freedom for the t-distribution under the null hypothesis. The number of degrees of freedom equals the number of observations minus the number of parameters estimated. In a regression with one independent variable, there are two estimated parameters, the intercept term and the coefficient on the independent variable. For 62 observations and two parameters estimated in this example, we have 60 degrees of freedom (62 – 2). For 60 degrees of freedom, the table of critical values in the back of the book shows that the critical t-value at the 0.05 significance level is 2.00. Substituting the values from our example into Equation 6 gives us the interval

$$\begin{aligned}\hat{b}_1 \pm t_c s_{\hat{b}_1} &= 1.5 \pm 2.00(0.200)\\ &= 1.5 \pm 0.400\\ &= 1.10 \text{ to } 1.90\end{aligned}$$

A 95% confidence interval is the interval, based on the sample value, that we would expect to include the population value with a 95% degree of confidence. Because we are testing the null hypothesis that $b_1 = 1.0$ and because our confidence interval does not include 1.0, we can reject the null hypothesis.

In practice, the most common way to test a hypothesis using a regression model is with a t-test of significance. To test the hypothesis, we can compute the statistic

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \tag{7}$$

18 We use the t-distribution for this test because we are using a sample estimate of the standard error, s_b, rather than its true (population) value.

This test statistic has a t-distribution with $n - 2$ degrees of freedom because two parameters were estimated in the regression. We compare the absolute value of the t-statistic to t_c. If the absolute value of t is greater than t_c, then we can reject the null hypothesis. Substituting the values from the above example into this relationship gives the t-statistic associated with the test that the stock's beta equals 1.0 (b_1 = 1.0).

$$\begin{aligned} t &= \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \\ &= (1.5 - 1.0)/0.200 \\ &= 2.50 \end{aligned}$$

Because $t > t_c$, we reject the null hypothesis that b_1 = 1.0.

The t-statistic in the example above is 2.50, and at the 0.05 significance level, t_c = 2.00; thus we reject the null hypothesis because $t > t_c$. This statement is equivalent to saying that we are 95 percent confident that the interval for the slope coefficient does not contain the value 1.0. If we were performing this test at the 0.01 level, however, t_c would be 2.66 and we would not reject the hypothesis because t would not be greater than t_c at this significance level. A 99 percent confidence interval for the slope coefficient does contain the value 1.0.

The choice of significance level is always a matter of judgment. When we use higher levels of confidence, the t_c increases. This choice leads to wider confidence intervals and to a decreased likelihood of rejecting the null hypothesis. Analysts often choose the 0.05 level of significance, which indicates a 5 percent chance of rejecting the null hypothesis when, in fact, it is true (a Type I error). Of course, decreasing the level of significance from 0.05 to 0.01 decreases the probability of Type I error, but it increases the probability of Type II error—failing to reject the null hypothesis when, in fact, it is false.

Often, financial analysts do not simply report whether or not their tests reject a particular hypothesis about a regression parameter. Instead, they report the p-value or probability value for a particular hypothesis. The p-value is the smallest level of significance at which the null hypothesis can be rejected. It allows the reader to interpret the results rather than be told that a certain hypothesis has been rejected or accepted. In most regression software packages, the p-values printed for regression coefficients apply to a test of null hypothesis that the true parameter is equal to 0 against the alternative that the parameter is not equal to 0, given the estimated coefficient and the standard error for that coefficient. For example, if the p-value is 0.005, we can reject the hypothesis that the true parameter is equal to 0 at the 0.5 percent significance level (99.5 percent confidence).

The standard error of the estimated coefficient is an important input for a hypothesis test concerning the regression coefficient (and for a confidence interval for the estimated coefficient). Stronger regression results lead to smaller standard errors of an estimated parameter and result in tighter confidence intervals. If the standard error ($s_{\hat{b}_1}$) in the above example were 0.100 instead of 0.200, the confidence interval range would be half as large and the t-statistic twice as large. With a standard error this small, we would reject the null hypothesis even at the 0.01 significance level because we would have t = (1.5 − 1)/0.1 = 5.00 and t_c = 2.66.

With this background, we can turn to hypothesis tests using actual regression results. The next three examples illustrate hypothesis tests in a variety of typical investment contexts.

EXAMPLE 4

Estimating Beta for Royal Bank of Canada Stock

Royal Bank of Canada (RBC) is one of Canada's largest banks. It provides financial products and services on a global basis. Its stock trades on the Toronto Stock Exchange. Suppose you are an investor in RBC's stock and want an estimate of its beta. As in the text example, you hypothesize that RBC has an average level of market risk and that its required return in excess of the risk-free rate is the same as the market's required excess return. One regression that summarizes these statements is

$$(R - R_F) = \alpha + \beta(R_M - R_F) + \varepsilon \quad (8)$$

where R_F is the periodic risk-free rate of return (known at the beginning of the period), R_M is the periodic return on the market, R is the periodic return to the stock of the company, and β measures the sensitivity of the required excess return to the excess return to market. Estimating this equation with linear regression provides an estimate of β, $\hat{\beta}$, which tells us the size of the required return premium for the security, given expectations about market returns.[19]

Suppose we want to test the null hypothesis, H_0, that β = 1 for RBC stock to see whether RBC stock has the same required return premium as the market as a whole. We need data on returns to RBC stock, a risk-free interest rate, and the returns to the market index. For this example, we use data from January 2013 through December 2017 (n = 60). The return to RBC stock is R. The monthly return to 1-month Canadian Treasury bills is R_F. The return to the S&P/TSX Composite Index is R_M.[20] This index is the primary broad measure of the Canadian equity market. We are estimating two parameters, so the number of degrees of freedom is $n - 2 = 60 - 2 = 58$. Table 5 shows the results from the regression $(R - R_F) = \alpha + \beta\,(R_M - R_F) + \varepsilon$.

Table 5 Estimating Beta for Royal Bank of Canada

Regression Statistics	
Multiple R	0.7131
R-squared	0.5086
Standard error of estimate	0.0269
Observations	60

(continued)

19 Beta (β) is typically estimated using 60 months of historical data, but the data-sample length sometimes varies. Although monthly data is typically used, some financial analysts estimate β using daily data. The expected excess return for RBC stock above the risk-free rate $(R - R_F)$ is $\beta(R_M - R_F)$, given a particular excess return to the market above the risk-free rate $(R_M - R_F)$. This result holds because we regress $(R - R_F)$ against $(R_M - R_F)$. For example, if a stock's beta is 1.5, its expected excess return is 1.5 times that of the market portfolio.

20 Data on RBC stock returns and S&P/TSX Composite Index returns came from ca.finance.yahoo.com. Data on Canadian T-bill returns came from the Bank of Canada.

Table 5 (Continued)

	Coefficients	Standard Error	*t*-Statistic
Alpha	0.0031	0.0070	0.4429
Beta	0.9068	0.1170	7.7504

Sources: Bank of Canada and ca.finance.yahoo.com.

1 Test the null hypothesis, H_0, that β for RBC equals 1 (β = 1) against the alternative hypothesis that β does not equal 1 (β ≠ 1) using the confidence interval approach.
2 Test the above hypothesis using a *t*-test.
3 How much of RBC stock's excess return variation can be attributed to company-specific risk?

Solution to 1:

The estimated $\hat{\beta}$ from the regression is 0.9068. The estimated standard error for that coefficient in the regression, $s_{\hat{\beta}}$ is 0.1170. The regression equation has 58 degrees of freedom (60 − 2), so the critical value for the test statistic is approximately t_c = 2.00 at the 0.05 significance level. Therefore, the 95 percent confidence interval for the data for any hypothesized value of β is shown by the range

$$\hat{\beta} \pm t_c s_{\hat{\beta}}$$

$$0.9068 \pm 2.00(0.1170)$$

$$0.6728 \text{ to } 1.1408$$

In this case, the hypothesized parameter value is β = 1, and the value 1 falls inside this confidence interval, so we cannot reject the hypothesis at the 0.05 significance level. This means that we cannot reject the hypothesis that RBC stock has the same systematic risk as the market as a whole.

Solution to 2:

The t-statistic for the test of whether the slope is equal to the average of the stocks in the market is computed using Equation:

$$t = \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} = \frac{0.9068 - 1.0}{0.1170} = -0.7966$$

The absolute value of this *t*-statistic is less than the critical *t*-value of 2.00. Therefore, neither approach allows us to reject the null hypothesis. Note that the *t*-statistic associated with $\hat{\beta}$ in the regression results in Table 5 is 7.7504. Given the significance level we are using, we cannot reject the null hypothesis that β = 1, but we can reject the hypothesis that β = 0.[21]

21 The *t*-statistics for a coefficient automatically reported by statistical software programs assume that the null hypothesis states that the coefficient is equal to 0. If you have a different null hypothesis, as we do in this example (β = 1), then you must either construct the correct test statistic yourself or instruct the program to compute it.

Solution to 3:

The R^2 in this regression is 0.5086. This result suggests that about 51 percent of the total variation in the excess return to RBC stock (the return to RBC above the risk-free rate) can be explained by excess return to the market portfolio. The remaining 49 percent of RBC stock's excess return variation is the nonsystematic component, which can be attributed to company-specific risk.

In the next example, we show a regression hypothesis test with a one-sided alternative.

EXAMPLE 5

Explaining Company Value Based on Returns to Invested Capital

Some financial analysts have argued that one good way to measure a company's ability to create wealth is to compare the company's return on invested capital (ROIC) to its weighted-average cost of capital (WACC). If a company has an ROIC greater than its cost of capital, the company is creating wealth; if its ROIC is less than its cost of capital, it is destroying wealth.[22]

Enterprise value (EV) is a market-price-based measure of company value defined as the market value of equity and debt minus the value of cash and investments. Invested capital (IC) is an accounting measure of company value defined as the sum of the book values of equity and debt. Higher ratios of EV to IC should reflect greater success at wealth creation in general. Mauboussin (1996) argued that the spread between ROIC and WACC helps explain the ratio of EV to IC. Using data on companies in the food-processing industry, we can test the relationship between EV/IC and (ROIC–WACC) using the regression model given in Equation 9.

$$\text{EV}_i/\text{IC}_i = b_0 + b_1(\text{ROIC}_i - \text{WACC}_i) + \varepsilon_i \quad (9)$$

where the subscript i is an index to identify the company. Our null hypothesis is H_0: $b_1 \le 0$, and we specify a significance level of 0.05. If we reject the null hypothesis, we have evidence of a statistically significant relationship between EV/IC and (ROIC–WACC). Equation 9 is estimated using data from nine food-processing companies.[23] The results of this regression are displayed in Table 6 and Figure 4.

Table 6 Explaining Enterprise Value/Invested Capital by the ROIC–WACC Spread

Regression Statistics	
Multiple R	0.9469
R-squared	0.8966
Standard error of estimate	0.7422
Observations	9

(continued)

22 See, for example, Sonkin and Johnson (2017).

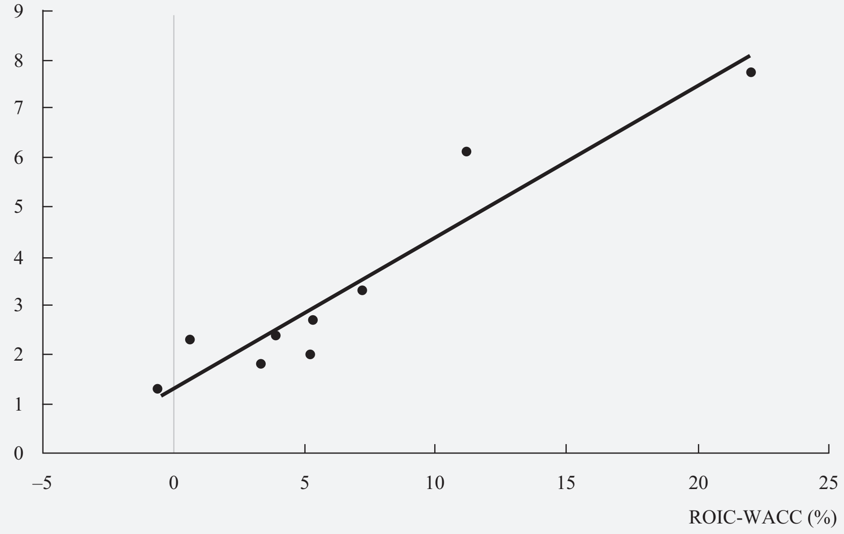
23 Our data come from Nelson, Moskow, Lee, and Valentine (2003) and relate to 2001. Many sell-side analysts use this type of regression. It is one of the most frequently used cross-sectional regressions in published analyst reports.

Table 6 (Continued)

	Coefficients	Standard Error	*t*-Statistic
Intercept	1.3478	0.3511	3.8391
Spread	30.0169	3.8519	7.7928

Source: Nelson, Moskow, Lee and Valentine (2003).

Figure 4 Fitted Regression Line Explaining Enterprise Value/Invested Capital Using ROIC–WACC Spread for the Food Industry



Source: Nelson et al. (2003).

We reject the null hypothesis based on the *t*-statistic of 7.7928 on estimated slope coefficient. There is a statistically significant positive relationship between the return spread (ROIC–WACC) and the ratio of EV to IC in our sample of companies. Figure 4 illustrates the strong positive relationship. The R^2 of 0.8966 indicates that the return spread explains about 90 percent of the variation in the ratio of EV to IC among the food-processing companies in the sample in 2001. The coefficient on the return spread of 30.0169 implies that the predicted increase in EV/IC is 0.01(30.0169) = 0.3002 or about 30 percent for a 1-percentage-point increase in the return spread, for our sample of companies.

In the final example of this section, the null hypothesis for a *t*-test of the slope coefficient is that the value of slope equals 1 in contrast to the null hypothesis that it equals 0 as in prior examples.

EXAMPLE 6

Testing whether Inflation Forecasts Are Unbiased

Example 1 introduced the concept of testing for bias in forecasts. That example showed that if a forecast is unbiased, its expected error is 0. We can examine whether a time-series of forecasts for a particular economic variable is unbiased by comparing the forecast at each date with the actual value of the economic variable announced after the forecast. If the forecasts are unbiased, then, by definition, the average realized forecast error should be close to 0. In that case, the value of b_0 (the intercept) should be 0 and the value of b_1 (the slope) should be 1, as discussed in Example 1.

Refer once again to Figure 2, which shows the mean forecast made by professional economic forecasters in the first quarter of a year for the percentage change in euro area HICP during that year and the actual percentage change from 1999 through 2017 (n = 19). To test whether the forecasts are unbiased, we must estimate the regression shown in Example 1. We report the results of this regression in Table 7. The equation to be estimated is

$$\text{Actual percentage change in HICP}_t = b_0 + b_1(\text{Predicted change}_t) + \varepsilon_t$$

This regression estimates two parameters (the intercept and the slope); therefore, the regression has $n - 2 = 19 - 2 = 17$ degrees of freedom.

Table 7 Testing whether Forecasts of Euro Area HICP Are Unbiased (Dependent Variable: CPI Change Expressed in Percent)

Regression Statistics	
Multiple R	0.9392
R-squared	0.8821
Standard error of estimate	0.3282
Observations	19

	Coefficients	Standard Error	t-Statistic
Intercept	–0.7679	0.2332	–3.2929
Forecast (slope)	1.5922	0.1412	11.2762

Source: European Central Bank.

We can now test two null hypotheses about the parameters in this regression. Our first null hypothesis is that the intercept in this regression is 0 (H_0: b_0 = 0). The alternative hypothesis is that the intercept does not equal 0 (H_a: $b_0 \neq 0$). Our second null hypothesis is that the slope coefficient in this regression is 1 (H_0: b_1 = 1). The alternative hypothesis is that the slope coefficient does not equal 1 (H_a: $b_1 \neq 1$).

To test the hypotheses about b_0 and b_1, we must first decide on a critical value based on a particular significance level and then construct the confidence intervals for each parameter. If we choose the 0.05 significance level, with 17 degrees of freedom, the critical value, t_c, is approximately 2.11. The estimated value of the parameter $\hat{b}_0$ is –0.7679, and the estimated value of the standard error for $\hat{b}_0\left(s_{\hat{b}_0}\right)$ is 0.2332. Let B_0 stand for any particular hypothesized value. Therefore, under the null hypothesis that $b_0 = B_0$, a 95 percent confidence interval for b_0 is $\hat{b}_0 \pm t_c s_{\hat{\beta}_0}$.

$-0.7679 \pm 2.11(0.2332)$

-1.2600 to -0.2758

In this case, B_0 is 0. The value of 0 falls outside this confidence interval, so we can reject the first null hypothesis that $b_0 = 0$. We will explain how to interpret this result shortly.

Our second null hypothesis is based on the same sample as our first null hypothesis. Therefore, the critical value for testing that hypothesis is the same as the critical value for testing the first hypothesis ($t_c = 2.11$). The estimated value of the parameter $\hat{b}_1$ is 1.5922, and the estimated value of the standard error for $\hat{b}_1$, $s_{\hat{b}_1}$, is 0.1412. Therefore, the 95 percent confidence interval for any particular hypothesized value of b_1 can be constructed as follows:

$\hat{b}_1 \pm t_c s_{\hat{b}_1}$

$1.5922 \pm 2.11(0.1412)$

1.2943 to 1.8901

In this case, our hypothesized value of b_1 is 1. The value 1 falls outside this confidence interval, so we can reject the null hypothesis that $b_1 = 1$ at the 0.05 significance level. Because we did reject at least one[24] of the two null hypotheses ($b_0 = 0$, $b_1 = 1$) about the parameters in this model, we can reject the hypothesis that the forecasts of HICP change were unbiased.[25]

As an analyst, you often will need forecasts of economic variables to help you make recommendations about asset allocation, expected returns, and other investment decisions. The hypothesis tests just conducted suggest that you can reject the hypothesis that the HICP predictions in the Survey of Professional Forecasters are unbiased. If you need an unbiased forecast of future percentage change in HICP for your asset-allocation decision, you might not want to use these forecasts. You may also want to do some further exploration. For example, a study of forecasts in the European Central Bank Survey of Professional Forecasters by Genre, Kenny, Meyler, and Timmerman (2013) finds that the performance of inflation forecasts is lowered when the financial crisis period is included. Therefore, you may want to re-estimate the regression equation after excluding the financial crisis period and retest the null hypotheses that $b_0 = 0$ and $b_1 = 1$. Similarly, you may want to look for evidence of any outliers in a scatter plot of realized and forecasted inflations, and redo the above analysis after excluding these outliers.[26]

24 In this example, we have rejected both the null hypotheses.

25 Jointly testing the hypothesis $b_0 = 0$ and $b_1 = 1$ would require us to take into account the covariance of $\hat{b}_0$ and $\hat{b}_1$. For information on testing joint hypotheses of this type, see Greene (2018).

26 Outliers are small numbers of observations at either extreme (small or large) of a sample. Correlation and linear regression may be unreliable measures when outliers are present in one or both of the series.

ANALYSIS OF VARIANCE IN A REGRESSION WITH ONE INDEPENDENT VARIABLE

7

Analysis of variance (ANOVA) is a statistical procedure for dividing the total variability of a variable into components that can be attributed to different sources.[27] In regression analysis, we use ANOVA to determine the usefulness of the independent variable or variables in explaining variation in the dependent variable. An important statistical test conducted in analysis of variance is the *F*-test. The *F*-statistic tests whether all the slope coefficients in a linear regression are equal to 0. In a regression with one independent variable, this is a test of the null hypothesis H_0: $b_1 = 0$ against the alternative hypothesis H_a: $b_1 \neq 0$.

To correctly determine the test statistic for the null hypothesis that the slope coefficient equals 0, we need to know the following:

- the total number of observations (n);
- the total number of parameters to be estimated (in a one-independent-variable regression, this number is two: the intercept and the slope coefficient);
- the sum of squared errors or residuals, $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$, abbreviated SSE. This value is also known as the residual sum of squares; and
- the regression sum of squares, $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$, abbreviated RSS. This value is the amount of total variation in Y that is explained in the regression equation. Total variation (TSS) is the sum of SSE and RSS.

The *F*-test for determining whether the slope coefficient equals 0 is based on an *F*-statistic, constructed using these four values. The *F*-statistic measures how well the regression equation explains the variation in the dependent variable. The *F*-statistic is the ratio of the average regression sum of squares to the average sum of the squared errors. The average regression sum of squares is computed by dividing the regression sum of squares by the number of slope parameters estimated (in this case, one). The average sum of squared errors is computed by dividing the sum of squared errors by the number of observations, n, minus the total number of parameters estimated (in this case, two: the intercept and the slope). These two divisors are the degrees of freedom for an *F*-test. If there are n observations, the *F*-test for the null hypothesis that the slope coefficient is equal to 0 is here denoted $F_{(\#\text{ slope parameters}),(n - \#\text{ parameters})} = F_{1,n-2}$, and the test has 1 and $n - 2$ degrees of freedom.

Suppose, for example, that the independent variable in a regression model explains none of the variation in the dependent variable. Then the predicted value for the regression model, $\hat{Y}_i$, is the average value of the dependent variable $\overline{Y}$. In this case, the regression sum of squares $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$ is 0. Therefore, the *F*-statistic is 0. If the independent variable explains little of the variation in the dependent variable, the value of the *F*-statistic will be very small.

27 In this reading, we focus on regression applications of ANOVA, the most common context in which financial analysts will encounter this tool. In this context, ANOVA is used to test whether all the regression slope coefficients are equal to 0. Analysts also use ANOVA to test a hypothesis that the means of two or more populations are equal. See Daniel and Terrell (1995) for details.

The formula for the F-statistic in a regression with one independent variable is

$$F = \frac{\text{RSS}/1}{\text{SSE}/(n-2)} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}} \quad (10)$$

If the regression model does a good job of explaining variation in the dependent variable, then this ratio should be high. The explained regression sum of squares per estimated parameter will be high relative to the unexplained variation for each degree of freedom. Critical values for this F-statistic are given in Appendix D at the end of this volume.

Even though the F-statistic is commonly computed by regression software packages, analysts typically do not use ANOVA and F-tests in regressions with just one independent variable. Why not? In such regressions, the F-statistic is the square of the t-statistic for the slope coefficient. Therefore, the F-test duplicates the t-test for the significance of the slope coefficient. This relation is not true for regressions with two or more slope coefficients. Nevertheless, the one-slope coefficient case gives a foundation for understanding the multiple-slope coefficient cases.

Often, mutual fund performance is evaluated based on whether the fund has positive alpha—significantly positive excess risk-adjusted returns.[28] One commonly used method of risk adjustment is based on the capital asset pricing model. Consider the regression

$$(R_i - R_F) = \alpha_i + \beta_i(R_M - R_F) + \varepsilon_i \quad (11)$$

where R_F is the periodic risk-free rate of return (known at the beginning of the period), R_M is the periodic return on the market, R_i is the periodic return to Mutual Fund i, and β_i is the fund's beta. A fund has zero risk-adjusted excess return if $\alpha_i = 0$. If $\alpha_i = 0$, then $(R_i - R_F) = \beta_i(R_M - R_F) + \varepsilon_i$ and taking expectations, $E(R_i) = R_F + \beta_i(R_M - R_F)$, implying that β_i completely explains the fund's mean excess returns. If, for example, $\alpha_i > 0$, the fund is earning higher returns than expected given its beta.

In summary, to test whether a fund has a positive alpha, we must test the null hypothesis that the fund has no risk-adjusted excess returns (H_0: $\alpha = 0$) against the alternative hypothesis of nonzero risk-adjusted returns (H_a: $\alpha \neq 0$).

EXAMPLE 7

Performance Evaluation: The Dreyfus Appreciation Fund

Table 8 presents results evaluating the excess return to the Dreyfus Appreciation Fund from January 2013 through December 2017. Note that the estimated beta in this regression, $\hat{\beta}_i$, is 0.9306. The Dreyfus Appreciation Fund was estimated to be less risky than the market as a whole.

Table 8 Performance Evaluation of Dreyfus Appreciation Fund, January 2013 to December 2017

Regression Statistics	
Multiple R	0.9699
R-squared	0.9407

28 Note that the Greek letter alpha, α, is traditionally used to represent the intercept in Equation 11 and should not be confused with another traditional usage of α to represent a significance level.

Table 8 (Continued)

Regression Statistics	
Standard error of estimate	0.0089
Observations	60

ANOVA	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	*F*
Regression	1	0.0729	0.0729	919.17
Residual	58	0.0046	0.0001	
Total	59	0.0997		

	Coefficients	Standard Error	*t*-Statistic
Alpha	–0.0035	0.0012	-2.9167
Beta	0.9306	0.0307	30.3127

Sources: finance.yahoo.com 21st June 2018; and the Federal Reserve.

1 Test whether the fund had a significant excess return beyond the return associated with the market risk of the fund.

2 Based on the *t*-test, discuss whether the beta of the fund is likely to be zero.

3 Use Equation 10 to compute the *F*-statistic. Based on the *F*-test, determine whether the beta of the fund is likely to be zero.

Solution to 1:

The estimated alpha ($\hat{\alpha}$) in this regression is negative (–0.0035). The absolute value of the coefficient is almost three times the size of the standard error for that coefficient (0.0012), so the *t*-statistic for the coefficient is –2.9167. Therefore, we can reject the null hypothesis ($\alpha = 0$) that the fund did not have a significant excess return beyond the return associated with the market risk of the fund. This result of a significant negative alpha means that the fund is earning lower returns than expected given its beta during this period.[29]

Solution to 2:

Because the *t*-statistic for the slope coefficient in this regression is 30.3127, the *p*-value for that coefficient is less than 0.0001 and is approximately zero. Therefore, the probability that the true value of this coefficient is 0 is microscopic.

29 This example introduces a well-known investment use of regression involving the capital asset pricing model. Researchers, however, recognize qualifications to the interpretation of alpha from a linear regression. The systematic risk of a managed portfolio is controlled by the portfolio manager. If, as a consequence, portfolio beta is correlated with the return on the market (as could result from market timing), inferences on alpha based on least-squares beta, as here, can be mistaken. This advanced subject is discussed in Dybvig and Ross (1985a) and (1985b).

Solution to 3:

The ANOVA portion of Table 8 provides the data we need to compute the *F*-statistic. In this case:

- the total number of observations (n) is 60;
- the total number of parameters to be estimated is 2 (intercept and slope);
- the sum of squared errors or residuals, SSE, is 0.0046; and
- the regression sum of squares, RSS, is 0.0729.

Therefore, the *F*-statistic to test whether the slope coefficient is equal to 0 is

$$\frac{0.0729/1}{0.0046/(60-2)} = 919.17$$

The ANOVA output would show that the p-value for this F-statistic is less than 0.0001 and is exactly the same as the p-value for the t-statistic for the slope coefficient. Therefore, the F-test tells us nothing more than we already knew from the t-test. Note also that the F-statistic (919.17) is the square of the t-statistic (30.3127). (The slight difference is due to rounding.)

8 PREDICTION INTERVALS

Financial analysts often want to use regression results to make predictions about a dependent variable. For example, we might ask, "How fast will the sales of XYZ Corporation grow this year if real GDP grows by 4 percent?" But we are not merely interested in making these forecasts; we also want to know how certain we should be about the forecasts' results. For example, if we predicted that sales for XYZ Corporation would grow by 6 percent this year, our prediction would mean more if we were 95 percent confident that sales growth would fall in the interval from 5 percent to 7 percent, rather than only 25 percent confident that this outcome would occur. Therefore, we need to understand how to compute confidence intervals around regression forecasts.

We must take into account two sources of uncertainty when using the regression model $Y_i = b_0 + b_1X_i + \varepsilon_i$, $i = 1, \ldots, n$ and the estimated parameters, $\hat{b}_0$ and $\hat{b}_1$, to make a prediction. First, the error term itself contains uncertainty. The standard deviation of the error term, σ_ε, can be estimated from the standard error of estimate for the regression equation. A second source of uncertainty in making predictions about Y, however, comes from uncertainty in the estimated parameters $\hat{b}_0$ and $\hat{b}_1$.

If we knew the true values of the regression parameters, b_0 and b_1, then the variance of our prediction of Y, given any particular predicted (or assumed) value of X, would simply be s^2, the squared standard error of estimate. The variance would be s^2 because the prediction, $\hat{Y}$, would come from the equation $\hat{Y} = b_0 + b_1X$ and $\left(Y - \hat{Y}\right) = \varepsilon$.

Because we must estimate the regression parameters $\hat{b}_0$ and $\hat{b}_1$ however, our prediction of Y, $\hat{Y}$, given any particular predicted value of X, is actually $\hat{Y} = \hat{b}_0 + \hat{b}_1X$. The estimated variance of the prediction error, s_f^2 of Y, given X, is

$$s_f^2 = s^2\left[1 + \frac{1}{n} + \frac{\left(X - \bar{X}\right)^2}{(n-1)s_x^2}\right] \tag{12}$$

This estimated variance depends on:

- the squared standard error of estimate, s^2;
- the number of observations, n;
- the value of the independent variable, X, used to predict the dependent variable;
- the estimated mean, $\bar{X}$; and
- variance, s_x^2 of the independent variable.

Once we have this estimate of the variance of the prediction error, determining a prediction interval around the prediction is very similar to estimating a confidence interval around an estimated parameter, as shown earlier in this reading. We need to take the following four steps to determine the prediction interval for the prediction:

1. Make the prediction.
2. Compute the variance of the prediction error using Equation 12.
3. Choose a significance level, α, for the forecast. For example, the 0.05 level, given the degrees of freedom in the regression, determines the critical value for the forecast interval, t_c.
4. Compute the (1 − α) percent prediction interval for the prediction, namely $\hat{Y} \pm t_c s_f$.

EXAMPLE 8

Predicting the Ratio of Enterprise Value to Invested Capital

We continue with the example of explaining the ratio of enterprise value to invested capital among food-processing companies by the spread between the return to invested capital and the weighted-average cost of capital (ROIC–WACC). In Example 5, we estimated the regression given in Table 6.

Table 6 Explaining Enterprise Value/Invested Capital by the ROIC–WACC Spread (repeated)

Regression Statistics	
Multiple R	0.9469
R-squared	0.8966
Standard error of estimate	0.7422
Observations	9

	Coefficients	Standard Error	t-Statistic
Intercept	1.3478	0.3511	3.8391
Spread	30.0169	3.8519	7.7928

Source: Nelson, Moskow, Lee, and Valentine (2003).

You are interested in predicting the ratio of enterprise value to invested capital for a company if the return spread between ROIC and WACC is 10 percentage points. What is the 95 percent prediction interval for the ratio of enterprise value to invested capital for that company?

Using the data provided in Table 6, take the following steps:

1 Make the prediction: Expected EV/IC = 1.3478 + 30.0169(0.10) = 4.3495. This regression suggests that if the return spread between ROIC and WACC (X_i) is 10 percent, the predicted EV/IC ratio will be 4.3495.

2 Compute the variance of the prediction error. To compute the variance of the forecast error, we must know:

- the standard error of the estimate of the equation, s = 0.7422 (as shown in Table 6);
- the mean return spread, $\overline{X}$ = 0.0647 (this computation is not shown in the table); and
- the variance of the mean return spread in the sample, s_x^2 = 0.004641 (this computation is not shown in the table).

Using these data, you can compute the variance of the forecast error (s_f^2) for predicting EV/IC for a company with a 10 percent spread between ROIC and WACC.

$$s_f^2 = 0.7422^2\left[1 + \frac{1}{9} + \frac{(0.10 - 0.0647)^2}{(9 - 1)0.004641}\right]$$
$$= 0.630556$$

In this example, the variance of the forecast error is 0.630556, and the standard deviation of the forecast error is $s_f = (0.630556)^{1/2} = 0.7941$.

3 Determine the critical value of the t-statistic. Given a 95 percent confidence interval and 9 − 2 = 7 degrees of freedom, the critical value of the t-statistic, t_c, is 2.365 using the tables in the back of this volume.

4 Compute the prediction interval. The 95 percent confidence interval for EV/IC extends from 4.3495 − 2.365(0.7941) to 4.3495 + 2.365(0.7941), or 2.4715 to 6.2275.

In summary, if the spread between the ROIC and the WACC is 10 percent, the 95 percent prediction interval for EV/IC will extend from 2.4715 to 6.2275. The small sample size is reflected in the relatively large prediction interval.

9 LIMITATIONS OF REGRESSION ANALYSIS

Although this reading has shown many of the uses of regression models for financial analysis, regression models do have limitations. First, regression relations can change over time, just as correlations can. This fact is known as the issue of **parameter instability**, and its existence should not be surprising as the economic, tax, regulatory, political, and institutional contexts in which financial markets operate change. Whether considering cross-sectional or time-series regression, the analyst will probably face this issue. As one example, cross-sectional regression relationships between stock characteristics may differ between growth-led and value-led markets. As a second example, the time-series regression estimating the beta often yields significantly different estimated betas depending on the time period selected. In both cross-sectional and time-series contexts, the most common problem is sampling from more than one population, with the challenge of identifying when doing so is an issue.

A second limitation to the use of regression results specific to investment contexts is that public knowledge of regression relationships may negate their future usefulness. Suppose, for example, an analyst discovers that stocks with a certain characteristic have had historically very high returns. If other analysts discover and act upon this relationship, then the prices of stocks with that characteristic will be bid up. The knowledge of the relationship may result in the relation no longer holding in the future.

Finally, if the regression assumptions listed in Section 2.2 are violated, hypothesis tests and predictions based on linear regression will not be valid. Although there are tests for violations of regression assumptions, often uncertainty exists as to whether an assumption has been violated. This limitation will be discussed in detail in the reading on multiple regression.

SUMMARY

- The dependent variable in a linear regression is the variable that the regression model tries to explain. The independent variables are the variables that a regression model uses to explain the dependent variable.
- If there is one independent variable in a linear regression and there are n observations on the dependent and independent variables, the regression model is $Y_i = b_0 + b_1X_i + \varepsilon_i$, $i = 1, \ldots, n$, where Y_i is the dependent variable, X_i is the independent variable, and ε_i is the error term. In this model, the coefficient b_0 is the intercept. The intercept is the predicted value of the dependent variable when the independent variable has a value of zero. In this model, the coefficient b_1 is the slope of the regression line. If the value of the independent variable increases by one unit, then the model predicts that the value of the dependent variable will increase by b_1 units.
- The assumptions of the classic normal linear regression model are the following:
  - A linear relation exists between the dependent variable and the independent variable.
  - The independent variable is not random.
  - The expected value of the error term is 0.
  - The variance of the error term is the same for all observations (homoskedasticity).
  - The error term is uncorrelated across observations.
  - The error term is normally distributed.
- The estimated parameters in a linear regression model minimize the sum of the squared regression residuals.
- The standard error of estimate measures how well the regression model fits the data. If the SEE is small, the model fits well.
- The coefficient of determination measures the fraction of the total variation in the dependent variable that is explained by the independent variable. In a linear regression with one independent variable, the simplest way to compute the coefficient of determination is to square the correlation of the dependent and independent variables.
- To calculate a confidence interval for an estimated regression coefficient, we must know the standard error of the estimated coefficient and the critical value for the t-distribution at the chosen level of significance, t_c.

- To test whether the population value of a regression coefficient, b_1, is equal to a particular hypothesized value, B_1, we must know the estimated coefficient, $\hat{b}_1$, the standard error of the estimated coefficient, $s_{\hat{b}_1}$, and the critical value for the t-distribution at the chosen level of significance, t_c. The test statistic for this hypothesis is $\left(\hat{b}_1 - B_1\right)/s_{\hat{b}_1}$. If the absolute value of this statistic is greater than t_c, then we reject the null hypothesis that $b_1 = B_1$.
- In the regression model $Y_i = b_0 + b_1X_i + \varepsilon_i$, if we know the estimated parameters, $\hat{b}_0$ and $\hat{b}_1$, for any value of the independent variable, X, then the predicted value of the dependent variable Y is $\hat{Y} = \hat{b}_0 + \hat{b}_1X$.
- The prediction interval for a regression equation for a particular predicted value of the dependent variable is $\hat{Y} \pm t_c s_f$ where s_f is the square root of the estimated variance of the prediction error and t_c is the critical level for the t-statistic at the chosen significance level. This computation specifies a (1 − α) percent confidence interval. For example, if α = 0.05, then this computation yields a 95 percent confidence interval.

REFERENCES

Buetow, Gerald W., Jr, Robert R. Johnson, and David E. Runkle. 2000. “The Inconsistency of Returns-Based Style Analysis.” *Journal of Portfolio Management* 26 (3): 61–77.

Campbell, John Y., Karine Serfaty-de Medeiros, and Luis M. Viceira. 2010. “Global Currency Hedging.” *Journal of Finance* 65 (1): 87–121.

Chan, Louis K. C., Stephen G. Dimmock, and Josef Lakonishok. 2009. “Benchmarking Money Manager Performance: Issues and Evidence.” *Review of Financial Studies* 22 (11): 4553–99.

Daniel, Wayne W., and James C. Terrell. 1995. *Business Statistics for Management and Economics*. 7th ed. Boston: Houghton-Mifflin.

Dybvig, Philip H., and Stephen A. Ross. 1985a. “Differential Information and Performance Measurement Using a Security Market Line.” *Journal of Finance* 40 (2): 383–99.

Dybvig, Philip H., and Stephen A. Ross. 1985b. “The Analytics of Performance Measurement Using a Security Market Line.” *Journal of Finance* 40 (2): 401–16.

Genre, Veronique, Geoff Kenny, Aidan Meyler, and Allan Timmermann. 2013. “Combining expert forecasts: Can anything beat the simple average?” *International Journal of Forecasting* 29 (1): 108–21.

Greene, William H. 2018. Economic Analysis. 8th ed. Upper Saddle River, NJ: Prentice-Hall.

Keane, Michael P., and David E. Runkle. 1990. “Testing the Rationality of Price Forecasts: New Evidence from Panel Data.” *American Economic Review* 80 (4): 714–35.

Nelson, David C., Robert B. Moskow, Tiffany Lee, and Gregg Valentine. 2003. Food Investor’s Handbook. New York: Credit Suisse First Boston.

Sharpe, William F. 1992. “Asset Allocation: Management Style and Performance Measurement.” *Journal of Portfolio Management* 18 (2): 7–19.

Sonkin, Paul D., and Paul Johnson. 2017. Pitch the Perfect Investment. New York: Wiley.

PRACTICE PROBLEMS

1 Julie Moon is an energy analyst examining electricity, oil, and natural gas consumption in different regions over different seasons. She ran a regression explaining the variation in energy consumption as a function of temperature. The total variation of the dependent variable was 140.58, the explained variation was 60.16, and the unexplained variation was 80.42. She had 60 monthly observations.

A Compute the coefficient of determination.

B What was the sample correlation between energy consumption and temperature?

C Compute the standard error of the estimate of Moon's regression model.

D Compute the sample standard deviation of monthly energy consumption.

2 You are examining the results of a regression estimation that attempts to explain the unit sales growth of a business you are researching. The analysis of variance output for the regression is given in the table below. The regression was based on five observations ($n = 5$).

ANOVA	df	SS	MSS	F	Significance F
Regression	1	88.0	88.0	36.667	0.00904
Residual	3	7.2	2.4		
Total	4	95.2			

A How many independent variables are in the regression to which the ANOVA refers?

B Define Total SS.

C Calculate the sample variance of the dependent variable using information in the above table.

D Define Regression SS and explain how its value of 88 is obtained in terms of other quantities reported in the above table.

E What hypothesis does the F-statistic test?

F Explain how the value of the F-statistic of 36.667 is obtained in terms of other quantities reported in the above table.

G Is the F-test significant at the 5 percent significance level?

3 An economist collected the monthly returns for KDL's portfolio and a diversified stock index. The data collected are shown below:

Month	Portfolio Return (%)	Index Return (%)
1	1.11	–0.59
2	72.10	64.90
3	5.12	4.81
4	1.01	1.68
5	–1.72	–4.97
6	4.06	–2.06

The economist calculated the correlation between the two returns and found it to be 0.996. The regression results with the KDL return as the dependent variable and the index return as the independent variable are given as follows:

Regression Statistics	
Multiple R	0.996
R-squared	0.992
Standard error	2.861
Observations	6

ANOVA	**df**	**SS**	**MSS**	***F***	**Significance *F***
Regression	1	4101.62	4101.62	500.79	0
Residual	4	32.76	8.19		
Total	5	4134.38			

	Coefficients	**Standard Error**	***t*-Statistic**	***p*-Value**
Intercept	2.252	1.274	1.768	0.1518
Slope	1.069	0.0477	22.379	0

When reviewing the results, Andrea Fusilier suspected that they were unreliable. She found that the returns for Month 2 should have been 7.21 percent and 6.49 percent, instead of the large values shown in the first table. Correcting these values resulted in a revised correlation of 0.824 and the revised regression results shown as follows:

Regression Statistics	
Multiple R	0.824
R-squared	0.678
Standard error	2.062
Observations	6

ANOVA	**df**	**SS**	**MSS**	***F***	**Significance *F***
Regression	1	35.89	35.89	8.44	0.044
Residual	4	17.01	4.25		
Total	5	52.91			

	Coefficients	**Standard Error**	***t*-Statistic**	***p*-Value**
Intercept	2.242	0.863	2.597	0.060
Slope	0.623	0.214	2.905	0.044

Explain how the bad data affected the results.

The following information relates to Questions 4–9

Kenneth McCoin, CFA, is a fairly tough interviewer. Last year, he handed each job applicant a sheet of paper with the information in the following table, and he then asked several questions about regression analysis. Some of McCoin's questions, along with a sample of the answers he received to each, are given below. McCoin told the applicants that the independent variable is the ratio of net income to sales for restaurants with a market cap of more than $100 million and the dependent variable is the ratio of cash flow from operations to sales for those restaurants. Which of the choices provided is the best answer to each of McCoin's questions?

Regression Statistics	
Multiple *R*	0.8623
R-squared	0.7436
Standard error	0.0213
Observations	24

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	0.029	0.029000	63.81	0
Residual	22	0.010	0.000455		
Total	23	0.040			

	Coefficients	Standard Error	*t*-Statistic	*p*-Value
Intercept	0.077	0.007	11.328	0
Slope	0.826	0.103	7.988	0

4 What is the value of the coefficient of determination?

A 0.8261.

B 0.7436.

C 0.8623.

5 Suppose that you deleted several of the observations that had small residual values. If you re-estimated the regression equation using this reduced sample, what would likely happen to the standard error of the estimate and the *R*-squared?

	Standard Error of the Estimate	*R*-Squared
A	Decrease	Decrease
B	Decrease	Increase
C	Increase	Decrease

6 What is the correlation between *X* and *Y*?

A −0.7436.

B 0.7436.

C 0.8623.

7 Where did the *F*-value in the ANOVA table come from?

A You look up the *F*-value in a table. The *F* depends on the numerator and denominator degrees of freedom.

B Divide the "Mean Square" for the regression by the "Mean Square" of the residuals.

C The F-value is equal to the reciprocal of the t-value for the slope coefficient.

8 If the ratio of net income to sales for a restaurant is 5 percent, what is the predicted ratio of cash flow from operations to sales?

A 0.007 + 0.103(5.0) = 0.524.

B 0.077 − 0.826(5.0) = −4.054.

C 0.077 + 0.826(5.0) = 4.207.

9 Is the relationship between the ratio of cash flow to operations and the ratio of net income to sales significant at the 5 percent level?

A No, because the R-squared is greater than 0.05.

B No, because the p-values of the intercept and slope are less than 0.05.

C Yes, because the p-values for F and t for the slope coefficient are less than 0.05.

The following information relates to Questions 10–14

Howard Golub, CFA, is preparing to write a research report on Stellar Energy Corp. common stock. One of the world's largest companies, Stellar is in the business of refining and marketing oil. As part of his analysis, Golub wants to evaluate the sensitivity of the stock's returns to various economic factors. For example, a client recently asked Golub whether the price of Stellar Energy Corporation stock has tended to rise following increases in retail energy prices. Golub believes the association between the two variables to be negative, but he does not know the strength of the association.

Golub directs his assistant, Jill Batten, to study the relationships between Stellar monthly common stock returns versus the previous month's percent change in the US Consumer Price Index for Energy (CPIENG), and Stellar monthly common stock returns versus the previous month's percent change in the US Producer Price Index for Crude Energy Materials (PPICEM). Golub wants Batten to run both a correlation and a linear regression analysis. In response, Batten compiles the summary statistics shown in Exhibit 1 for the 248 months between January 1980 and August 2000. All of the data are in decimal form, where 0.01 indicates a 1 percent return. Batten also runs a regression analysis using Stellar monthly returns as the dependent variable and the monthly change in CPIENG as the independent variable. Exhibit 2 displays the results of this regression model.

Exhibit 1 Descriptive Statistics

	Monthly Return Stellar Common Stock	Lagged Monthly Change	
		CPIENG	PPICEM
Mean	0.0123	0.0023	0.0042
Standard Deviation	0.0717	0.0160	0.0534
Covariance, Stellar vs. CPIENG	−0.00017		

Exhibit 1 (Continued)

	Monthly Return Stellar Common Stock	Lagged Monthly Change CPIENG	Lagged Monthly Change PPICEM
Covariance, Stellar vs. PPICEM	−0.00048		
Covariance, CPIENG vs. PPICEM	0.00044		
Correlation, Stellar vs. CPIENG	−0.1452		

Exhibit 2 Regression Analysis with CPIENG

Regression Statistics	
Multiple R	0.1452
R-squared	0.0211
Standard error of the estimate	0.0710
Observations	248

	Coefficients	Standard Error	t-Statistic
Intercept	0.0138	0.0046	3.0275
Slope coefficient	−0.6486	0.2818	−2.3014

10 Did Batten's regression analyze cross-sectional or time-series data, and what was the expected value of the error term from that regression?

	Data Type	Expected Value of Error Term
A	Time-series	0
B	Time-series	ε_i
C	Cross-sectional	0

11 Based on the regression, which used data in decimal form, if the CPIENG *decreases* by 1.0 percent, what is the expected return on Stellar common stock during the next period?

A 0.0073 (0.73 percent).

B 0.0138 (1.38 percent).

C 0.0203 (2.03 percent).

12 Based on Batten's regression model, the coefficient of determination indicates that:

A Stellar's returns explain 2.11 percent of the variability in CPIENG.

B Stellar's returns explain 14.52 percent of the variability in CPIENG.

C Changes in CPIENG explain 2.11 percent of the variability in Stellar's returns.

13 For Batten's regression model, the standard error of the estimate shows that the standard deviation of:

A the residuals from the regression is 0.0710.

B values estimated from the regression is 0.0710.

C Stellar's observed common stock returns is 0.0710.

14 For the analysis run by Batten, which of the following is an *incorrect* conclusion from the regression output?

A The estimated intercept coefficient from Batten's regression is statistically significant at the 0.05 level.

B In the month after the CPIENG declines, Stellar's common stock is expected to exhibit a positive return.

C Viewed in combination, the slope and intercept coefficients from Batten's regression are not statistically significant at the 0.05 level.

The following information relates to Questions 15–24

Anh Liu is an analyst researching whether a company's debt burden affects investors' decision to short the company's stock. She calculates the short interest ratio (the ratio of short interest to average daily share volume, expressed in days) for 50 companies as of the end of 2016 and compares this ratio with the companies' debt ratio (the ratio of total liabilities to total assets, expressed in decimal form).

Liu provides a number of statistics in Exhibit 1. She also estimates a simple regression to investigate the effect of the debt ratio on a company's short interest ratio. The results of this simple regression, including the analysis of variance (ANOVA), are shown in Exhibit 2.

In addition to estimating a regression equation, Liu graphs the 50 observations using a scatterplot, with the short interest ratio on the vertical axis and the debt ratio on the horizontal axis.

Exhibit 1 Summary Statistics

Statistic	Debt Ratio X_i	Short Interest Ratio Y_i
Sum	19.8550	192.3000
Average	0.3971	3.8460
Sum of squared deviations from the mean	$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 2.2225$	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 412.2042$
Sum of cross-products of deviations from the mean	$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = -9.2430$	

Exhibit 2 Regression of the Short Interest Ratio on the Debt Ratio

ANOVA	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)
Regression	1	38.4404	38.4404
Residual	48	373.7638	7.7867
Total	49	412.2042	

Regression Statistics	
Multiple R	0.3054
R^2	0.0933
Standard error of estimate	2.7905
Observations	50

	Coefficients	Standard Error	*t*-Statistic
Intercept	5.4975	0.8416	6.5322
Debt ratio	−4.1589	1.8718	−2.2219

Liu is considering three interpretations of these results for her report on the relationship between debt ratios and short interest ratios:

Interpretation 1	Companies' higher debt ratios cause lower short interest ratios.
Interpretation 2	Companies' higher short interest ratios cause higher debt ratios.
Interpretation 3	Companies with higher debt ratios tend to have lower short interest ratios.

She is especially interested in using her estimation results to predict the short interest ratio for MQD Corporation, which has a debt ratio of 0.40.

15 Based on Exhibits 1 and 2, if Liu were to graph the 50 observations, the scatterplot summarizing this relation would be *best* described as:

A horizontal.

B upward sloping.

C downward sloping.

16 Based on Exhibit 1, the sample covariance is *closest to*:

A −9.2430.

B −0.1886.

C 8.4123.

17 Based on Exhibit 1, the correlation between the debt ratio and the short interest ratio is *closest to*:

A −0.3054.

B 0.0933.

C 0.3054.

18 Which of the interpretations *best* describes Liu's findings for her report?

A Interpretation 1

B Interpretation 2

C Interpretation 3

19 The dependent variable in Liu's regression analysis is the:

A intercept.

B debt ratio.

C short interest ratio.

20 Based on Exhibit 2, the degrees of freedom for the *t*-test of the slope coefficient in this regression are:

A 48.

B 49.

C 50.

21 The upper bound for the 95% confidence interval for the coefficient on the debt ratio in the regression is *closest* to:

A −1.0199.

B −0.3947.

C 1.4528.

22 Which of the following should Liu conclude from these results shown in Exhibit 2?

A The average short interest ratio is 5.4975.

B The estimated slope coefficient is statistically significant at the 0.05 level.

C The debt ratio explains 30.54% of the variation in the short interest ratio.

23 Based on Exhibit 2, the short interest ratio expected for MQD Corporation is *closest to*:

A 3.8339.

B 5.4975.

C 6.2462.

24 Based on Liu's regression results in Exhibit 2, the *F*-statistic for testing whether the slope coefficient is equal to zero is *closest* to:

A −2.2219.

B 3.5036.

C 4.9367.

The following information relates to Questions 25–30

Elena Vasileva recently joined EnergyInvest as a junior portfolio analyst. Vasileva's supervisor asks her to evaluate a potential investment opportunity in Amtex, a multinational oil and gas corporation based in the US. Vasileva's supervisor suggests using regression analysis to examine the relation between Amtex shares and returns on crude oil.

Vasileva notes the following assumptions of regression analysis:

Assumption 1 The error term is uncorrelated across observations.

Assumption 2 The variance of the error term is the same for all observations.

Assumption 3 The expected value of the error term is equal to the mean value of the dependent variable.

Vasileva runs a regression of Amtex share returns on crude oil returns using the monthly data she collected. Selected data used in the regression are presented in Exhibit 1, and selected regression output is presented in Exhibit 2.

Exhibit 1 Selected Data for Crude Oil Returns and Amtex Share Returns

	Oil Return (X_i)	Amtex Return (Y_i)	Cross-Product $(X_i - \bar{X})(Y_i - \bar{Y})$	Predicted Amtex Return $(\hat{Y})$	Regression Residual $(Y_i - \hat{Y})$	Squared Residual $(Y_i - \hat{Y})^2$
Month 1	–0.032000	0.033145	–0.000388	0.002011	–0.031134	0.000969
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Month 36	0.028636	0.062334	0.002663	0.016282	–0.046053	0.002121
Sum			0.085598			0.071475
Average	–0.018056	0.005293				

Exhibit 2 Selected Regression Output Dependent Variable: Amtex Share Return

	Coefficient	Standard Error
Intercept	0.0095	0.0078
Oil return	0.2354	0.0760

Note: The critical *t*-value for a one-sided *t*-test at the 5% significance level is 1.691.

Vasileva expects the crude oil return next month, Month 37, to be –0.01. She computes the variance of the prediction error to be 0.0022.

25 Which of Vasileva's assumptions regarding regression analysis is *incorrect*?

A Assumption 1

B Assumption 2

C Assumption 3

26 Based on Exhibit 1, the standard error of the estimate is *closest* to:

A 0.044558.

B 0.045850.

C 0.050176.

27 Based on Exhibit 2, Vasileva should reject the null hypothesis that:

A the slope is less than or equal to 0.15.

B the intercept is less than or equal to 0.

C crude oil returns do not explain Amtex share returns.

28 Based on Exhibit 2, Vasileva should compute the:

A coefficient of determination to be 0.4689.

B 95% confidence interval for the intercept to be –0.0037 to 0.0227.

C 95% confidence interval for the slope coefficient to be 0.0810 to 0.3898.

29 Based on Exhibit 2 and Vasileva's prediction of the crude oil return for month 37, the estimate of Amtex share return for month 37 is *closest* to:

A –0.0024.

B 0.0071.

C 0.0119.

30 Using information from Exhibit 2, Vasileva should compute the 95% prediction interval for Amtex share return for month 37 to be:

A –0.0882 to 0.1025.

B –0.0835 to 0.1072.

C 0.0027 to 0.0116.

The following information relates to Question 31–33

Doug Abitbol is a portfolio manager for Polyi Investments, a hedge fund that trades in the United States. Abitbol manages the hedge fund with the help of Robert Olabudo, a junior portfolio manager.

Abitbol looks at economists' inflation forecasts and would like to examine the relationship between the US Consumer Price Index (US CPI) consensus forecast and actual US CPI using regression analysis. Olabudo estimates regression coefficients to test whether the consensus forecast is unbiased. Regression results are presented in Exhibit 1. Additionally, Olabudo calculates the 95% prediction interval of the actual CPI using a US CPI consensus forecast of 2.8.

Exhibit 1 Regression Output: Estimating US CPI

Regression Statistics	
Multiple *R*	0.9929
R-squared	0.9859
Standard error of estimate	0.0009
Observations	60

Exhibit 1 (Continued)

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.0001	0.0002	0.5351
US CPI consensus forecast	0.9830	0.0155	63.6239

Notes:

1 The absolute value of the critical value for the *t*-statistic is 2.0 at the 5% level of significance.

2 The standard deviation of the US CPI consensus forecast is s_x = 0.7539.

3 The mean of US CPI consensus forecast is $\overline{X}$ = 1.3350.

To conclude their meeting, Abitbol and Olabudo discuss the limitations of regression analysis. Olabudo notes the following limitations of regression analysis:

Limitation 1: Public knowledge of regression relationships may negate their future usefulness.

Limitation 2: Hypothesis tests and predictions based on linear regression will not be valid if regression assumptions are violated.

31 Based on Exhibit 1, Olabudo should:

A conclude that the inflation predictions are unbiased.

B reject the null hypothesis that the slope coefficient equals 1.

C reject the null hypothesis that the intercept coefficient equals 0.

32 Based on Exhibit 1, Olabudo should calculate a prediction interval for the actual US CPI *closest* to:

A 2.7506 to 2.7544.

B 2.7521 to 2.7529.

C 2.7981 to 2.8019.

33 Which of Olabudo's noted limitations of regression analysis is correct?

A Only Limitation 1

B Only Limitation 2

C Both Limitation 1 and Limitation 2

SOLUTIONS

1 **A** The coefficient of determination is

$$\frac{\text{Explained variation}}{\text{Total variation}} = \frac{60.16}{140.58} = 0.4279$$

B For a linear regression with one independent variable, the absolute value of correlation between the independent variable and the dependent variable equals the square root of the coefficient of determination, so the correlation is $\sqrt{0.4279} = 0.6542$. (The correlation will have the same sign as the slope coefficient.)

C The standard error of the estimate is

$$\left(\sum_{i=1}^{n}\frac{\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2}{n-2}\right)^{1/2} = \left(\frac{\text{Unexplained variation}}{n-2}\right)^{1/2}$$

$$= \sqrt{\frac{80.42}{60-2}} = 1.178$$

D The sample variance of the dependent variable is

$$\sum_{i=1}^{n}\frac{\left(Y_i - \overline{Y}\right)^2}{n-1} = \frac{\text{Total variation}}{n-1} = \frac{140.58}{60-1} = 2.3827$$

The sample standard deviation is $\sqrt{2.3827} = 1.544$.

2 **A** The degrees of freedom for the regression is the number of slope parameters in the regression, which is the same as the number of independent variables in the regression. Because regression df = 1, we conclude that there is one independent variable in the regression.

B Total SS is the sum of the squared deviations of the dependent variable Y about its mean.

C The sample variance of the dependent variable is the total SS divided by its degrees of freedom ($n - 1 = 5 - 1 = 4$ as given). Thus the sample variance of the dependent variable is 95.2/4 = 23.8.

D The Regression SS is the part of total sum of squares explained by the regression. Regression SS equals the sum of the squared differences between predicted values of the Y and the sample mean of Y: $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$. In terms of other values in the table, Regression SS is equal to Total SS minus Residual SS: 95.2 − 7.2 = 88.

E The F-statistic tests whether all the slope coefficients in a linear regression are equal to 0.

F The calculated value of F in the table is equal to the Regression MSS divided by the Residual MSS: 88/2.4 = 36.667.

G Yes. The significance of 0.00904 given in the table is the p-value of the test (the smallest level at which we can reject the null hypothesis). This value of 0.00904 is less than the specified significance level of 0.05, so we reject the null hypothesis. The regression equation has significant explanatory power.

3 The Month 2 data point is an outlier, lying far away from the other data values. Because this outlier was caused by a data entry error, correcting the outlier improves the validity and reliability of the regression. In this case, the true correlation is reduced from 0.996 to 0.824. The revised *R*-squared is substantially lower (0.678 versus 0.992). The significance of the regression is also lower, as can be seen in the decline of the *F*-value from 500.79 to 8.44 and the decline in the *t*-statistic of the slope coefficient from 22.379 to 2.905.

The total sum of squares and regression sum of squares were greatly exaggerated in the incorrect analysis. With the correction, the slope coefficient changes from 1.069 to 0.623. This change is important. When the index moves up or down, the original model indicates that the portfolio return goes up or down by 1.069 times as much, while the revised model indicates that the portfolio return goes up or down by only 0.623 times as much. In this example, incorrect data entry caused the outlier. Had it been a valid observation, not caused by a data error, then the analyst would have had to decide whether the results were more reliable including or excluding the outlier.

4 B is correct. The coefficient of determination is the same as *R*-squared.

5 C is correct. Deleting observations with small residuals will degrade the strength of the regression, resulting in an *increase* in the standard error and a *decrease* in *R*-squared.

6 C is correct. For a regression with one independent variable, the correlation is the same as the Multiple *R* with the sign of the slope coefficient. Because the slope coefficient is positive, the correlation is 0.8623.

7 B is correct. This answer describes the calculation of the *F*-statistic.

8 C is correct. To make a prediction using the regression model, multiply the slope coefficient by the forecast of the independent variable and add the result to the intercept.

9 C is correct. The *p*-value is the smallest level of significance at which the null hypotheses concerning the slope coefficient can be rejected. In this case the *p*-value is less than 0.05, and thus the regression of the ratio of cash flow from operations to sales on the ratio of net income to sales is significant at the 5 percent level.

10 A is correct because the data are time series, and the expected value of the error term, $E(\varepsilon)$, is 0.

11 C is correct. From the regression equation, Expected return = 0.0138 + −0.6486(−0.01) = 0.0138 + 0.006486 = 0.0203, or 2.03 percent.

12 C is correct. *R*-squared is the coefficient of determination. In this case, it shows that 2.11 percent of the variability in Stellar's returns is explained by changes in CPIENG.

153 A is correct, because the standard error of the estimate is the standard deviation of the regression residuals.

14 C is the correct response, because it is a false statement. The slope and intercept are both statistically significant.

15 C is correct because the slope coefficient (Exhibit 2) and the cross-product (Exhibit 1) are negative.

16 B is correct. The sample covariance is calculated as

$$\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1} = -9.2430 \div 49 = -0.1886$$

17 A is correct. For a regression with one independent variable, the correlation is the same as the Multiple R with the sign of the slope coefficient. Because the slope coefficient is negative, the correlation is −0.3054.

18 C is correct. Conclusions cannot be drawn regarding causation, only about association.

19 C is correct. Liu explains the short interest ratio using the debt ratio.

20 A is correct. The degrees of freedom are the number of observations minus the number of parameters estimated, which equals two in this case (the intercept and the slope coefficient). The number of degrees of freedom is 50 − 2 = 48.

21 B is correct. The calculation for the confidence interval is −4.1589 ± (2.011 × 1.8718). The upper bound is −0.3947. The 2.011 is the critical t-value for the 5% level of significance (2.5% in one tail) for 48 degrees of freedom.

22 B is correct. The t-statistic is −2.2219, which is outside of the bounds created by the critical t-values of ± 2.011 for a two-tailed test with a 5% significance level. The 2.011 is the critical t-value for the 5% level of significance (2.5% in one tail) for 48 degrees of freedom.

23 A is correct because Predicted value = 5.4975 + (−4.1589 × 0.40) = 5.4975 − 1.6636 = 3.8339.

24 C is correct because $F = \dfrac{\text{Mean regression sum of squares}}{\text{Mean squared error}} = \dfrac{38.4404}{7.7867} = 4.9367$.

25 C is correct. The assumptions of the linear regression model are that the 1) the relationship between the dependent variable and the independent variable is linear in the parameters b_0 and b_1; 2) the independent variable is not random; 3) the expected value of the error term is 0; 4) the variance of the error term is the same for all observations; 5) the error term is uncorrelated across observations; and 6) the error term is normally distributed. Assumption 3 is incorrect because the expected value of the error term is assumed to be zero, not equal to the mean of the dependent variable.

26 B is correct. The standard error of the estimate (SEE) for a linear regression model with one independent variable is calculated as:

$$\text{SEE} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y - \hat{b}_0 - \hat{b}_1 X_i\right)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y - \hat{Y}\right)^2}{n-2}}$$

$$= \sqrt{\frac{0.071475}{34}}$$

$$= 0.045850$$

27 C is correct. Crude oil returns explain the Amtex share returns if the slope coefficient is statistically different from zero. The slope coefficient is 0.2354 and is statistically different from zero because the absolute value of the t-statistic of 3.0974 is higher than the critical t-value of 2.032 (two-sided test for n − 2 = 34 degrees of freedom and a 5% significance level):

$$t\text{-statistic} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{0.2354 - 0.0000}{0.0760} = 3.0974$$

Therefore, Vasileva should reject the null hypothesis that crude oil returns do not explain Amtex share returns because the slope coefficient is statistically different from zero.

28 C is correct. The confidence interval for the slope coefficient is calculated as:

$$\text{Confidence interval} = \hat{b}_1 \pm t_c s_{b_1}$$

Where $\hat{b}_1 = 0.2354$, $s_{b_1} = 0.0760$ and $t_c = 2.032$

The lower limit for the confidence interval = 0.2354 – (2.032 × 0.0760) = 0.0810

The upper limit for the confidence interval = 0.2354 + (2.032 × 0.0760) = 0.3898

29 B is correct. The predicted value of the dependent variable, Amtex share return, given the value of the independent variable, crude oil return, of –0.01, is calculated as:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_i = 0.0095 + \left(0.2354 \times (-0.01)\right) = 0.0071$$

30 A is correct. The 95% prediction interval for the dependent variable given a certain value of the independent variable is calculated as:

$$\text{Prediction interval} = \hat{Y} \pm t_c s_f \text{ and the predicted value } \hat{Y} = \hat{b}_0 + \hat{b}_1 X_i$$

Therefore:

Predicted value = 0.0095 + (0.2354 × (–0.01)) = 0.0071

$s_f = (0.0022)^{0.5} = 0.0469$

$t_c = 2.032$

The lower limit for the prediction interval = 0.0071 – (2.032 × 0.0469) = –0.0882

The upper limit for the prediction interval = 0.0071 + (2.032 × 0.0469) = 0.1025

31 A is correct. If the consensus inflation forecast is unbiased, then the intercept, b_0, should equal 0, and the slope coefficient, b_1, should equal 1. The *t*-statistic for the intercept coefficient is 0.5351, which is less than the critical *t*-value of 2.0, so the intercept coefficient is not statistically different than 0. To test whether the slope coefficient equals 1, the *t*-statistic is calculated as:

$$t = \left(\hat{b}_1 - b_1\right) / s_{\hat{b}_1} = (0.9830 - 1)/0.0155 = -1.0968$$

Because the absolute value of the t-statistic of –1.0968 is less than the critical *t*-value of 2.0, the slope coefficient is not statistically different than 1. Therefore, Olabudo can conclude that the inflation forecasts are unbiased.

32 A is correct. The prediction interval for inflation is calculated in three steps:

Step 1 – Make the prediction given the US CPI forecast of 2.8:

$$\begin{aligned}\hat{Y} &= b_0 + b_1 X \\ &= 0.0001 + (0.9830 \times 2.8) \\ &= 2.7525\end{aligned}$$

Step 2 – Compute the variance of the prediction error:

$$s_f^2 = s^2\left[1 + (1/n) + \left((X - \bar{X})^2\right) / \left((n-1) \times s_x^2\right)\right]$$

$$s_f^2 = 0.0009^2\left[1 + (1/60) + \left((2.8 - 1.3350)^2\right) / \left((60-1) \times 0.7539^2\right)\right]$$

$$s_f^2 = 0.00000088$$

$$s_f = 0.0009$$

Step 3 – Compute the prediction interval:

$\hat{Y} \pm t_c \times s_f$

2.7525 ± (2.0 x 0.0009)

2.7525 – (2.0 x 0.0009) = 2.7506; lower bound

2.7525 + (2.0 x 0.0009) = 2.7544; upper bound

So, given the US CPI forecast of 2.8, the 95% prediction interval is 2.7506 to 2.7544.

33 C is correct. Public knowledge of regression relationships may negate their future usefulness in an investment context. Also, if regression assumptions are violated, hypothesis tests and predictions based on linear regression will not be valid.

READING

5

Multiple Regression

by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA

Richard A. DeFusco, PhD, CFA, is at the University of Nebraska-Lincoln (USA). Dennis W. McLeavey, DBA, CFA, is at the University of Rhode Island (USA). Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). David E. Runkle, PhD, CFA, is at Trilogy Global Advisors (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	**a.** formulate a multiple regression equation to describe the relation between a dependent variable and several independent variables and determine the statistical significance of each independent variable;
☐	**b.** interpret estimated regression coefficients and their *p*-values;
☐	**c.** formulate a null and an alternative hypothesis about the population value of a regression coefficient, calculate the value of the test statistic, and determine whether to reject the null hypothesis at a given level of significance;
☐	**d.** interpret the results of hypothesis tests of regression coefficients;
☐	**e.** calculate and interpret 1) a confidence interval for the population value of a regression coefficient and 2) a predicted value for the dependent variable, given an estimated regression model and assumed values for the independent variables;
☐	**f.** explain the assumptions of a multiple regression model;
☐	**g.** calculate and interpret the *F*-statistic, and describe how it is used in regression analysis;
☐	**h.** distinguish between and interpret the R^2 and adjusted R^2 in multiple regression;
☐	**i.** evaluate how well a regression model explains the dependent variable by analyzing the output of the regression equation and an ANOVA table;
☐	**j.** formulate a multiple regression equation by using dummy variables to represent qualitative factors and interpret the coefficients and regression results;

(continued)

LEARNING OUTCOMES	
Mastery	*The candidate should be able to:*
☐	**k.** explain the types of heteroskedasticity and how heteroskedasticity and serial correlation affect statistical inference;
☐	**l.** describe multicollinearity and explain its causes and effects in regression analysis;
☐	**m.** describe how model misspecification affects the results of a regression analysis and describe how to avoid common forms of misspecification;
☐	**n.** describe models with qualitative dependent variables;
☐	**o.** evaluate and interpret a multiple regression model and its results.

1 INTRODUCTION

As financial analysts, we often need to use more-sophisticated statistical methods than correlation analysis or regression involving a single independent variable. For example, a trading desk interested in the costs of trading NASDAQ stocks might want information on the determinants of the bid–ask spread on the NASDAQ. A mutual fund analyst might want to know whether returns to a technology mutual fund behaved more like the returns to a growth stock index or like the returns to a value stock index. An investor might be interested in the factors that determine whether analysts cover a stock. We can answer these questions using linear regression with more than one independent variable—multiple linear regression.

In Sections 2 and 3, we introduce and illustrate the basic concepts and models of multiple regression analysis. These models rest on assumptions that are sometimes violated in practice. In Section 4, we discuss three commonly occurring violations of regression assumptions. We address practical concerns such as how to diagnose an assumption violation and what remedial steps to take when a model assumption has been violated. Section 5 outlines some guidelines for building good regression models and discusses ways that analysts sometimes go wrong in this endeavor. In Section 6, we discuss a class of models whose dependent variable is qualitative in nature. These models are useful when the concern is over the occurrence of some event, such as whether a stock has analyst coverage or not.

2 MULTIPLE LINEAR REGRESSION

As investment analysts, we often hypothesize that more than one variable explains the behavior of a variable in which we are interested. The variable we seek to explain is called the dependent variable. The variables that we believe explain the dependent variable are called the independent variables.[1] A tool that permits us to examine the relationship (if any) between the two types of variables is multiple linear regression. **Multiple linear regression** allows us to determine the effect of more than one independent variable on a particular dependent variable.

1 Independent variables are also called explanatory variables or regressors.

A **multiple linear regression model** has the general form

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + \varepsilon_i, i = 1, 2, \ldots n \quad (1)$$

where

Y_i = the ith observation of the dependent variable Y

X_{ji} = the ith observation of the independent variable X_j, $j = 1, 2, \ldots, k$

b_0 = the intercept of the equation

$b_1, \ldots, b_k$ = the slope coefficients for each of the independent variables

ε_i = the error term

n = the number of observations

A slope coefficient, b_j, measures how much the dependent variable, Y, changes when the independent variable, X_j, changes by one unit, holding all other independent variables constant. For example, if $b_1 = 1$ and all of the other independent variables remain constant, then we predict that if X_1 increases by one unit, Y will also increase by one unit. If $b_1 = -1$ and all of the other independent variables are held constant, then we predict that if X_1 increases by one unit, Y will decrease by one unit. Multiple linear regression estimates $b_0, \ldots, b_k$. In this reading, we will refer to both the intercept, b_0, and the slope coefficients, $b_1, \ldots, b_k$, as **regression coefficients**. As we proceed with our discussion, keep in mind that a regression equation has k slope coefficients and $k + 1$ regression coefficients.

Although Equation 1 may seem to apply only to cross-sectional data because the notation for the observations is the same ($i = 1, \ldots, n$), all of these results apply to time-series data as well. For example, if we analyze data from many time periods for one company, we would typically use the notation $Y_t, X_{1t}, X_{2t}, \ldots, X_{kt}$, in which the first subscript denotes the variable and the second denotes the tth time period.

In practice, we use software to estimate a multiple regression model. Example 1 presents an application of multiple regression analysis in investment practice. In the course of discussing a hypothesis test, Example 1 presents typical regression output and its interpretation.

EXAMPLE 1

Explaining the Bid–Ask Spread

As the manager of the trading desk at an investment management firm, you have noticed that the average bid–ask spreads of different NASDAQ-listed stocks can vary widely. When the ratio of a stock's bid–ask spread to its price is higher than for another stock, your firm's costs of trading in that stock tend to be higher. You have formulated the hypothesis that NASDAQ stocks' percentage bid–ask spreads are related to the number of market makers and the company's stock market capitalization. You have decided to investigate your hypothesis using multiple regression analysis.

You specify a regression model in which the dependent variable measures the percentage bid–ask spread and the independent variables measure the number of market makers and the company's stock market capitalization. The regression is estimated using data from 31 December 2013 for 2,587 NASDAQ-listed stocks. Based on earlier published research exploring bid–ask spreads, you express the dependent and independent variables as natural logarithms, a so-called **log-log regression model**. A log-log regression model may be appropriate when one believes that proportional changes in the dependent variable bear a constant relationship to proportional changes in the independent variable(s), as we illustrate below. You formulate the multiple regression:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \varepsilon_i \quad (2)$$

where

Y_i = the natural logarithm of (Bid–ask spread/Stock price) for stock i

X_{1i} = the natural logarithm of the number of NASDAQ market makers for stock i

X_{2i} = the natural logarithm of the market capitalization (measured in millions of US$) of company i

In a log-log regression such as Equation 2, the slope coefficients are interpreted as elasticities, assumed to be constant. For example, a value of b_2 = −0.75 would mean that for a 1 percent increase in the market capitalization, we expect Bid–ask spread/Stock price to decrease by 0.75 percent, holding all other independent variables constant.[2]

Reasoning that greater competition tends to lower costs, you suspect that the greater the number of market makers, the smaller the percentage bid–ask spread. Therefore, you formulate a first null hypothesis (H_0) and alternative hypothesis (H_a):

H_0: $b_1 \geq 0$

H_a: $b_1 < 0$

The null hypothesis is the hypothesis that the "suspected" condition is not true. If the evidence supports rejecting the null hypothesis and accepting the alternative hypothesis, you have statistically confirmed your suspicion.[3]

You also believe that the stocks of companies with higher market capitalization may have more-liquid markets, tending to lower percentage bid–ask spreads. Therefore, you formulate a second null hypothesis and alternative hypothesis:

H_0: $b_2 \geq 0$

H_a: $b_2 < 0$

For both tests, we use a t-test, rather than a z-test, because we do not know the population variance of b_1 and b_2. Suppose that you choose a 0.01 significance level for both tests.

Exhibit 1 Results from Regressing ln(Bid–Ask Spread/Price) on ln(Number of Market Makers) and ln(Market Capitalization)

	Coefficient	Standard Error	*t*-Statistic
Intercept	1.5949	0.2275	7.0105
ln(Number of NASDAQ market makers)	−1.5186	0.0808	−18.7946
ln(Company's market capitalization)	−0.3790	0.0151	−25.0993

2 Note that $\Delta(\ln X) \approx \Delta X/X$, where Δ represents "change in" and $\Delta X/X$ is a proportional change in X. We discuss the model further in Example 11.

3 An alternative valid formulation is a two-sided test (H_0: $b_1 = 0$ versus H_a: $b_1 \neq 0$) which reflects the beliefs of the researcher less strongly. A two-sided test could also be conducted for the hypothesis on market capitalization that we discuss next.

Exhibit 1 (Continued)

ANOVA	df	SS	MSS	F	Significance F
Regression	2	3,728.1334	1,864.0667	2,216.75	0.00
Residual	2,584	2,172.8870	0.8409		
Total	2,586	5,901.0204			
Residual standard error			0.9170		
Multiple R-squared			0.6318		
Observations			2,587		

Source: Center for Research in Security Prices, University of Chicago.

Exhibit 1 shows the results of estimating this linear regression using data from 31 December 2013.

If the regression result is not significant, we follow the useful principle of not proceeding to interpret the individual regression coefficients. Thus the analyst might look first at the **analysis of variance (ANOVA)** section, which addresses the regression's overall significance.

- The ANOVA (analysis of variance) section reports quantities related to the overall explanatory power and significance of the regression. SS stands for sum of squares, and MSS stands for mean sum of squares (SS divided by df). The F-test reports the overall significance of the regression. For example, an entry of 0.01 for the significance of F means that the regression is significant at the 0.01 level. In Exhibit 1, the regression is even more significant because the significance of F is 0 at two decimal places. Later in the reading, we will present more information on the F-test.

Having ascertained that the overall regression is highly significant, an analyst might turn to the first listed column in the first section of the regression output.

- The Coefficient column gives the estimates of the intercept, b_0, and the slope coefficients, b_1 and b_2. The estimated intercept is positive, but both estimated slope coefficients are negative. Are these estimated regression coefficients significantly different from zero? The Standard Error column gives the standard error (the standard deviation) of the estimated regression coefficients. The test statistic for hypotheses concerning the population value of a regression coefficient has the form (Estimated regression coefficient – Hypothesized population value of the regression coefficient)/(Standard error of the regression coefficient). This is a t-test. Under the null hypothesis, the hypothesized population value of the regression coefficient is 0. Thus (Estimated regression coefficient)/(Standard error of the regression coefficient) is the t-statistic given in the third column. For example, the t-statistic for the intercept is 1.5949/0.2275 = 7.0105. To evaluate the significance of the t-statistic we need to determine a quantity called degrees of freedom (df).[4] The calculation is Degrees of freedom = Number of observations − (Number of independent variables + 1) = $n - (k + 1)$.

- The final section of Exhibit 1 presents two measures of how well the estimated regression fits or explains the data. The first is the standard deviation of the regression residual, the residual standard error. This standard deviation is called the standard error of estimate (SEE). The second measure quantifies the degree of linear association between the dependent variable and all of the independent variables jointly. This measure is known as multiple R^2 or simply R^2 (the square of the correlation between predicted and actual values of the dependent variable).[5] A value of 0 for R^2 indicates no linear association; a value of 1 indicates perfect linear association. The final item in Exhibit 1 is the number of observations in the sample (2,587).

Having reviewed the meaning of typical regression output, we can return to complete the hypothesis tests. The estimated regression supports the hypothesis that the greater the number of market makers, the smaller the percentage bid–ask spread: We reject H_0: $b_1 \geq 0$ in favor of H_a: $b_1 < 0$. The results also support the belief that the stocks of companies with higher market capitalization have lower percentage bid–ask spreads: We reject H_0: $b_2 \geq 0$ in favor of H_a: $b_2 < 0$.

To see that the null hypothesis is rejected for both tests, we can use t-test tables. For both tests, df = 2,587 − 3 = 2,584. The tables do not give critical values for degrees of freedom that large. The critical value for a one-tailed test with df = 200 at the 0.01 significance level is 2.345; for a larger number of degrees of freedom, the critical value would be even smaller in magnitude. Therefore, in our one-sided tests, we reject the null hypothesis in favor of the alternative hypothesis if

$$t = \frac{\hat{b}_j - b_j}{s_{\hat{b}_j}} = \frac{\hat{b}_j - 0}{s_{\hat{b}_j}} < -2.345$$

where

$\hat{b}_j$ = the regression estimate of b_j, j = 1, 2

b_j = the hypothesized value[6] of the coefficient (0)

$s_{\hat{b}_j}$ = the estimated standard error of $\hat{b}_j$

The t-values of −18.7946 and −25.0993 for the estimates of b_1 and b_2, respectively, are both less than −2.345.

Before proceeding further, we should address the interpretation of a prediction stated in natural logarithm terms. We can convert a natural logarithm to the original units by taking the antilogarithm. To illustrate this conversion, suppose that a particular stock has 20 NASDAQ market makers and a market capitalization of \$100 million. The natural logarithm of the number of NASDAQ market makers is equal to ln 20 = 2.9957, and the natural logarithm of the company's market cap (in millions) is equal to ln 100 = 4.6052. With these values, the regression model predicts that the natural log of the ratio of the bid–ask spread

4 To calculate the degrees of freedom lost in the regression, we add 1 to the number of independent variables to account for the intercept term.

5 Multiple R^2 is also known as the multiple coefficient of determination, or simply the coefficient of determination.

6 To economize on notation in stating test statistics, in this context we use b_j to represent the hypothesized value of the parameter (elsewhere we use it to represent the unknown population parameter).

to the stock price will be 1.5949 + (−1.5186 × 2.9957) + (−0.3790 × 4.6052) = −4.6997. We take the antilogarithm of −4.6997 by raising e to that power: $e^{-4.6997}$ = 0.0091. The predicted bid–ask spread will be 0.91 percent of the stock price.[7] Later we state the assumptions of the multiple regression model; before using an estimated regression to make predictions in actual practice, we should assure ourselves that those assumptions are satisfied.

In Exhibit 1, we presented output common to most regression software programs. Many software programs also report p-values for the regression coefficients.[8] For each regression coefficient, the p-value would be the smallest level of significance at which we can reject a null hypothesis that the population value of the coefficient is 0, in a two-sided test. The lower the p-value, the stronger the evidence against that null hypothesis. A p-value quickly allows us to determine if an independent variable is significant at a conventional significance level such as 0.05, or at any other standard we believe is appropriate.

Having estimated Equation 1, we can write

$$\begin{aligned}\hat{Y}_i &= \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} \\ &= 1.5949 - 1.5186 X_{1i} - 0.3790 X_{2i}\end{aligned}$$

where $\hat{Y}_i$ stands for the predicted value of Y_i, and $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$, stand for the estimated values of b_0, b_1, and b_2, respectively. How should we interpret the estimated slope coefficients −1.5186 and −0.3790?

Interpreting the slope coefficients in a multiple linear regression model is different than doing so in the one-independent-variable regressions explored in the reading on correlation and regression. Suppose we have a one-independent-variable regression that we estimate as $\hat{Y}_i = 0.50 + 0.75X_{1i}$. The interpretation of the slope estimate 0.75 is that for every 1-unit increase in X_1, we expect Y to increase by 0.75 units. If we were to add a second independent variable to the equation, we would generally find that the estimated coefficient on X_1 is *not* 0.75 unless the second independent variable were uncorrelated with X_1. The slope coefficients in a multiple regression are known as **partial regression coefficients** or **partial slope coefficients** and need to be interpreted with care.[9] Suppose the coefficient on X_1 in a regression with the second independent variable was 0.60. Can we say that for every 1-unit increase in X_1, we expect Y to increase by 0.60 units? Not without qualification. For every 1-unit increase in X_1, we still expect Y to increase by 0.75 units when X_2 is not held constant. We would interpret 0.60 as the expected increase in Y for a 1-unit increase X_1 *holding the second independent variable constant.*

To explain what the shorthand reference "holding the second independent constant" refers to, if we were to regress X_1 on X_2, the residuals from that regression would represent the part of X_1 that is uncorrelated with X_2. We could then regress Y on those residuals in a one-independent-variable regression. We would find that the slope coefficient on the residuals would be 0.60; by construction, 0.60 would represent the expected effect on Y of a 1-unit increase in X_1 after removing the part of X_1 that is correlated with X_2. Consistent with this explanation, we can view 0.60 as the expected net effect on Y of a 1-unit increase in X_1, after accounting for any

7 The operation illustrated (taking the antilogarithm) recovers the value of a variable in the original units as $e^{\ln X} = X$.

8 The entry 0.00 for the significance of F was a p-value for the F-test.

9 The terminology comes from the fact that they correspond to the partial derivatives of Y with respect to the independent variables. Note that in this usage, the term "regression coefficients" refers just to the slope coefficients.

effects of the other independent variables on the expected value of Y. To reiterate, a partial regression coefficient measures the expected change in the dependent variable for a 1-unit increase in an independent variable, holding all the other independent variables constant.

To apply this process to the regression in Exhibit 1, we see that the estimated coefficient on the natural logarithm of market capitalization is −0.3790. Therefore, the model predicts that an increase of 1 in the natural logarithm of the company's market capitalization is associated with a −0.3790 change in the natural logarithm of the ratio of the bid–ask spread to the stock price, holding the natural logarithm of the number of market makers constant. We need to be careful not to expect that the natural logarithm of the ratio of the bid–ask spread to the stock price would differ by −0.3790 if we compared two stocks for which the natural logarithm of the company's market capitalization differed by 1, because in all likelihood the number of market makers for the two stocks would differ as well, which would affect the dependent variable. The value −0.3790 is the expected net effect of difference in log market capitalizations, net of the effect of the log number of market makers on the expected value of the dependent variable.

2.1 Assumptions of the Multiple Linear Regression Model

Before we can conduct correct statistical inference on a multiple linear regression model (a model with more than one independent variable estimated using ordinary least squares), we need to know the assumptions underlying that model.[10] Suppose we have n observations on the dependent variable, Y, and the independent variables, X_1, X_2, ..., X_k, and we want to estimate the equation $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + ... + b_kX_{ki} + \varepsilon_i$.

In order to make a valid inference from a multiple linear regression model, we need to make the following six assumptions, which as a group define the classical normal multiple linear regression model:

1 The relationship between the dependent variable, Y, and the independent variables, X_1, X_2, ..., X_k, is linear as described in Equation 1.

2 The independent variables (X_1, X_2, ..., X_k) are not random.[11] Also, no exact linear relation exists between two or more of the independent variables.[12]

3 The expected value of the error term, conditioned on the independent variables, is 0: $E(\varepsilon \mid X_1,X_2, \ldots, X_k) = 0$.

4 The variance of the error term is the same for all observations:[13] $E\left(\varepsilon_i^2\right) = \sigma_\varepsilon^2$.

5 The error term is uncorrelated across observations: $E(\varepsilon_i\varepsilon_j) = 0$, $j \neq i$.

6 The error term is normally distributed.

Note that these assumptions are almost exactly the same as those for the single-variable linear regression model. Assumption 2 is modified such that no exact linear relation exists between two or more independent variables or combinations of

10 Ordinary least squares (OLS) is an estimation method based on the criterion of minimizing the sum of the squared residuals of a regression.

11 As discussed in the reading on correlation and regression, even though we assume that independent variables in the regression model are not random, often that assumption is clearly not true. For example, the monthly returns to the S&P 500 are not random. If the independent variable is random, then is the regression model incorrect? Fortunately, no. Even if the independent variable is random but uncorrelated with the error term, we can still rely on the results of regression models. See, for example, Greene (2018).

12 No independent variable can be expressed as a linear combination of any set of the other independent variables. Technically, a constant equal to 1 is included as an independent variable associated with the intercept in this condition.

13 $\text{Var}(\varepsilon) = E(\varepsilon^2)$ and $\text{Cov}(\varepsilon_i\varepsilon_j) = E(\varepsilon_i\varepsilon_j)$ because $E(\varepsilon) = 0$.

independent variables. If this part of Assumption 2 is violated, then we cannot compute linear regression estimates.[14] Also, even if no exact linear relationship exists between two or more independent variables, or combinations of independent variables, linear regression may encounter problems if two or more of the independent variables or combinations thereof are highly correlated. Such a high correlation is known as multicollinearity, which we will discuss later in this reading. We will also discuss the consequences of conducting regression analysis premised on Assumptions 4 and 5 being met when, in fact, they are violated.

Although Equation 1 may seem to apply only to cross-sectional data because the notation for the observations is the same ($i = 1, ..., n$), all of these results apply to time-series data as well. For example, if we analyze data from many time periods for one company, we would typically use the notation $Y_t, X_{1t}, X_{2t}, ..., X_{kt}$, in which the first subscript denotes the variable and the second denotes the tth time period.

EXAMPLE 2

Factors Explaining the Valuations of Multinational Corporations

Kyaw, Manley, and Shetty (2011) examined which factors affect the valuation of a multinational corporation (MNC). Specifically, they wanted to know whether political risk, transparency, and geographic diversification affected the valuations of MNCs. They used data for 450 US MNCs from 1998 to 2003. The valuations of these corporations were measured using Tobin's q, a commonly used measure of corporate valuation that is calculated as the ratio of the sum of the market value of a corporation's equity and the book value of long-term debt to the sum of the book values of equity and long-term debt. The authors regressed Tobin's q of MNCs on variables representing political risk, transparency, and geographic diversification. The authors also included some additional variables that may affect company valuation, including size, leverage, and beta.[15] They used the equation

$$\text{Tobin's } q_{i,t} = b_0 + b_1(\text{Size}_{i,t}) + b_2(\text{Leverage}_{i,t}) + b_3(\text{Beta}_{i,t}) + b_4(\text{Political risk}_{i,t}) + b_5(\text{Transparency}_{i,t}) + b_6(\text{Geographic diversification}_{i,t}) + \varepsilon_{i,t}$$

14 When we encounter this kind of linear relationship (called perfect collinearity), we cannot compute the matrix inverse needed to compute the linear regression estimates. See Greene (2018) for a further description of this issue.

15 As mentioned in an earlier footnote, technically a constant equal to 1 is included as an independent variable associated with the intercept term in a regression. Because all the regressions reported in this reading include an intercept term, we will not separately mention a constant as an independent variable in the remainder of this reading.

where

Tobin's $q_{i,t}$ = the Tobin's q for MNC i in year t, with Tobin's q computed as (Market value of equity + Book value of long-term debt)/(Book value of equity + Book value of long-term debt)

$\text{Size}_{i,t}$ = the natural log of the total sales of MNC i in the year t in millions of US$

$\text{Leverage}_{i,t}$ = the ratio of total debt to total assets of MNC i in year t

$\text{Beta}_{i,t}$ = the beta of the stock of MNC i in year t

$\text{Political risk}_{i,t}$ = the at-risk-proportion of international operations of MNC i in year t, calculated as [1 – (number of safe countries/total number of foreign countries in which the firm has operations)], using national risk coding from *Euromoney*

$\text{Transparency}_{i,t}$ = the "transparency percent" (representing the level of disclosure) of MNC i in year t, using survey data from *S&P Transparency & Disclosure*

$\text{Geographic diviersification}_{i,t}$ = foreign sales of MNC i in year t expressed as a percentage of its total sales in that year

Exhibit 2 shows the results of their analysis.[16]

Exhibit 2 Results from Regressing Tobin's *q* on Factors Affecting the Value of Multinational Corporations

	Coefficient	Standard Error*	*t*-Statistic
Intercept	19.829	4.798	4.133
Size	–0.712	0.228	–3.123
Leverage	–3.897	0.987	–3.948
Beta	–1.032	0.261	–3.954
Political risk	–2.079	0.763	–2.725
Transparency	–0.129	0.050	–2.580
Geographic diversification	0.021	0.010	2.100

* This study combines time series observations with cross-sectional observations; such data are commonly referred to as panel data. In such a setting, the standard errors need to be corrected for bias by using a clustered standard error approach as in Petersen (2009). The standard errors reported in this exhibit are clustered standard errors.

Source: Kyaw, Manley, and Shetty (2011).

16 Size is the natural log of total sales. A log transformation (either natural log or log base 10) is commonly used for independent variables that can take a wide range of values; company size and fund size are two such variables. One reason to use the log transformation is to improve the statistical properties of the residuals. If the authors had not taken the log of sales and instead used sales as the independent variable, the regression model probably would not have explained Tobin's *q* as well.

Suppose that we use the results in Exhibit 2 to test the null hypothesis that the size of a multinational corporation has no effect on its value. Our null hypothesis is that the coefficient on the size variable equals 0 (H_0: b_1 = 0), and our alternative hypothesis is that the coefficient does not equal 0 (H_a: $b_1 \neq 0$). The t-statistic for testing that hypothesis is

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{-0.712 - 0}{0.228} = -3.12$$

With 450 observations and seven coefficients, the t-statistic has 450 − 7 = 443 degrees of freedom. At the 0.05 significance level, the critical value for t is about 1.97. The absolute value of computed t-statistic on the size coefficient is 3.12, which suggests strongly that we can reject the null hypothesis that size is unrelated to MNC value. In fact, the critical value for t is about 2.6 at the 0.01 significance level.

Because $\text{Size}_{i,t}$ is the natural (base e or 2.72) log of sales, an increase of 1 in $\text{Size}_{i,t}$ is the same as a 2.72-fold increase in sales. Thus, the estimated coefficient of approximately −0.7 for $\text{Size}_{i,t}$ implies that every 2.72-fold increase in sales of the MNC (an increase of 1 in $\text{Size}_{i,t}$) is associated with an expected decrease of 0.7 in Tobin's $q_{i,t}$ of the MNC, *holding constant the other five independent variables in the regression*.

Now suppose we want to test the null hypothesis that geographic diversification is not related to Tobin's q; we want to test whether the coefficient on geographic diversification equals 0 (H_0: b_6 = 0) against the alternative hypothesis that the coefficient on geographic diversification does not equal 0 (H_a: $b_6 \neq 0$). The t-statistic to test this hypothesis is

$$t = \frac{\hat{b}_6 - b_6}{s_{\hat{b}_6}} = \frac{0.021 - 0}{0.010} = 2.10$$

The critical value of the t-test is 1.97 at the 0.05 significance level. Therefore, at the 0.05 significance level, we can reject the null hypothesis that geographic diversification has no effect on MNC valuation. We can interpret the coefficient on geographic diversification of 0.021 as implying that an increase of 1 in the percentage of MNC's sales that are foreign sales is associated with an expected 0.021 increase in Tobin's q for the MNC, holding all other independent variables constant.

EXAMPLE 3

Explaining Returns to the Fidelity Select Technology Portfolio

Suppose you are considering an investment in the Fidelity Select Technology Portfolio (FSPTX), a US mutual fund specializing in technology stocks. You want to know whether the fund behaves more like a large-cap growth fund or a large-cap value fund.[17] You decide to estimate the regression

$$Y_t = b_0 + b_1X_{1t} + b_2X_{2t} + \varepsilon_t$$

17 This regression is related to return-based style analysis, one of the most frequent applications of regression analysis in the investment profession. For more information, see Sharpe (1988), who pioneered this field, and Buetow, Johnson, and Runkle (2000).

where

Y_t = the monthly return to the FSPTX
X_{1t} = the monthly return to the S&P 500 Growth Index
X_{2t} = the monthly return to the S&P 500 Value Index

The S&P 500 Growth and S&P 500 Value indices represent predominantly large-cap growth and value stocks, respectively.

Exhibit 3 shows the results of this linear regression using monthly data from January 2009 through December 2013. The estimated intercept in the regression is 0.0018. Thus, if both the return to the S&P 500 Growth Index and the return to the S&P 500 Value Index equal 0 in a specific month, the regression model predicts that the return to the FSPTX will be 0.18 percent. The coefficient on the large-cap growth index is 1.4697, and the coefficient on the large-cap value index return is −0.1833. Therefore, if in a given month the return to the S&P 500 Growth Index was 1 percent and the return to the S&P 500 Value Index was −2 percent, the model predicts that the return to the FSPTX would be 0.0018 + 1.4697(0.01) − 0.1833(−0.02) = 2.02 percent.

Exhibit 3 Results from Regressing the FSPTX Returns on the S&P 500 Growth and S&P 500 Value Indices

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0018	0.0038	0.4737
S&P 500 Growth Index	1.4697	0.2479	5.9286
S&P 500 Value Index	−0.1833	0.2034	−0.9012

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	0.1653	0.0826	113.7285	1.27E-20
Residual	57	0.0414	0.0007		
Total	59	0.2067			
Residual standard error			0.0270		
Multiple *R*-squared			0.7996		
Observations			60		

Source: Bloomberg, finance.yahoo.com.

We may want to know whether the coefficient on the returns to the S&P 500 Value Index is statistically significant. Our null hypothesis states that the coefficient equals 0 (H_0: $b_2 = 0$); our alternative hypothesis states that the coefficient does not equal 0 (H_a: $b_2 \neq 0$).

Our test of the null hypothesis uses a t-test constructed as follows:

$$t = \frac{\hat{b}_2 - b_2}{s_{\hat{b}_2}} = \frac{-0.1833 - 0}{0.2034} = -0.9012$$

where

$\hat{b}_2$ = the regression estimate of b_2

b_2 = the hypothesized value[18] of the coefficient (0)

$s_{\hat{b}_2}$ = the estimated standard error of $\hat{b}_2$

This regression has 60 observations and three coefficients (two independent variables and the intercept); therefore, the t-test has 60 − 3 = 57 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is about 2.00. The absolute value of the test statistic is 0.9012. Because the test statistic's absolute value is less than the critical value (0.9012 < 2.00), we fail to reject the null hypothesis that $b_2 = 0$. (Note that the t-tests reported in Exhibit 3, as well as the other regression tables, are tests of the null hypothesis that the population value of a regression coefficient equals 0.)

Similar analysis shows that at the 0.05 significance level, we cannot reject the null hypothesis that the intercept equals 0 (H_0: $b_0 = 0$) in favor of the alternative hypothesis that the intercept does not equal 0 (H_a: $b_0 \neq 0$). Exhibit 3 shows that the t-statistic for testing that hypothesis is 0.4737, a result smaller in absolute value than the critical value of 2.00. However, at the 0.05 significance level we *can* reject the null hypothesis that the coefficient on the S&P 500 Growth Index equals 0 (H_0: $b_1 = 0$) in favor of the alternative hypothesis that the coefficient does not equal 0 (H_a: $b_1 \neq 0$). As Exhibit 3 shows, the t-statistic for testing that hypothesis is 5.928, a result far above the critical value of 2.00. Thus multiple regression analysis suggests that returns to the FSPTX are very closely associated with the returns to the S&P 500 Growth Index, but they are not related to S&P 500 Value Index (the t-statistic of 0.9012 is not statistically significant).

2.2 Predicting the Dependent Variable in a Multiple Regression Model

Financial analysts often want to predict the value of the dependent variable in a multiple regression based on assumed values of the independent variables. We have previously discussed how to make such a prediction in the case of only one independent variable. The process for making that prediction with multiple linear regression is very similar.

To predict the value of a dependent variable using a multiple linear regression model, we follow these three steps:

1 Obtain estimates $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$ of the regression parameters $b_0, b_1, b_2, \ldots, b_k$.

2 Determine the assumed values of the independent variables, $\hat{X}_{1i}, \hat{X}_{2i}, \ldots, \hat{X}_{ki}$.

3 Compute the predicted value of the dependent variable, $\hat{Y}_i$, using the equation

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1\hat{X}_{1i} + \hat{b}_2\hat{X}_{2i} + \ldots + \hat{b}_k\hat{X}_{ki} \quad (3)$$

Two practical points concerning using an estimated regression to predict the dependent variable are in order. First, we should be confident that the assumptions of the regression model are met. Second, we should be cautious about predictions based on values of the independent variables that are outside the range of the data on which the model was estimated; such predictions are often unreliable.

18 To economize on notation in stating test statistics, in this context we use b_2 to represent the hypothesized value of the parameter (elsewhere we use it to represent the unknown population parameter).

EXAMPLE 4

Predicting a Multinational Corporation's Tobin's *q*

In Example 2, we explained the Tobin's q for US multinational corporations (MNC) based on the natural log of sales, leverage, beta, political risk, transparency, and geographic diversification. To review the regression equation:

$$\text{Tobin's } q_{i,t} = b_0 + b_1(\text{Size}_{i,t}) + b_2(\text{Leverage}_{i,t}) + b_3(\text{Beta}_{i,t}) + b_4(\text{Political risk}_{i,t}) + b_5(\text{Transparency}_{i,t}) + b_6(\text{Geographic diversification}_{i,t}) + \varepsilon_i$$

Now we can use the results of the regression reported in Exhibit 2 (excerpted here) to predict the Tobin's q for a US MNC.

Exhibit 2 (excerpt)

	Coefficient
Intercept	19.829
Size	–0.712
Leverage	–3.897
Beta	–1.032
Political risk	–2.079
Transparency	–0.129
Geographic diversification	0.021

1 Suppose that a particular MNC has the following data for a given year.

- Total sales of $7,600 million. The natural log of total sales in millions of US$ equals ln(7,600) = 8.94.
- Leverage (Total debt/Total assets) of 0.45.
- Beta of 1.30.
- Political risk of 0.47, implying that the ratio of the number of safe countries to the total number of foreign countries in which the MNC has operations is 0.53.
- Transparency score of 65, indicating 65% "yes" answers to survey questions related to the corporation's transparency.
- Geographic diversification of 30, indicating that 30% of the corporation's sales are in foreign countries.

What is the predicted Tobin's q for the above MNC?

Solution to 1:

The predicted Tobin's q for the MNC, based on the regression, is:

$$19.829 + (-0.712 \times 8.94) + (-3.897 \times 0.45) + (-1.032 \times 1.30) + (-2.079 \times 0.47) + (-0.129 \times 65) + (0.021 \times 30) = 1.64$$

When predicting the dependent variable using a linear regression model, we encounter two types of uncertainty: uncertainty in the regression model itself, as reflected in the standard error of estimate, and uncertainty about the estimates of the regression model's parameters. In the reading on correlation and regression, we presented procedures for constructing a prediction interval for linear regression with one independent variable. For multiple regression, however, computing a prediction interval to properly incorporate both types of uncertainty requires matrix algebra, which is outside the scope of this reading.[19]

2.3 Testing whether All Population Regression Coefficients Equal Zero

Earlier, we illustrated how to conduct hypothesis tests on regression coefficients individually. What if we now want to test the significance of the regression as a whole? As a group, do the independent variables help explain the dependent variable? To address this question, we test the null hypothesis that all the slope coefficients in a regression are simultaneously equal to 0. In this section, we further discuss ANOVA with regard to a regression's explanatory power and the inputs for an F-test of the above null hypothesis.

If none of the independent variables in a regression model helps explain the dependent variable, the slope coefficients should all equal 0. In a multiple regression, however, we cannot test the null hypothesis that *all* slope coefficients equal 0 based on t-tests that *each individual* slope coefficient equals 0, because the individual tests do not account for the effects of interactions among the independent variables. For example, a classic symptom of multicollinearity is that we can reject the hypothesis that all the slope coefficients equal 0 even though none of the t-statistics for the individual estimated slope coefficients is significant. Conversely, we can construct unusual examples in which the estimated slope coefficients are significantly different from 0 although jointly they are not.

To test the null hypothesis that all of the slope coefficients in the multiple regression model are jointly equal to 0 (H_0: $b_1 = b_2 = \dots = b_k = 0$) against the alternative hypothesis that at least one slope coefficient is not equal to 0 we must use an F-test. The F-test is viewed as a test of the regression's overall significance.

To correctly calculate the test statistic for the null hypothesis, we need four inputs:

- total number of observations, n;
- total number of regression coefficients to be estimated, $k + 1$, where k is the number of slope coefficients;
- sum of squared errors or residuals, $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$, abbreviated SSE, also known as the residual sum of squares (unexplained variation);[20] and
- regression sum of squares, $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$, abbreviated RSS.[21] This amount is the variation in Y from its mean that the regression equation explains (explained variation).

The F-test for determining whether the slope coefficients equal 0 is based on an F-statistic calculated using the four values listed above. The F-statistic measures how well the regression equation explains the variation in the dependent variable; it is the ratio of the mean regression sum of squares to the mean squared error.

19 For more information, see Greene (2018).

20 In a table of regression output, this is the number under "SS" column in the row "Residual."

21 In a table of regression output, this is the number under the "SS" column in the row "Regression."

We compute the mean regression sum of squares by dividing the regression sum of squares by the number of slope coefficients estimated, k. We compute the mean squared error by dividing the sum of squared errors by the number of observations, n, minus (k + 1). The two divisors in these computations are the degrees of freedom for calculating an F-statistic. For n observations and k slope coefficients, the F-test for the null hypothesis that the slope coefficients are all equal to 0 is denoted $F_{k,n-(k+1)}$. The subscript indicates that the test should have k degrees of freedom in the numerator (numerator degrees of freedom) and $n - (k + 1)$ degrees of freedom in the denominator (denominator degrees of freedom).

The formula for the F-statistic is

$$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n-(k+1)\right]} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}} = \frac{\text{MSR}}{\text{MSE}} \quad (4)$$

where MSR is the mean regression sum of squares and MSE is the mean squared error. In our regression output tables, MSR and MSE are the first and second quantities under the MSS (mean sum of squares) column in the ANOVA section of the output. If the regression model does a good job of explaining variation in the dependent variable, then the ratio MSR/MSE will be large.

What does this F-test tell us when the independent variables in a regression model explain none of the variation in the dependent variable? In this case, each predicted value in the regression model, $\hat{Y}_i$, has the average value of the dependent variable, $\overline{Y}$, and the regression sum of squares, $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$ is 0. Therefore, the F-statistic for testing the null hypothesis (that all the slope coefficients are equal to 0) has a value of 0 when the independent variables do not explain the dependent variable at all.

To specify the details of making the statistical decision when we have calculated F, we reject the null hypothesis at the α significance level if the calculated value of F is greater than the upper α critical value of the F distribution with the specified numerator and denominator degrees of freedom. Note that we use a one-tailed F-test.[22]

We can illustrate the test using Example 1, in which we investigated whether the natural log of the number of NASDAQ market makers and the natural log of the stock's market capitalization explained the natural log of the bid–ask spread divided by price. Assume that we set the significance level for this test to α = 0.05 (i.e., a 5 percent probability that we will mistakenly reject the null hypothesis if it is true). Exhibit 1 (excerpted here) presents the results of variance computations for this regression.

Exhibit 1 (excerpt)

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	3,728.1334	1,864.0667	2,216.7505	0.00
Residual	2,584	2,172.8870	0.8409		
Total	2,586	5,901.0204			

This model has two slope coefficients (k = 2), so there are two degrees of freedom in the numerator of this F-test. With 2,587 observations in the sample, the number of degrees of freedom in the denominator of the F-test is $n - (k + 1) = 2{,}587 - 3 =$

22 We use a one-tailed test because MSR necessarily increases relative to MSE as the explanatory power of the regression increases.

2,584. The sum of the squared errors is 2,172.8870. The regression sum of squares is 3,728.1334. Therefore, the F-test for the null hypothesis that the two slope coefficients in this model equal 0 is

$$\frac{3,728.1334/2}{2,172.8870/2,584} = 2,216.7505$$

This test statistic is distributed as an $F_{2,2,584}$ random variable under the null hypothesis that the slope coefficients are equal to 0. In the Exhibit for the 0.05 significance level, we look at the second column, which shows F-distributions with two degrees of freedom in the numerator. Near the bottom of the column, we find that the critical value of the F-test needed to reject the null hypothesis is between 3.00 and 3.07.[23] The actual value of the F-test statistic at 2,216.75 is much greater, so we reject the null hypothesis that coefficients of both independent variables equal 0. In fact, Exhibit 1 under "Significance F," reports a p-value of 0. This p-value means that the smallest level of significance at which the null hypothesis can be rejected is practically 0. The large value for this F-statistic implies a minuscule probability of incorrectly rejecting the null hypothesis (a mistake known as a Type I error).

2.4 Adjusted R^2

In the reading on correlation and regression, we presented the coefficient of determination, R^2, as a measure of the goodness of fit of an estimated regression to the data. In a multiple linear regression, however, R^2 is less appropriate as a measure of whether a regression model fits the data well (goodness of fit). Recall that R^2 is defined as

$$\frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}}$$

The numerator equals the regression sum of squares, RSS. Thus R^2 states RSS as a fraction of the total sum of squares, $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$. If we add regression variables to the model, the amount of unexplained variation will decrease, and RSS will increase, if the new independent variable explains any of the unexplained variation in the model. Such a reduction occurs when the new independent variable is even slightly correlated with the dependent variable and is not a linear combination of other independent variables in the regression.[24] Consequently, we can increase R^2 simply by including many additional independent variables that explain even a slight amount of the previously unexplained variation, even if the amount they explain is not statistically significant.

Some financial analysts use an alternative measure of goodness of fit called **adjusted R^2**, or $\overline{R}^2$. This measure of fit does not automatically increase when another variable is added to a regression; it is adjusted for degrees of freedom. Adjusted R^2 is typically part of the multiple regression output produced by statistical software packages.

The relation between R^2 and $\overline{R}^2$ is

$$\overline{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1 - R^2\right)$$

23 We see a range of values because the denominator has more than 120 degrees of freedom but less than an infinite number of degrees of freedom.

24 We say that variable y is a linear combination of variables x and z if $y = ax + bz$ for some constants a and b. A variable can also be a linear combination of more than two variables.

where n is the number of observations and k is the number of independent variables (the number of slope coefficients). Note that if $k \geq 1$, then R^2 is strictly greater than adjusted R^2. When a new independent variable is added, $\bar{R}^2$ can decrease if adding that variable results in only a small increase in R^2. In fact, $\bar{R}^2$ can be negative, although R^2 is always nonnegative.[25] If we use $\bar{R}^2$ to compare regression models, it is important that the dependent variable be defined the same way in both models and that the sample sizes used to estimate the models are the same.[26] For example, it makes a difference for the value of $\bar{R}^2$ if the dependent variable is GDP (gross domestic product) or ln(GDP), even if the independent variables are identical. Furthermore, we should be aware that a high $\bar{R}^2$ does not necessarily indicate that the regression is well specified in the sense of including the correct set of variables.[27] One reason for caution is that a high $\bar{R}^2$ may reflect peculiarities of the dataset used to estimate the regression. To evaluate a regression model, we need to take many other factors into account, as we discuss in Section 5.1.

3 USING DUMMY VARIABLES IN REGRESSIONS

Often, financial analysts need to use qualitative variables as independent variables in a regression. One type of qualitative variable, called a **dummy variable**, takes on a value of 1 if a particular condition is true and 0 if that condition is false.[28] For example, suppose we want to test whether stock returns were different in January than during the remaining months of a particular year. We include one independent variable in the regression, X_{1t}, that has a value of 1 for each January and a value of 0 for every other month of the year. We estimate the regression model

$$Y_t = b_0 + b_1 X_{1t} + \varepsilon_t$$

In this equation, the coefficient b_0 is the average value of Y_t in months other than January, and b_1 is the difference between the average value of Y_t in January and the average value of Y_t in months other than January.

We need to exercise care in choosing the number of dummy variables in a regression. The rule is that if we want to distinguish among n categories, we need $n - 1$ dummy variables. For example, to distinguish between *during January* and *not during January* above ($n = 2$ categories), we used one dummy variable ($n - 1 = 2 - 1 = 1$). If we want to distinguish between each of the four quarters in a year, we would include dummy variables for three of the four quarters in a year. If we make the mistake of including dummy variables for four rather than three quarters, we have violated Assumption 2 of the multiple regression model and cannot estimate the regression. The next example illustrates the use of dummy variables in a regression with monthly data.

---

25 When $\bar{R}^2$ is negative, we can effectively consider its value to be 0.

26 See Gujarati, Porter, and Gunasekar (2011). The value of adjusted R^2 depends on sample size. These points hold if we are using R^2 to compare two regression models.

27 See Mayer (1980).

28 Not all qualitative variables are simple dummy variables. For example, in a trinomial choice model (a model with three choices), a qualitative variable might have the value 0, 1, or 2.

EXAMPLE 5

Month-of-the-Year Effects on Japanese Small-Stock Returns

For many years, financial analysts have been concerned about seasonality in stock returns.[29] In particular, analysts have researched whether returns to small stocks differ during various months of the year. Suppose we want to test whether total returns to one small-stock index, the MSCI Japan Small Cap Index, differ by month. Using data from January 2001 (the first available date for these data) through the end of 2013, we can estimate a regression including an intercept and 11 dummy variables, one for each of the first 11 months of the year. The equation that we estimate is

$$\text{Returns}_t = b_0 + b_1\text{Jan}_t + b_2\text{Feb}_t + \ldots + b_{11}\text{Nov}_t + \varepsilon_t$$

where each monthly dummy variable has a value of 1 when the month occurs (e.g., $\text{Jan}_1 = \text{Jan}_{13} = 1$, as the first observation is a January) and a value of 0 for the other months. Exhibit 4 shows the results of this regression.

The intercept, b_0, measures the average return for stocks in December because there is no dummy variable for December.[30] This equation estimates that the average return in December is 2.73 percent ($\hat{b}_0 = 0.0273$). Each of the estimated coefficients for the dummy variables shows the estimated difference between returns in that month and returns for December. So, for example, the estimated additional return in January is 2.13 percent lower than December ($\hat{b}_1 = -0.0213$). This gives a January return prediction of 0.60 percent (2.73 in December – 2.13 corresponding to the January coefficient).

Exhibit 4 Results from Regressing MSCI Japan Small Cap Index Returns on Monthly Dummy Variables

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0273	0.0149	1.8322
January	–0.0213	0.0210	1.0143
February	–0.0112	0.0210	–0.5333
March	0.0101	0.0210	0.4810
April	–0.0012	0.0210	–0.0571
May	–0.0425	0.0210	–2.0238
June	–0.0065	0.0210	–0.3095
July	–0.0481	0.0210	–2.2905
August	–0.0367	0.0210	–1.7476
September	–0.0285	0.0210	–1.3571
October	–0.0429	0.0210	–2.0429
November	–0.0339	0.0210	–1.6143

(continued)

29 For a discussion of this issue, see Siegel (2014).

30 When $\text{Jan}_t = \text{Feb}_t = \ldots = \text{Nov}_t = 0$, the return is not associated with January through November so the month is December and the regression equation simplifies to $\text{Returns}_1 = b_0 + \varepsilon_t$. Because $E(\text{Returns}_1) = b_0 + E(\varepsilon_t) = b_0$, the intercept b_0 represents the mean return for December.

Exhibit 4 (Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	11	0.0551	0.0050	1.7421	0.0698
Residual	144	0.4142	0.0029		
Total	155	0.4693			
Residual standard error		0.0536			
Multiple *R*-squared		0.1174			
Observations		156			

Source: Morgan Stanley Capital International.

The low R^2 in this regression (0.1174), however, suggests that a month-of-the-year effect in small-stock returns may not be very important for explaining small-stock returns. We can use the *F*-test to analyze the null hypothesis that jointly, the monthly dummy variables all equal 0 (H_0: $b_1 = b_2 = \ldots = b_{11} = 0$). We are testing for significant monthly variation in small-stock returns. Exhibit 4 shows the data needed to perform an analysis of variance. The number of degrees of freedom in the numerator of the *F*-test is 11; the number of degrees of freedom in the denominator is [156 − (11 + 1)] = 144. The regression sum of squares equals 0.0551, and the sum of squared errors equals 0.4142. Therefore, the *F*-statistic to determine whether all of the regression slope coefficients are jointly equal to 0 is

$$\frac{0.0551/11}{0.4142/144} = 1.74$$

Appendix D (the *F*-distribution table) at the end of this volume shows the critical values for this *F*-test. If we choose a significance level of 0.05 and look in Column 11 (because the numerator has 11 degrees of freedom), we see that the critical value is 1.87 when the denominator has 120 degrees of freedom. The denominator actually has 144 degrees of freedom, so the critical value of the *F*-statistic is smaller than 1.87 (for df = 120) but larger than 1.79 (for an infinite number of degrees of freedom). The value of the test statistic is 1.74, so we cannot reject the null hypothesis that all of the coefficients jointly are equal to 0.

The *p*-value of 0.0698 shown for the *F*-test in Exhibit 4 means that the smallest level of significance at which we can reject the null hypothesis is roughly 0.07, or 7 percent—which is above the conventional level of 5 percent. Among the 11 monthly dummy variables, May, July, and October have a *t*-statistic with an absolute value greater than 2. Although the coefficients for these dummy variables are statistically significant, we have so many insignificant estimated coefficients that we cannot reject the null hypothesis that returns are equal across the months. This test suggests that the significance of a few coefficients in this regression model may be the result of random variation. We may thus want to avoid portfolio strategies calling for differing investment weights for small stocks in different months.

EXAMPLE 6

Determinants of Short-Term Stock Return Performance in Mergers and Acquisitions by Chinese Companies

Bhabra and Huang (2013) examined short-term market reaction to mergers and acquisition deals initiated by Chinese companies listed on the Shanghai Stock Exchange and the Shenzhen Stock Exchange. They examined those deals during 1997 to 2007 in which the acquirer gained complete control of the target. As the measure of short-term stock return performance around the announcement day, they used the cumulative abnormal return on the acquirer's stock during a five-day window from day –2 to day +2, where day 0 is the acquisition announcement day. Cumulative abnormal return is the excess return achieved over a stated period measured in relation to the return expected given a security's risk. The independent variables in their model included the following firm- and deal-related factors that may affect short-term stock return performance:

- profit margin: ratio of acquiring firm's net income to revenue prior to the acquisition;
- sales growth: acquiring firm's annual sales growth rate prior to the acquisition;
- change in leverage: change in the ratio of debt to total assets for the acquiring firm due to the acquisition;
- firm value: natural logarithm of the market value of the acquirer in the announcement year;
- same industry: Dummy variable (1 = the acquiring and target firms are in the same industry, 0 = in different industries);
- state-owned enterprise: Dummy variable (1 = acquiring firm is a state-owned firm, 0 = not a state-owned firm);
- cash: Dummy variable (1 = the form of payment in the transaction is cash, 0 = other forms of payment);
- cross-border: Dummy variable (1 = cross-border deal, 0 = domestic deal);
- private: Dummy variable (1 = target firm is a stand-alone private firm, 0 = not a stand-alone private firm);
- missing method of payment: Dummy variable (1 = the form of payment information is not available, 0 = form of payment information is available).

Exhibit 5 shows the authors' results.

Exhibit 5 Multiple Regression Model of Cumulative Abnormal Returns for Chinese Acquisitions, 1997–2007

	Coefficient	*p*-Value
Intercept	–0.0543	0.2316
Profit margin	0.0000	0.1522
Sales growth	–0.0180	0.2774
Change in leverage	–0.0136	0.5887
Firm value	0.0024	0.2348
Same industry	0.0012	0.4296
State-owned enterprise	0.0333	0.0435

(continued)

Exhibit 5 (Continued)

	Coefficient	*p*-Value
Cash	0.0041	0.7912
Cross-border	–0.0311	0.3376
Private	–0.0336	0.0204
Missing method of payment	0.0013	0.9399
R-squared	0.2194	
Observations	87	

Source: Bhabra and Huang (2013).

We can summarize Bhabra and Huang's findings as follows:

- The coefficient of state-owned enterprise in this regression model is positive and statistically significant at the 0.05 level as the *p*-value is less than 0.05. State-owned firms play a very important role in the Chinese economy. Non-state-owned firms are relatively new entrants in the Chinese market and are smaller firms. The statistically significant coefficient of state-owned enterprise suggests that the short-term increase in firm value is greater when the acquiring firm is state owned.
- The coefficient of private targets is also statistically significant at the 0.05 level. The sign of this coefficient is negative. Bhabra and Huang point out that the vast majority of target firms in Chinese M&As are unlisted firms, either stand-alone private firms or subsidiaries of listed firms. All the firms included in their sample are either subsidiaries or stand-alone private firms. The significantly negative coefficient of the dummy for private targets tends to result in lower cumulative abnormal returns compared to acquisitions of unlisted subsidiaries. The authors point out that a possible reason could be relatively limited data accessibility for stand-alone private firms as compared with subsidiaries of listed parents. Because of the challenges faced by acquirers when estimating the value and prospects of the private firms, acquisitions of subsidiaries elicit a more positive stock price response.
- Although none of the other coefficients are statistically significant in the above model, in some of the other models estimated in the study (not included in this reading), the authors find some evidence that the stock price response is more positive when the target is in the same industry as the acquirer and the change in leverage is low.

4 VIOLATIONS OF REGRESSION ASSUMPTIONS

In Section 2.1, we presented the assumptions of the multiple linear regression model. Inference based on an estimated regression model rests on those assumptions being satisfied. In applying regression analysis to financial data, analysts need to be able to diagnose violations of regression assumptions, understand the consequences of violations, and know the remedial steps to take. In the following sections we discuss three regression violations: **heteroskedasticity**, serial correlation, and multicollinearity.

4.1 Heteroskedasticity

So far, we have made an important assumption that the variance of error in a regression is constant across observations. In statistical terms, we assumed that the errors were homoskedastic. Errors in financial data, however, are often **heteroskedastic**: the variance of the errors differs across observations. In this section, we discuss how heteroskedasticity affects statistical analysis, how to test for heteroskedasticity, and how to correct for it.

We can see the difference between homoskedastic and heteroskedastic errors by comparing two graphs. Exhibit 6 shows the values of the dependent and independent variables and a fitted regression line for a model with homoskedastic errors. There is no systematic relationship between the value of the independent variable and the regression residuals (the vertical distance between a plotted point and the fitted regression line). Exhibit 7 shows the values of the dependent and independent variables and a fitted regression line for a model with heteroskedastic errors. Here, a systematic relationship is visually apparent: On average, the regression residuals grow much larger as the size of the independent variable increases.

4.1.1 *The Consequences of Heteroskedasticity*

What are the consequences when the assumption of constant error variance is violated? Although heteroskedasticity does not affect the consistency[31] of the regression parameter estimators, it can lead to mistakes in inference. When errors are heteroskedastic, the *F*-test for the overall significance of the regression is unreliable.[32] Furthermore, *t*-tests for the significance of individual regression coefficients are unreliable because heteroskedasticity introduces bias into estimators of the standard error of regression coefficients. If a regression shows significant heteroskedasticity, the standard errors and test statistics computed by regression programs will be incorrect unless they are adjusted for heteroskedasticity.

31 Informally, an estimator of a regression parameter is consistent if the probability that estimates of a regression parameter differ from the true value of the parameter decreases as the number of observations used in the regression increases. The regression parameter estimates from ordinary least squares are consistent regardless of whether the errors are heteroskedastic or homoskedastic. For a more advanced discussion, see Greene (2018).

32 This unreliability occurs because the mean squared error is a biased estimator of the true population variance given heteroskedasticity.

Exhibit 6 Regression with Homoskedasticity

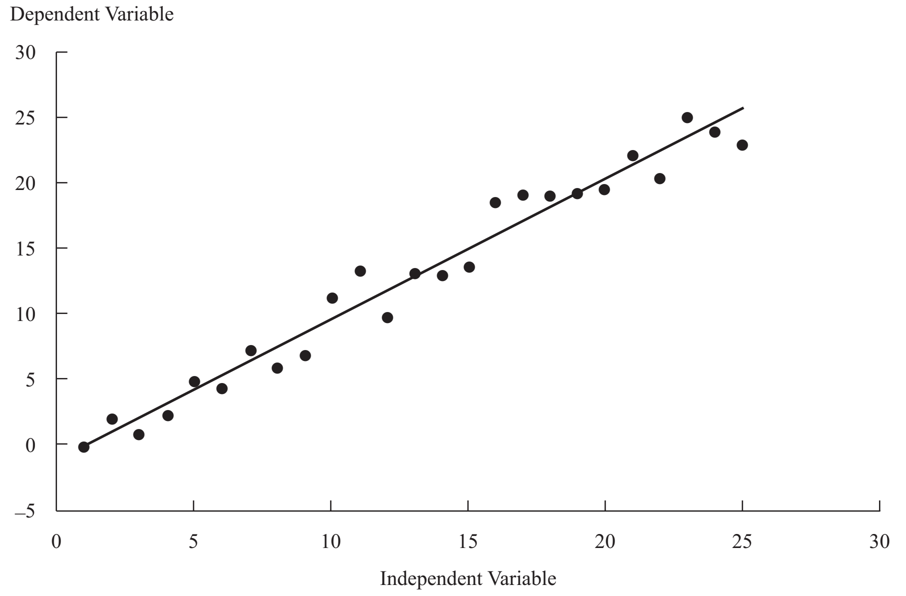
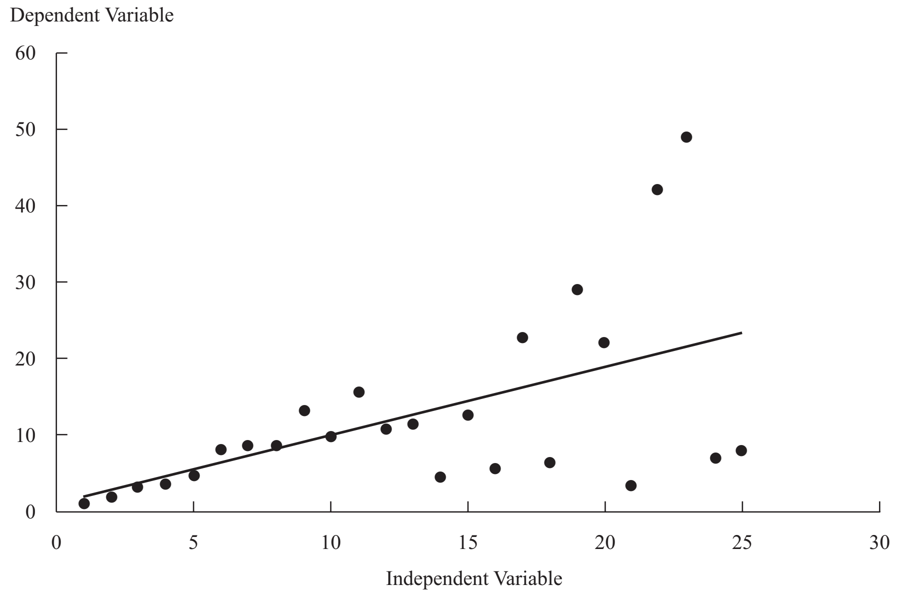


Exhibit 7 Regression with Heteroskedasticity



In regressions with financial data, the most likely result of heteroskedasticity is that the estimated standard errors will be underestimated and the *t*-statistics will be inflated. When we ignore heteroskedasticity, we tend to find significant relationships where none actually exist.[33] The consequences in practice may be serious if we are using regression analysis in the development of investment strategies. As Example 7 shows, the issue impinges even on our understanding of financial models.

EXAMPLE 7

Heteroskedasticity and Tests of an Asset Pricing Model

MacKinlay and Richardson (1991) examined how heteroskedasticity affects tests of the capital asset pricing model (CAPM). These authors argued that if the CAPM is correct, they should find no significant differences between the risk-adjusted returns for holding small stocks versus large stocks. To implement their test, MacKinlay and Richardson grouped all stocks on the New York Stock Exchange and the American Stock Exchange (now called NYSE MKT) by market-value decile with annual reassignment. They then tested for systematic differences in risk-adjusted returns across market-capitalization-based stock portfolios. They estimated the following regression:

$$r_{i,t} = \alpha_i + \beta_i r_{m,t} + \varepsilon_{i,t}$$

where

$r_{i,t}$ = excess return (return above the risk-free rate) to portfolio i in period t

$r_{m,t}$ = excess return to the market as a whole in period t

The CAPM formulation hypothesizes that excess returns on a portfolio are explained by excess returns on the market as a whole. That hypothesis implies that $\alpha_i = 0$ for every portfolio i; on average, no excess return accrues to any portfolio after taking into account its systematic (market) risk.

Using data from January 1926 to December 1988 and a market index based on equal-weighted returns, MacKinlay and Richardson failed to reject the CAPM at the 0.05 level when they assumed that the errors in the regression model are normally distributed and homoskedastic. They found, however, that they could reject the CAPM when they corrected their test statistics to account for heteroskedasticity. They rejected the hypothesis that there are no size-based, risk-adjusted excess returns in historical data.[34]

We have stated that effects of heteroskedasticity on statistical inference can be severe. To be more precise about this concept, we should distinguish between two broad kinds of heteroskedasticity: unconditional and conditional.

Unconditional heteroskedasticity occurs when heteroskedasticity of the error variance is not correlated with the independent variables in the multiple regression. Although this form of heteroskedasticity violates Assumption 4 of the linear regression model, it creates no major problems for statistical inference.

The type of heteroskedasticity that causes the most problems for statistical inference is **conditional heteroskedasticity**—heteroskedasticity in the error variance that is correlated with (conditional on) the values of the independent variables in the regression. Fortunately, many statistical software packages easily test and correct for conditional heteroskedasticity.

33 Sometimes, however, failure to adjust for heteroskedasticity results in standard errors that are too large (and t-statistics that are too small).

34 MacKinlay and Richardson also show that when using value-weighted returns, one can reject the CAPM whether or not one assumes normally distributed returns and homoskedasticity.

4.1.2 *Testing for Heteroskedasticity*

Because of conditional heteroskedasticity's consequences on inference, the analyst must be able to diagnose its presence. The Breusch–Pagan test is widely used in finance research because of its generality.[35]

Breusch and Pagan (1979) suggested the following test for conditional heteroskedasticity: Regress the squared residuals from the estimated regression equation on the independent variables in the regression. If no conditional heteroskedasticity exists, the independent variables will not explain much of the variation in the squared residuals. If conditional heteroskedasticity is present in the original regression, however, the independent variables will explain a significant portion of the variation in the squared residuals. The independent variables can explain the variation because each observation's squared residual will be correlated with the independent variables if the independent variables affect the variance of the errors.

Breusch and Pagan showed that under the null hypothesis of no conditional heteroskedasticity, nR^2 (from the regression of the squared residuals on the independent variables from the original regression) will be a χ^2 random variable with the number of degrees of freedom equal to the number of independent variables in the regression.[36] Therefore, the null hypothesis states that the regression's squared error term is uncorrelated with the independent variables. The alternative hypothesis states that the squared error term is correlated with the independent variables. Example 8 illustrates the Breusch–Pagan test for conditional heteroskedasticity.

EXAMPLE 8

Testing for Conditional Heteroskedasticity in the Relation between Interest Rates and Expected Inflation

Suppose an analyst wants to know how closely nominal interest rates are related to expected inflation to determine how to allocate assets in a fixed income portfolio. The analyst wants to test the Fisher effect, the hypothesis suggested by Irving Fisher that nominal interest rates increase by 1 percentage point for every 1 percentage point increase in expected inflation.[37] The Fisher effect assumes the following relation between nominal interest rates, real interest rates, and expected inflation:

$$i = r + \pi^e$$

where

i = the nominal rate

r = the real interest rate (assumed constant)

π^e = the expected rate of inflation

To test the Fisher effect using time-series data, we could specify the following regression model for the nominal interest rate:

$$i_t = b_0 + b_1\pi_t^e + \varepsilon_t \tag{5}$$

35 Some other tests require more-specific assumptions about the functional form of the heteroskedasticity. For more information, see Greene (2018).

36 The Breusch–Pagan test is distributed as a χ^2 random variable in large samples. The constant 1 technically associated with the intercept term in a regression is not counted here in computing the number of independent variables. For more on the Breusch–Pagan test, see Greene (2018).

37 For more on the Fisher effect, see, for example, Mankiw (2015).

Noting that the Fisher effect predicts that the coefficient on the inflation variable is 1, we can state the null and alternative hypotheses as

H_0: $b_1 = 1$

H_a: $b_1 \neq 1$

We might also specify a 0.05 significance level for the test. Before we estimate Equation 5, we must decide how to measure expected inflation (π_t^e) and the nominal interest rate (i_t).

The Survey of Professional Forecasters (SPF) has compiled data on the quarterly inflation expectations of professional forecasters.[38] We use those data as our measure of expected inflation. We use three-month Treasury bill returns as our measure of the (risk-free) nominal interest rate.[39] We use quarterly data from the fourth quarter of 1968 to the fourth quarter of 2013 to estimate Equation 5. Exhibit 8 shows the regression results.

To make the statistical decision on whether the data support the Fisher effect, we calculate the following t-statistic, which we then compare to its critical value.

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{1.1744 - 1}{0.0761} = 2.29$$

With a 0.05 significance level and 181 − 2 = 179 degrees of freedom, the critical t-value is about 1.97. If we have conducted a valid test, we can reject at the 0.05 significance level the hypothesis that the true coefficient in this regression is 1 and that the Fisher effect holds. The t-test assumes that the errors are homoskedastic. Before we accept the validity of the t-test, therefore, we should test whether the errors are conditionally heteroskedastic. If those errors prove to be conditionally heteroskedastic, then the test is invalid.

Exhibit 8 Results from Regressing T-Bill Returns on Predicted Inflation

	Coefficient	Standard Error	t-Statistic
Intercept	0.0116	0.0033	3.5152
Inflation prediction	1.1744	0.0761	15.4323
Residual standard error	0.0233		
Multiple R-squared	0.5708		
Observations	181		
Durbin–Watson statistic*	0.2980		

Note: The Durbin–Watson statistic will be explained in Section 4.2.2.
Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

We can perform the **Breusch–Pagan test** for conditional heteroskedasticity on the squared residuals from the Fisher effect regression. The test regresses the squared residuals on the predicted inflation rate. The R^2 in the squared residuals regression (not shown here) is 0.0666. The test statistic from this regression,

38 For this example, we use the annualized median SPF prediction of current-quarter growth in the GDP deflator (GNP deflator before 1992).

39 Our data on Treasury bill returns are based on three-month T-bill yields in the secondary market. Because those yields are stated on a discount basis, we convert them to a compounded annual rate so they will be measured on the same basis as our data on inflation expectations. These returns are risk-free because they are known at the beginning of the quarter and there is no default risk.

nR^2, is 181 × 0.0666 = 12.0546. Under the null hypothesis of no conditional heteroskedasticity, this test statistic is a χ^2 random variable with one degree of freedom (because there is only one independent variable).

We should be concerned about heteroskedasticity only for large values of the test statistic. Therefore, we should use a one-tailed test to determine whether we can reject the null hypothesis. The critical value of the test statistic for a variable from a χ^2 distribution with one degree of freedom at the 0.05 significance level is 3.84. The test statistic from the Breusch–Pagan test is 12.0546, so we can reject the hypothesis of no conditional heteroskedasticity at the 0.05 level. In fact, we can even reject the hypothesis of no conditional heteroskedasticity at the 0.01 significance level, because the critical value of the test statistic in the case is 6.63. As a result, we conclude that the error term in the Fisher effect regression is conditionally heteroskedastic. The standard errors computed in the original regression are not correct, because they do not account for heteroskedasticity. Therefore, we cannot accept the *t*-test as valid.

In Example 8, we concluded that a *t*-test that we might use to test the Fisher effect was not valid. Does that mean that we cannot use a regression model to investigate the Fisher effect? Fortunately, no. A methodology is available to adjust regression coefficients' standard error to correct for heteroskedasticity. Using an adjusted standard error for $\hat{b}_1$, we can reconduct the *t*-test. As we shall see in the next section, using this valid *t*-test we will not reject the null hypothesis in Example 8. That is, our statistical conclusion will change after we correct for heteroskedasticity.

4.1.3 *Correcting for Heteroskedasticity*

Financial analysts need to know how to correct for heteroskedasticity, because such a correction may reverse the conclusions about a particular hypothesis test—and thus affect a particular investment decision. In Example 7, for instance, MacKinlay and Richardson reversed their investment conclusions after correcting their model's significance tests for heteroskedasticity.

We can use two different methods to correct the effects of conditional heteroskedasticity in linear regression models. The first method, computing **robust standard errors**, corrects the standard errors of the linear regression model's estimated coefficients to account for the conditional heteroskedasticity. The second method, **generalized least squares**, modifies the original equation in an attempt to eliminate the heteroskedasticity. The new, modified regression equation is then estimated under the assumption that heteroskedasticity is no longer a problem.[40] The technical details behind these two methods of correcting for conditional heteroskedasticity are outside the scope of this reading.[41] Many statistical software packages can easily compute robust standard errors, however, and we recommend using them.[42]

Returning to the subject of Example 8 concerning the Fisher effect, recall that we concluded that the error variance was heteroskedastic. If we correct the regression coefficients' standard errors for conditional heteroskedasticity, we get the results shown in Exhibit 9. In comparing the standard errors in Exhibit 9 with those in Exhibit 8, we see that the standard error for the intercept changes very little, but the standard error for the coefficient on predicted inflation (the slope coefficient) increases by

40 Generalized least squares requires econometric expertise to implement correctly on financial data. See Greene (2018), Hansen (1982), and Keane and Runkle (1998).

41 For more details on both methods, see Greene (2018).

42 Robust standard errors are also known as **heteroskedasticity-consistent standard errors** or **White-corrected standard errors**.

about 22 percent (from 0.0761 to 0.0931). Note also that the regression coefficients are the same in both tables, because the results in Exhibit 9 correct only the standard errors in Exhibit 8.

Exhibit 9 Results from Regressing T-Bill Returns on Predicted Inflation (Standard Errors Corrected for Conditional Heteroskedasticity)

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.0116	0.0034	3.4118
Inflation prediction	1.1744	0.0931	12.6144
Residual standard error	0.0233		
Multiple *R*-squared	0.5708		
Observations	181		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

We can now conduct a valid *t*-test of the null hypothesis that the slope coefficient has a true value of 1, using the robust standard error for $\hat{b}_1$. We find that t = (1.1744 − 1)/0.0931 = 1.8733. This number is smaller than the critical value of 1.97 needed to reject the null hypothesis that the slope equals 1.[43] So, we can no longer reject the null hypothesis that the slope equals 1. Thus, in this particular example, correcting for the statistically significant conditional heteroskedasticity had an effect on the result of the hypothesis test about the slope of the predicted inflation coefficient. Example 7 concerning tests of the CAPM is a similar case. In other cases, however, our statistical decision might not change based on using robust standard errors in the *t*-test.

4.2 Serial Correlation

A more common—and potentially more serious—problem than violation of the homoskedasticity assumption is the violation of the assumption that regression errors are uncorrelated across observations. Trying to explain a particular financial relation over a number of periods is risky, because errors in financial regression models are often correlated through time.

When regression errors are correlated across observations, we say that they are **serially correlated** (or autocorrelated). Serial correlation most typically arises in time-series regressions. In this section, we discuss three aspects of serial correlation: its effect on statistical inference, tests for it, and methods to correct for it.

4.2.1 *The Consequences of Serial Correlation*

As with heteroskedasticity, the principal problem caused by serial correlation in a linear regression is an incorrect estimate of the regression coefficient standard errors computed by statistical software packages. As long as none of the independent variables is a lagged value of the dependent variable (a value of the dependent variable from a previous period), then the estimated parameters themselves will be consistent and need not be adjusted for the effects of serial correlation. If, however, one of the independent variables is a lagged value of the dependent variable—for example, if the T-bill return from the previous month was an independent variable in the Fisher

43 Remember, this is a two-tailed test.

effect regression—then serial correlation in the error term will cause all the parameter estimates from linear regression to be inconsistent and they will not be valid estimates of the true parameters.[44]

In none of the regressions examined in this reading is an independent variable a lagged value of the dependent variable. Thus, in these regressions, any effect of serial correlation appears in the regression coefficient standard errors. We will examine here the positive serial correlation case, because that case is so common. **Positive serial correlation** is serial correlation in which a positive error for one observation increases the chance of a positive error for another observation. Positive serial correlation also means that a negative error for one observation increases the chance of a negative error for another observation.[45] In examining positive serial correlation, we make the common assumption that serial correlation takes the form of **first-order serial correlation**, or serial correlation between adjacent observations. In a time-series context, that assumption means the sign of the error term tends to persist from one period to the next.

Although positive serial correlation does not affect the consistency of the estimated regression coefficients, it does affect our ability to conduct valid statistical tests. First, the *F*-statistic to test for overall significance of the regression may be inflated because the mean squared error (MSE) will tend to underestimate the population error variance. Second, positive serial correlation typically causes the ordinary least squares (OLS) standard errors for the regression coefficients to underestimate the true standard errors. As a consequence, if positive serial correlation is present in the regression, standard linear regression analysis will typically lead us to compute artificially small standard errors for the regression coefficient. These small standard errors will cause the estimated *t*-statistics to be inflated, suggesting significance where perhaps there is none. The inflated *t*-statistics may, in turn, lead us to incorrectly reject null hypotheses about population values of the parameters of the regression model more often than we would if the standard errors were correctly estimated. This Type I error could lead to improper investment recommendations.[46]

4.2.2 *Testing for Serial Correlation*

We can choose from a variety of tests for serial correlation in a regression model,[47] but the most common is based on a statistic developed by Durbin and Watson (1951); in fact, many statistical software packages compute the Durbin–Watson statistic automatically. The equation for the Durbin–Watson test statistic is

$$DW = \frac{\sum_{t=2}^{T}\left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=1}^{T}\hat{\varepsilon}_t^2} \tag{6}$$

44 We address this issue in the reading on time-series analysis.

45 In contrast, with **negative serial correlation**, a positive error for one observation increases the chance of a negative error for another observation, and a negative error for one observation increases the chance of a positive error for another.

46 OLS standard errors need not be underestimates of actual standard errors if negative serial correlation is present in the regression.

47 See Greene (2018) for a detailed discussion of tests of serial correlation.

where $\hat{\varepsilon}_t$ is the regression residual for period t. We can rewrite this equation as

$$\frac{\frac{1}{T-1}\sum_{t=2}^{T}\left(\hat{\varepsilon}_t^2 - 2\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} + \hat{\varepsilon}_{t-1}^2\right)}{\frac{1}{T-1}\sum_{t=1}^{T}\hat{\varepsilon}_t^2}$$

$$\approx \frac{Var\left(\hat{\varepsilon}_t\right) - 2\ \mathrm{Cov}\left(\hat{\varepsilon}_t, \hat{\varepsilon}_{t-1}\right) + \mathrm{Var}\left(\hat{\varepsilon}_{t-1}\right)}{\mathrm{Var}\left(\hat{\varepsilon}_t\right)}$$

If the variance of the error is constant through time, then we expect $\mathrm{Var}\left(\hat{\varepsilon}_t\right) = \hat{\sigma}_\varepsilon^2$ for all t, where we use $\hat{\sigma}_\varepsilon^2$ to represent the estimate of the constant error variance. If, in addition, the errors are also not serially correlated, then we expect $\mathrm{Cov}\left(\hat{\varepsilon}_t, \hat{\varepsilon}_{t-1}\right) = 0$. In that case, the Durbin–Watson statistic is approximately equal to

$$\frac{\hat{\sigma}_\varepsilon^2 - 0 + \hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2} = 2$$

This equation tells us that if the errors are homoskedastic and not serially correlated, then the Durbin–Watson statistic will be close to 2. Therefore, we can test the null hypothesis that the errors are not serially correlated by testing whether the Durbin–Watson statistic differs significantly from 2.

If the sample is very large, the Durbin–Watson statistic will be approximately equal to $2(1 - r)$, where r is the sample correlation between the regression residuals from one period and those from the previous period. This approximation is useful because it shows the value of the Durbin–Watson statistic for differing levels of serial correlation. The Durbin–Watson statistic can take on values ranging from 0 (in the case of serial correlation of +1) to 4 (in the case of serial correlation of −1):

- If the regression has no serial correlation, then the regression residuals will be uncorrelated through time and the value of the Durbin–Watson statistic will be equal to $2(1 - 0) = 2$.
- If the regression residuals are positively serially correlated, then the Durbin–Watson statistic will be less than 2. For example, if the serial correlation of the errors is 1, then the value of the Durbin–Watson statistic will be 0.
- If the regression residuals are negatively serially correlated, then the Durbin–Watson statistic will be greater than 2. For example, if the serial correlation of the errors is −1, then the value of the Durbin–Watson statistic will be 4.

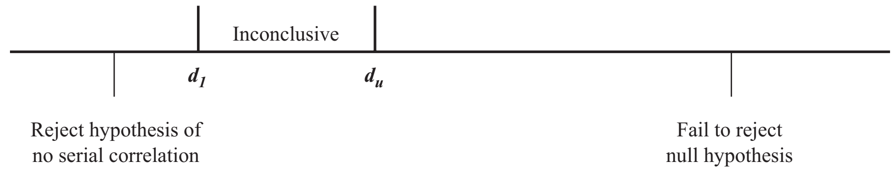
Returning to Example 8, which explored the Fisher effect, as shown in Exhibit 8 the Durbin–Watson statistic for the OLS regression is 0.2980. This result means that the regression residuals are positively serially correlated:

$$\begin{aligned} DW &= 0.2980 \\ &\approx 2\left(1 - r\right) \\ r &\approx 1 - \mathrm{DW}/2 \\ &= 1 - 0.2980/2 \\ &= 0.8510 \end{aligned}$$

This outcome raises the concern that OLS standard errors may be incorrect because of positive serial correlation. Does the observed Durbin–Watson statistic (0.2980) provide enough evidence to warrant rejecting the null hypothesis of no positive serial correlation?

We should reject the null hypothesis of no serial correlation if the Durbin–Watson statistic is below a critical value, d^*. Unfortunately, Durbin and Watson also showed that, for a given sample, we cannot know the true critical value, d^*. Instead, we can determine only that d^* lies either between two values, d_u (an upper value) and d_l (a lower value), or outside those values. Exhibit 10 depicts the upper and lower values of d^* as they relate to the results of the Durbin–Watson statistic.

Exhibit 10 Value of the Durbin–Watson Statistic



From Exhibit 10, we learn the following:

- When the Durbin–Watson (DW) statistic is less than d_l, we reject the null hypothesis of no positive serial correlation.
- When the DW statistic falls between d_l and d_u, the test results are inconclusive.
- When the DW statistic is greater than d_u, we fail to reject the null hypothesis of no positive serial correlation.[48]

Returning to Example 8, the Fisher effect regression has one independent variable and 181 observations. The Durbin–Watson statistic is 0.2980. We can reject the null hypothesis of no correlation in favor of the alternative hypothesis of positive serial correlation at the 0.05 level because the Durbin–Watson statistic is far below d_l for $k = 1$ and $n = 100$ (1.65). The level of d_l would be even higher for a sample of 181 observations. This finding of significant positive serial correlation suggests that the OLS standard errors in this regression probably significantly underestimate the true standard errors.

4.2.3 *Correcting for Serial Correlation*

We have two alternative remedial steps when a regression has significant serial correlation. First, we can adjust the coefficient standard errors for the linear regression parameter estimates to account for the serial correlation. Second, we can modify the regression equation itself to eliminate the serial correlation. We recommend using the first method for dealing with serial correlation; the second method may result in inconsistent parameter estimates unless implemented with extreme care.

Two of the most prevalent methods for adjusting standard errors were developed by Hansen (1982) and Newey and West (1987). These methods are standard features in many statistical software packages.[49] An additional advantage of these methods is that they simultaneously correct for conditional heteroskedasticity.[50]

48 Of course, sometimes serial correlation in a regression model is negative rather than positive. For a null hypothesis of no serial correlation, the null hypothesis is rejected if DW < d_l (indicating significant positive serial correlation) or if DW > 4 – d_l (indicating significant negative serial correlation).

49 This correction is known by various names, including serial-correlation consistent standard errors, serial correlation and heteroskedasticity adjusted standard errors, and robust standard errors. The Hansen standard errors are also known as Hansen–White standard errors.

50 We do not always use Hansen's method or Newey–West method to correct for serial correlation and heteroskedasticity because sometimes the errors of a regression are not serially correlated.

Exhibit 11 shows the results of correcting the standard errors from Exhibit 8 for serial correlation and heteroskedasticity using the Newey–West method. Note that the coefficients for both the intercept and the slope are exactly the same as in the original regression. The robust standard errors are now much larger, however—more than twice the OLS standard errors in Exhibit 8. Because of the severe serial correlation in the regression error, OLS greatly underestimates the uncertainty about the estimated parameters in the regression.

Note also that the serial correlation has not been eliminated, but the standard error has been corrected to account for the serial correlation.

Exhibit 11 Results from Regressing T-Bill Returns on Predicted Inflation (Standard Errors Corrected for Conditional Heteroskedasticity and Serial Correlation)

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0116	0.0067	1.7313
Inflation prediction	1.1744	0.1751	6.7070
Residual standard error	0.0233		
Multiple *R*-squared	0.5708		
Observations	181		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

Now suppose we want to test our original null hypothesis (the Fisher effect) that the coefficient on the predicted inflation term equals 1 (H_0: $b_1 = 1$) against the alternative that the coefficient on the inflation term is not equal to 1 (H_a: $b_1 \neq 1$). With the corrected standard errors, the value of the test statistic for this null hypothesis is

$$\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{1.1744 - 1}{0.1751} = 0.996$$

The critical values for both the 0.05 and 0.01 significance level are much larger than 0.996 (the *t*-test statistic), so we cannot reject the null hypothesis. This conclusion is the same as that reached in Example 7 where the correction was only for heteroskedasticity; but it is the opposite of the conclusion in Example 6 where there were no correlations.

This shows that for some hypotheses, serial correlation and conditional heteroskedasticity could have a big effect on whether we accept or reject those hypotheses.[51]

4.3 Multicollinearity

The second assumption of the multiple linear regression model is that no exact linear relationship exists between two or more of the independent variables. When one of the independent variables is an exact linear combination of other independent variables, it becomes mechanically impossible to estimate the regression. That case, known as perfect collinearity, is much less of a practical concern than

51 Serial correlation can also affect forecast accuracy.

multicollinearity.[52] **Multicollinearity** occurs when two or more independent variables (or combinations of independent variables) are highly (but not perfectly) correlated with each other. With multicollinearity we can estimate the regression, but the interpretation of the regression output becomes problematic. Multicollinearity is a serious practical concern because approximate linear relationships among financial variables are common.

4.3.1 *The Consequences of Multicollinearity*

Although the presence of multicollinearity does not affect the consistency of the OLS estimates of the regression coefficients, the estimates become extremely imprecise and unreliable. Furthermore, it becomes practically impossible to distinguish the individual impacts of the independent variables on the dependent variable. These consequences are reflected in inflated OLS standard errors for the regression coefficients. With inflated standard errors, t-tests on the coefficients have little power (ability to reject the null hypothesis).

4.3.2 *Detecting Multicollinearity*

In contrast to the cases of heteroskedasticity and serial correlation, we shall not provide a formal statistical test for multicollinearity. In practice, multicollinearity is often a matter of degree rather than of absence or presence.[53]

The analyst should be aware that using the magnitude of pairwise correlations among the independent variables to assess multicollinearity, as has occasionally been suggested, is generally not adequate. Although very high pairwise correlations among independent variables can indicate multicollinearity, it is not necessary for such pairwise correlations to be high for there to be a problem of multicollinearity.[54] Stated another way, high pairwise correlations among the independent variables are not a necessary condition for multicollinearity, and low pairwise correlations do not mean that multicollinearity is not a problem. The only case in which correlation between independent variables may be a reasonable indicator of multicollinearity occurs in a regression with exactly two independent variables.

The classic symptom of multicollinearity is a high R^2 (and significant F-statistic) even though the t-statistics on the estimated slope coefficients are not significant. The insignificant t-statistics reflect inflated standard errors. Although the coefficients might be estimated with great imprecision, as reflected in low t-statistics, the independent variables *as a group* may do a good job of explaining the dependent variable, and a high R^2 would reflect this effectiveness. Example 9 illustrates this diagnostic.

EXAMPLE 9

Multicollinearity in Explaining Returns to the Fidelity Select Technology Portfolio

In Example 3 we regressed returns to the Fidelity Select Technology Portfolio (FSPTX) on returns to the S&P 500 Growth Index and the S&P 500 Value Index. Exhibit 12 shows the results of our regression, which uses data from January 2009 through December 2013. The t-statistic of 5.9286 on the growth index return is

52 To give an example of perfect collinearity, suppose we tried to explain a company's credit ratings with a regression that included net sales, cost of goods sold, and gross profit as independent variables. Because Gross profit = Net sales – Cost of goods sold by definition, there is an exact linear relationship between these variables. This type of blunder is relatively obvious (and easy to avoid).

53 See Kmenta (1986).

54 Even if pairs of independent variables have low correlation, there may be linear combinations of the independent variables that are very highly correlated, creating a multicollinearity problem.

greater than 2, indicating that the coefficient on the growth index differs significantly from 0 at standard significance levels. On the other hand, the t-statistic on the value index return is –0.9012 and thus is not statistically significant. This result suggests that the returns to the FSPTX are linked to the returns to the growth index and not closely associated with the returns to the value index. The coefficient on the growth index, however, is 1.4697. This result implies that returns on the FSPTX are more volatile than are returns on the growth index.

Exhibit 12 Results from Regressing the FSPTX Returns on the S&P 500 Growth and Value Indices

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0018	0.0038	0.4737
S&P 500 Growth Index	1.4697	0.2479	5.9286
S&P 500 Value Index	–0.1833	0.2034	–0.9012

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	0.1653	0.0826	113.7285	1.27E-20
Residual	57	0.0414	0.0007		
Total	59	0.2067			
Residual standard error			0.0270		
Multiple R-squared			0.7996		
Observations			60		

Source: Bloomberg, finance.yahoo.com.

Note also that this regression explains a significant amount of the variation in the returns to the FSPTX. Specifically, the R^2 from this regression is 0.7996. Thus approximately 80 percent of the variation in the returns to the FSPTX is explained by returns to the S&P 500 Growth and S&P 500 Value indices.

Now suppose we run another linear regression that adds returns to the S&P 500 itself to the returns to the S&P 500 Growth and S&P 500 Value indices. The S&P 500 includes the component stocks of these two style indices, so we are introducing a severe multicollinearity problem.

Exhibit 13 shows the results of that regression. Note that the R^2 in this regression has changed almost imperceptibly from the R^2 in the previous regression (increasing from 0.7996 to 0.8084), but now the standard errors of the coefficients of the independent variables are much larger. Adding the return to the S&P 500 to the previous regression does not explain any more of the variance in the returns to the FSPTX than the previous regression did, but now none of the coefficients is statistically significant. This is the classic case of multicollinearity mentioned in the reading.

Exhibit 13 Results from Regressing the FSPTX Returns on Returns to the S&P 500 Growth and S&P 500 Value Indices and the S&P 500 Index

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0008	0.0038	0.2105
S&P 500 Growth Index	14.2444	7.9783	1.7854
S&P 500 Value Index	11.6955	7.4180	1.5766
S&P 500 Index	–24.6734	15.4022	–1.6019

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	3	0.1671	0.0557	78.7577	4.14E-20
Residual	56	0.0396	0.0007		
Total	59	0.2067			
Residual standard error			0.0266		
Multiple R-squared			0.8084		
Observations			60		

Source: Bloomberg, finance.yahoo.com, S&P Dow Jones Indices.

Multicollinearity may be a problem even when we do not observe the classic symptom of insignificant t-statistics but a highly significant F-test. Advanced textbooks provide further tools to help diagnose multicollinearity.[55]

4.3.3 *Correcting for Multicollinearity*

The most direct solution to multicollinearity is excluding one or more of the regression variables. In the example above, we can see that the S&P 500 total returns should not be included if both the S&P 500 Growth and S&P 500 Value indices are included, because the returns to the entire S&P 500 Index are a weighted average of the return to growth stocks and value stocks. In many cases, however, no easy solution is available to the problem of multicollinearity, and you will need to experiment with including or excluding different independent variables to determine the source of multicollinearity.

4.4 Heteroskedasticity, Serial Correlation, Multicollinearity: Summarizing the Issues

We have discussed some of the problems that heteroskedasticity, serial correlation, and multicollinearity may cause in interpreting regression results. These violations of regression assumptions, we have noted, all lead to problems in making valid inferences. The analyst should check that model assumptions are fulfilled before interpreting statistical tests.

Exhibit 14 gives a summary of these problems, the effect they have on the linear regression results (an analyst can see these effects using regression software), and the solutions to these problems.

55 See Greene (2018).

Exhibit 14 Problems in Linear Regression and Their Solutions

Problem	Effect	Solution
Heteroskedasticity	Incorrect standard errors	Use robust standard errors (corrected for conditional heteroskedasticity)
Serial correlation	Incorrect standard errors (additional problems if a lagged value of the dependent variable is used as an independent variable)	Use robust standard errors (corrected for serial correlation)
Multicollinearity	High R^2 and low t-statistics	Remove one or more independent variables; often no solution based in theory

MODEL SPECIFICATION AND ERRORS IN SPECIFICATION

5

Until now, we have assumed that whatever regression model we estimate is correctly specified. **Model specification** refers to the set of variables included in the regression and the regression equation's functional form. In the following, we first give some broad guidelines for correctly specifying a regression. Then we turn to three types of model misspecification: misspecified functional form, regressors that are correlated with the error term, and additional time-series misspecification. Each of these types of misspecification invalidates statistical inference using OLS; most of these misspecifications will cause the estimated regression coefficients to be inconsistent.

5.1 Principles of Model Specification

In discussing the principles of model specification, we need to acknowledge that there are competing philosophies about how to approach model specification. Furthermore, our purpose for using regression analysis may affect the specification we choose. The following principles have fairly broad application, however.

- *The model should be grounded in cogent economic reasoning.* We should be able to supply the economic reasoning behind the choice of variables, and the reasoning should make sense. When this condition is fulfilled, we increase the chance that the model will have predictive value with new data. This approach contrasts to the variable-selection process known as **data mining**. With data mining, the investigator essentially develops a model that maximally exploits the characteristics of a specific dataset. "Data mining" is used in the different sense of discovering patterns in large datasets in contexts discussed later in Section 7.
- *The functional form chosen for the variables in the regression should be appropriate given the nature of the variables.* As one illustration, consider studying mutual fund **market timing** based on fund and market returns alone. One might reason that for a successful timer, a plot of mutual fund returns against market returns would show curvature, because a successful timer would tend to

increase (decrease) beta when market returns were high (low). The model specification should reflect the expected nonlinear relationship.[56] In other cases, we may transform the data such that a regression assumption is better satisfied.

- *The model should be parsimonious.* In this context, "parsimonious" means accomplishing a lot with a little. We should expect each variable included in a regression to play an essential role.
- *The model should be examined for violations of regression assumptions before being accepted.* We have already discussed detecting the presence of heteroskedasticity, serial correlation, and multicollinearity. As a result of such diagnostics, we may conclude that we need to revise the set of included variables and/or their functional form.
- *The model should be tested and be found useful out of sample before being accepted.* The term "out of sample" refers to observations outside the dataset on which the model was estimated. A plausible model may not perform well out of sample because economic relationships have changed since the sample period. That possibility is itself useful to know. A second explanation, however, may be that relationships have not changed but that the model explains only a specific dataset.

Having given some broad guidance on model specification, we turn to a discussion of specific model specification errors. Understanding these errors will help an analyst develop better models and be a more informed consumer of investment research.

5.2 Misspecified Functional Form

Whenever we estimate a regression, we must assume that the regression has the correct functional form. This assumption can fail in several ways:

- One or more important variables could be omitted from regression.
- One or more of the regression variables may need to be transformed (for example, by taking the natural logarithm of the variable) before estimating the regression.
- The regression model pools data from different samples that should not be pooled.

First, consider the effects of omitting an important independent variable from a regression (omitted variable bias). If the true regression model was

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \varepsilon_i \quad (7)$$

but we estimate the model[57]

$$Y_i = a_0 + a_1X_{1i} + \varepsilon_i$$

then our regression model would be misspecified. What is wrong with the model?

If the omitted variable (X_2) is correlated with the remaining variable (X_1), then the error term in the model will be correlated with (X_1), and the estimated values of the regression coefficients a_0 and a_1 would be biased and inconsistent. In addition,

56 This example is based on Treynor and Mazuy (1966), an early regression study of mutual fund timing. To capture curvature, they included a term in the squared market excess return, which does not violate the assumption of the multiple linear regression model that relationship between the dependent and independent variables is linear *in the coefficients*.

57 We use a different regression coefficient notation when X_{2i} is omitted, because the intercept term and slope coefficient on X_{1i} will generally not be the same as when X_{2i} is included.

the estimates of the standard errors of those coefficients will also be inconsistent, so we can use neither the coefficients estimates nor the estimated standard errors to make statistical tests.

EXAMPLE 10

Omitted Variable Bias and the Bid–Ask Spread

In this example, we extend our examination of the bid–ask spread to show the effect of omitting an important variable from a regression. In Example 1, we showed that the natural logarithm of the ratio [(Bid–ask spread)/Price] was significantly related to both the natural logarithm of the number of market makers and the natural logarithm of the market capitalization of the company. We repeat Exhibit 1 from Example 1 below.

Exhibit 1 Results from Regressing ln(Bid–Ask Spread/Price) on ln(Number of Market Makers) and ln(Market Capitalization) (repeated)

	Coefficients	Standard Error	*t*-Statistic
Intercept	1.5949	0.2275	7.0105
ln(Number of NASDAQ market makers)	−1.5186	0.0808	−18.7946
ln(Company's market capitalization)	−0.3790	0.0151	−25.0993

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	3,728.1334	1,864.0667	2,216.7505	0.00
Residual	2,584	2,172.8870	0.8409		
Total	2,586	5,901.0204			
Residual standard error			0.9170		
Multiple *R*-squared			0.6318		
Observations			2,587		

Source: Center for Research in Security Prices, University of Chicago.

If we did not include the natural log of market capitalization as an independent variable in the regression, and we regressed the natural logarithm of the ratio [(Bid–ask spread)/Price] only on the natural logarithm of the number of market makers for the stock, the results would be as shown in Exhibit 15.

Exhibit 15 Results from Regressing ln(Bid–Ask Spread/Price) on ln(Number of Market Makers)

	Coefficients	Standard Error	*t*-Statistic
Intercept	5.0707	0.2009	25.2399
ln(Number of NASDAQ market makers)	−3.1027	0.0561	−55.3066

(continued)

Exhibit 15 (Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	3,200.3918	3,200.3918	3,063.3655	0.00
Residual	2,585	2,700.6287	1.0447		
Total	2,586	5,901.0204			
Residual standard error			1.0221		
Multiple R-squared			0.5423		
Observations			2,587		

Source: Center for Research in Security Prices, University of Chicago.

Note that the coefficient on ln(Number of NASDAQ market makers) changed from −1.5186 in the original (correctly specified) regression to −3.1027 in the misspecified regression. Also, the intercept changed from 1.5949 in the correctly specified regression to 5.0707 in the misspecified regression. These results illustrate that omitting an independent variable that should be in the regression can cause the remaining regression coefficients to be inconsistent.

A second common cause of misspecification in regression models is the use of the wrong form of the data in a regression, when a transformed version of the data is appropriate. For example, sometimes analysts fail to account for curvature or nonlinearity in the relationship between the dependent variable and one or more of the independent variables, instead specifying a linear relation among variables. When we are specifying a regression model, we should consider whether economic theory suggests a nonlinear relation. We can often confirm the nonlinearity by plotting the data, as we will illustrate in Example 11 below. If the relationship between the variables becomes linear when one or more of the variables is represented as a proportional change in the variable, we may be able to correct the misspecification by taking the natural logarithm of the variable(s) we want to represent as a proportional change. Other times, analysts use unscaled data in regressions, when scaled data (such as dividing net income or cash flow by sales) are more appropriate. In Example 1, we scaled the bid–ask spread by stock price because what a given bid–ask spread means in terms of transactions costs for a given size investment depends on the price of the stock; if we had not scaled the bid–ask spread, the regression would have been misspecified.

EXAMPLE 11

Nonlinearity and the Bid–Ask Spread

In Example 1, we showed that the natural logarithm of the ratio [(Bid–ask spread)/Price] was significantly related to both the natural logarithm of the number of market makers and the natural logarithm of the company's market capitalization. But why did we take the natural logarithm of each of the variables in the regression? We began a discussion of this question in Example 1, which we continue now.

What does theory suggest about the nature of the relationship between the ratio (Bid–ask spread)/Price, or the percentage bid–ask spread, and its determinants (the independent variables)? Stoll (1978) builds a theoretical model of the determinants of percentage bid–ask spread in a dealer market. In his model, the determinants enter multiplicatively in a particular fashion. In terms of the independent variables introduced in Example 1, the functional form assumed is

$$\left[(\text{Bid–ask spread})/\text{Price}\right]_i = c(\text{Number of market makers})_i^{b_1} \times (\text{Market capitalization})_i^{b_2}$$

where c is a constant. The relationship of the percentage bid–ask spread with the number of market makers and market capitalization is not linear in the original variables.[58] If we take the natural log of both sides of the above model, however, we have a log-log regression that is linear in the transformed variables:[59]

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \varepsilon_i$$

where

Y_i = the natural logarithm of the ratio (Bid–ask spread)/Price for stock i
b_0 = a constant that equals ln(c)
X_{1i} = the natural logarithm of the number of market makers for stock i
X_{2i} = the natural logarithm of the market capitalization of company i
ε_i = the error term

As mentioned in Example 1, a slope coefficient in the log-log model is interpreted as an elasticity, precisely, the partial elasticity of the dependent variable with respect to the independent variable ("partial" means holding the other independent variables constant).

We can plot the data to assess whether the variables are linearly related after the logarithmic transformation. For example Exhibit 16 shows a scatterplot of the natural logarithm of the number of market makers for a stock (on the X axis) and the natural logarithm of (Bid–ask spread)/Price (on the Y axis), as well as a regression line showing the linear relation between the two transformed variables. The relation between the two transformed variables is clearly linear.

58 The form of the model is analogous to the Cobb–Douglas production function in economics.
59 We have added an error term to the model.

Exhibit 16 Linear Regression When Two Variables Have a Linear Relation

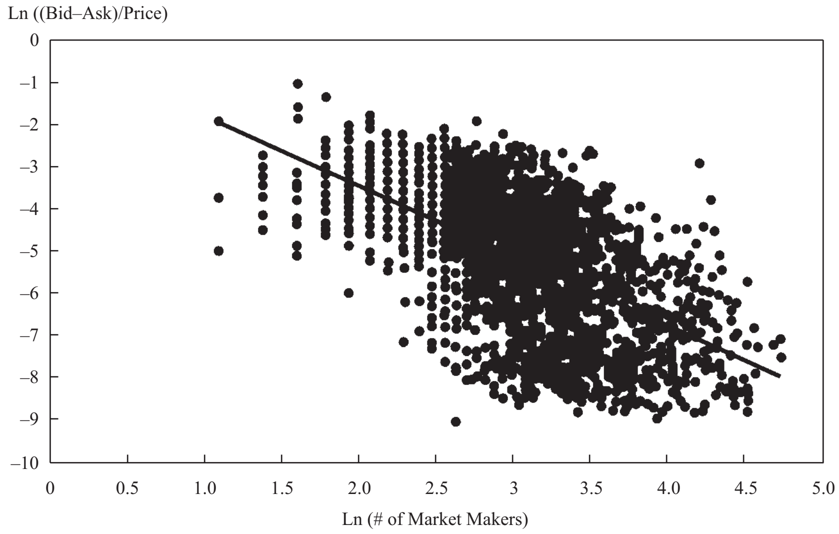
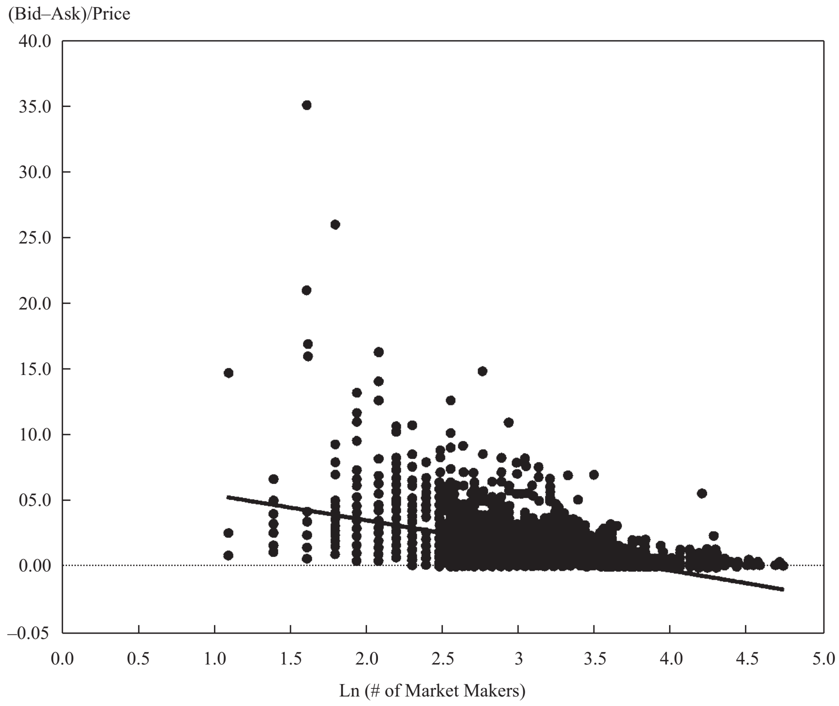


Exhibit 17 Linear Regression When Two Variables Have a Nonlinear Relation



If we do not take log of the ratio (Bid–ask spread)/Price, the plot is not linear. Exhibit 17 shows a plot of the natural logarithm of the number of market makers for a stock (on the *X* axis) and the ratio (Bid–ask spread)/Price expressed as a percentage (on the *Y* axis), as well as a regression line that attempts to show a linear relation between the two variables. We see that the relation between the two variables is very nonlinear.[60] Consequently, we should not estimate a regression with (Bid–ask spread)/Price as the dependent variable. Consideration of the need to ensure that predicted bid–ask spreads are positive would also lead us to not use (Bid–ask spread)/Price as the dependent variable. If we use the non-transformed ratio (Bid–ask spread)/Price as the dependent variable, the estimated model could predict negative values of the bid–ask spread. This result would be nonsensical; in reality, no bid–ask spread is negative (it is hard to motivate traders to simultaneously buy high and sell low), so a model that predicts negative bid–ask spreads is certainly misspecified.[61] We illustrate the problem of negative values of the predicted bid–ask spreads now.

Exhibit 18 shows the results of a regression with (Bid–ask spread)/ Price as the dependent variable and the natural logarithm of the number of market makers and the natural logarithm of the company's market capitalization as the independent variables.

Exhibit 18 Results from Regressing Bid–Ask Spread/Price on ln(Number of Market Makers) and ln(Market Cap)

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.0674	0.0035	19.2571
ln(Number of NASDAQ market makers)	−0.0142	0.0012	−11.8333
ln(Company's market cap)	−0.0016	0.0002	−8.0000

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	2	0.1539	0.0770	392.3338	0.00
Residual	2,584	0.5068	0.0002		
Total	2,586	0.6607			
Residual standard error			0.0140		
Multiple R-squared			0.2329		
Observations			2,587		

Source: Center for Research in Security Prices, University of Chicago.

1 Suppose that for a particular NASDAQ-listed stock, the number of market makers is 50 and the market capitalization is $6 billion. What is the predicted ratio of bid–ask spread to price for this stock based on the above model?

Solution to 1:

The natural log of the number of market makers equals ln 50 = 3.9120 and the natural log of the stock's market capitalization (in millions) is ln 6,000 = 8.6995. In this case, the predicted ratio of bid–ask spread to price is 0.0674 + (−0.0142 ×

60 The relation between (Bid–ask spread)/Price and ln(Market cap) is also nonlinear, while the relation between ln(Bid–ask spread)/Price and ln(Market cap) is linear. We omit these scatterplots to save space.

61 In our data sample, the bid–ask spread for each of the 2,587 companies is positive.

3.9120) + (−0.0016 × 8.6995) = −0.0021. Therefore, the model predicts that the ratio of bid–ask spread to stock price is −0.0021 or −0.21 percent of the stock price.

2 Does the predicted bid–ask spread for the above stock make sense? If not, how could this problem be avoided?

Solution to 2:

The predicted bid–ask spread is negative, which does not make economic sense. This problem could be avoided by using log of (Bid–ask spread)/Price as the dependent variable.[62]

Often, analysts must decide whether to scale variables before they compare data across companies. For example, in financial statement analysis, analysts often compare companies using **common size statements**. In a common size income statement, all the line items in a company's income statement are divided by the company's revenues. Common size statements make comparability across companies much easier. An analyst can use common size statements to quickly compare trends in gross margins (or other income statement variables) for a group of companies.

Issues of comparability also appear for analysts who want to use regression analysis to compare the performance of a group of companies. Example 12 illustrates this issue.

EXAMPLE 12

Scaling and the Relation between Cash Flow from Operations and Free Cash Flow

Suppose an analyst wants to explain free cash flow to the firm as a function of cash flow from operations in 2001 for 11 family clothing stores in the United States with market capitalizations of more than \$100 million as of the end of 2001.

To investigate this issue, the analyst might use free cash flow as the dependent variable and cash flow from operations as the independent variable in single-independent-variable linear regression. Exhibit 19 shows the results of that regression. Note that the t-statistic for the slope coefficient for cash flow from operations is quite high (6.5288), the significance level for the F-statistic for the regression is very low (0.0001), and the R-squared is quite high. We might be tempted to believe that this regression is a success and that for a family clothing store, if cash flow from operations increased by \$1.00, we could confidently predict that free cash flow to the firm would increase by \$0.3579.

Exhibit 19 Results from Regressing the Free Cash Flow on Cash Flow from Operations for Family Clothing Stores

	Coefficients	Standard Error	*t*-Statistic
Intercept	0.7295	27.7302	0.0263
Cash flow from operations	0.3579	0.0548	6.5288

62 Whether the natural log of the percentage bid–ask spread, Y, is positive or negative, the percentage bid–ask spread found as e^Y is positive, because a positive number raised to any power is positive. The constant e is positive ($e \approx 2.7183$).

Exhibit 19 (Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	245,093.7836	245,093.7836	42.6247	0.0001
Residual	9	51,750.3139	5,750.0349		
Total	10	296,844.0975			
Residual standard error			75.8290		
Multiple *R*-squared			0.8257		
Observations			11		

Source: Compustat.

But is this specification correct? The regression does not account for size differences among the companies in the sample.

We can account for size differences by using common size cash flow results across companies. We scale the variables by dividing cash flow from operations and free cash flow to the firm by the company's sales before using regression analysis. We will use (Free cash flow to the firm/Sales) as the dependent variable and (Cash flow from operations/Sales) as the independent variable. Exhibit 20 shows the results of this regression. Note that the t-statistic for the slope coefficient on (Cash flow from operations/Sales) is 1.6262, so it is not significant at the 0.05 level. Note also that the significance level of the F-statistic is 0.1383, so we cannot reject at the 0.05 level the hypothesis that the regression does not explain variation in (Free cash flow/Sales) among family clothing stores. Finally, note that the R-squared in this regression is much lower than that of the previous regression.

Exhibit 20 Results from Regressing the Free Cash Flow/Sales on Cash Flow from Operations/Sales for Family Clothing Stores

	Coefficient	Standard Error	*t*-Statistic
Intercept	−0.0121	0.0221	−0.5497
Cash flow from operations/Sales	0.4749	0.2920	1.6262

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	0.0030	0.0030	2.6447	0.1383
Residual	9	0.0102	0.0011		
Total	10	0.0131			
Residual standard error		0.0336			
Multiple *R*-squared		0.2271			
Observations		11			

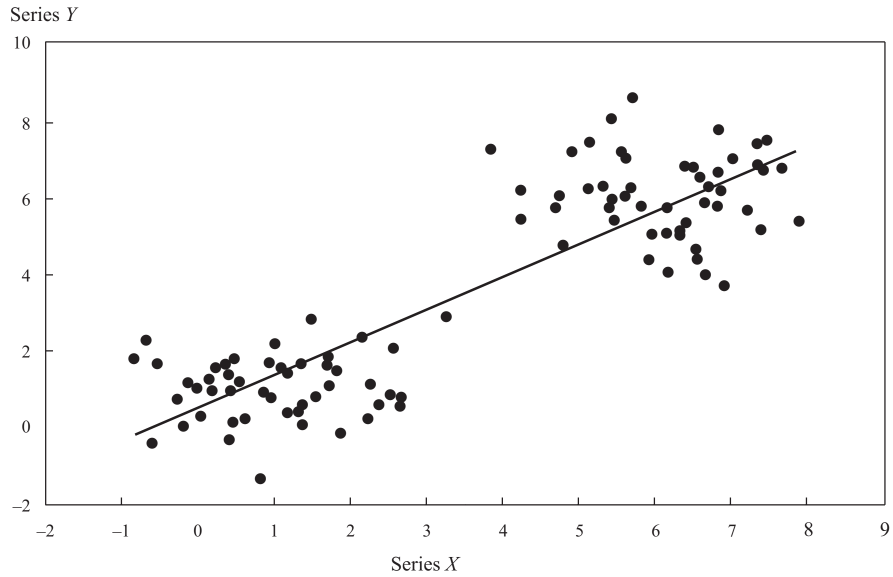
Source: Compustat.

Which regression makes more sense? Usually, the scaled regression makes more sense. We want to know what happens to free cash flow (as a fraction of sales) if a change occurs in cash flow from operations (as a fraction of sales).

Without scaling, the results of the regression can be based solely on scale differences across companies, rather than based on the companies' underlying economics.

A third common form of misspecification in regression models is pooling data from different samples that should not be pooled. This type of misspecification can best be illustrated graphically. Exhibit 21 shows two clusters of data on variables X and Y, with a fitted regression line. The data could represent the relationship between two financial variables at two different time periods, for example.

Exhibit 21 Plot of Two Series with Changing Means



In each cluster of data on X and Y, the correlation between the two variables is virtually 0. Because the means of both X and Y are different for the two clusters of data in the combined sample, X and Y are highly correlated. The correlation is spurious (misleading), however, because it reflects differences in the relationship between X and Y during two different time periods.

5.3 Time-Series Misspecification (Independent Variables Correlated with Errors)

In the previous section, we discussed the misspecification that arises when a relevant independent variable is omitted from a regression. In this section, we discuss problems that arise from the kinds of variables included in the regression, particularly in a time-series context. In models that use time-series data to explain the relations among different variables, it is particularly easy to violate Regression Assumption 3, that the error term has mean 0, conditioned on the independent variables. If this assumption is violated, the estimated regression coefficients will be biased and inconsistent.

Three common problems that create this type of time-series misspecification are:

- including lagged dependent variables as independent variables in regressions with serially correlated errors;

- including a function of a dependent variable as an independent variable, sometimes as a result of the incorrect dating of variables; and
- independent variables that are measured with error.

The next examples demonstrate these problems.

Suppose that an analyst includes the first lagged value of the dependent variable in a multiple regression that, as a result, has significant serial correlation in the errors. For example, the analyst might use the regression equation

$$Y_t = b_0 + b_1X_{1t} + b_2Y_{t-1} + \varepsilon_t \tag{8}$$

Because we assume that the error term is serially correlated, by definition the error term is correlated with the dependent variable. Consequently, the lagged dependent variable, Y_{t-1}, will be correlated with the error term, violating the assumption that the independent variables are uncorrelated with the error term. As a result, the estimates of the regression coefficients will be biased and inconsistent.

EXAMPLE 13

Fisher Effect with a Lagged Dependent Variable

In our discussion of serial correlation, we concluded from a test using the Durbin–Watson test that the error term in the Fisher effect equation (Equation 5) showed positive (first-order) serial correlation, using three-month T-bill returns as the dependent variable and inflation expectations of professional forecasters as the independent variable. Observations on the dependent and independent variables were quarterly. Exhibit 22 modifies that regression by including the previous quarter's three-month T-bill returns as an additional independent variable.

Exhibit 22 Results from Regressing T-Bill Returns on Predicted Inflation and Lagged T-Bill Returns

	Coefficient	Standard Error	*t*-Statistic
Intercept	–0.0005	0.0014	–0.3571
Inflation prediction	0.1843	0.0455	4.0505
Lagged T-bill return	0.8796	0.0295	29.8169
Residual standard error	0.0095		
Multiple R-squared	0.9285		
Observations	181		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

At first glance, these regression results look very interesting—the coefficient on the lagged T-bill return appears to be highly significant. But on closer consideration, we must ignore these regression results, because the regression is fundamentally misspecified. As long as the error term is serially correlated, including lagged T-bill returns as an independent variable in the regression will cause all the coefficient estimates to be biased and inconsistent. Therefore, this regression is not usable for either testing a hypothesis or for forecasting.

A second common time-series misspecification in investment analysis is to forecast the past. What does that mean? If we forecast the future (say we predict at time t the value of variable Y in period $t + 1$), we must base our predictions on information we knew at time t. We could use a regression to make that forecast using the equation

$$Y_{t+1} = b_0 + b_1X_{1t} + \varepsilon_{t+1} \tag{9}$$

In this equation, we predict the value of Y in time $t + 1$ using the value of X in time t. The error term, ε_{t+1}, is unknown at time t and thus should be uncorrelated with X_{1t}.

Unfortunately, analysts sometimes use regressions that try to forecast the value of a dependent variable at time $t + 1$ based on independent variable(s) that are functions of the value of the dependent variable at time $t + 1$. In such a model, the independent variable(s) would be correlated with the error term, so the equation would be misspecified. As an example, an analyst may try to explain the cross-sectional returns for a group of companies during a particular year using the market-to-book ratio and the market capitalization for those companies at the end of the year.[63] If the analyst believes that such a regression predicts whether companies with high market-to-book ratios or high market capitalizations will have high returns, the analyst is mistaken. This is because for any given period, the higher the return during the period, the higher the market capitalization and the market-to-book period will be at the end of the period. So in this case, if all the cross-sectional data come from period $t + 1$, a high value of the dependent variable (returns) actually causes a high value of the independent variables (market capitalization and the market-to-book ratio), rather than the other way around. In this type of misspecification, the regression model effectively includes the dependent variable on both the right-and left-hand sides of the regression equation.

The third common time-series misspecification arises when an independent variable is measured with error. Suppose a financial theory tells us that a particular variable X_t, such as expected inflation, should be included in the regression model. But we cannot directly observe X_t; instead, we can observe actual inflation, $Z_t = X_t + u_t$, where we assume u_t is an error term that is uncorrelated with X_t. Even in this best of circumstances, using Z_t in the regression instead of X_t will cause the regression coefficient estimates to be biased and inconsistent. To see why, assume we want to estimate the regression

$$Y_t = b_0 + b_1X_t + \varepsilon_t$$

but we substitute Z_t for X_t. Then we would estimate

$$Y_t = b_0 + b_1Z_t + (-b_1u_t + \varepsilon_t)$$

But $Z_t = X_t + u_t$, Z_t is correlated with the error term $(-b_1u_t + \varepsilon_t)$. Therefore, our estimated model violates the assumption that the error term is uncorrelated with the independent variable. Consequently, the estimated regression coefficients will be biased and inconsistent.

EXAMPLE 14

The Fisher Effect with Measurement Error

Recall from Example 8 on the Fisher effect that based on our initial analysis in which we did not correct for heteroskedasticity and serial correlation, we rejected the hypothesis that three-month T-bill returns moved one-for-one with expected inflation.

63 "Market-to-book ratio" is the ratio of price per share divided by book value per share.

Exhibit 8 Results from Regressing T-Bill Returns on Predicted Inflation (repeated)

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0116	0.0033	3.5152
Inflation prediction	1.1744	0.0761	15.4323
Residual standard error	0.0223		
Multiple R-squared	0.5708		
Observations	181		
Durbin–Watson statistic	0.2980		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

What if we used actual inflation instead of expected inflation as the independent variable? Note first that

$$\pi = \pi^e + \nu$$

where

π = actual rate of inflation

π^e = expected rate of inflation

ν = the difference between actual and expected inflation

Because actual inflation measures expected inflation with error, the estimators of the regression coefficients using T-bill yields as the dependent variable and actual inflation as the independent variable will not be consistent.[64]

Exhibit 23 shows the results of using actual inflation as the independent variable. The estimates in this exhibit are quite different from those presented in the previous exhibit. Note that the slope coefficient on actual inflation is much lower than the slope coefficient on predicted inflation in the previous regression. This result is an illustration of a general proposition: In a single-independent-variable regression, if we select a version of that independent variable that is measured with error, the estimated slope coefficient on that variable will be biased toward 0.[65]

Exhibit 23 Results from Regressing T-Bill Returns on Actual Inflation

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0227	0.0034	6.6765
Actual inflation	0.8946	0.0761	11.7556
Residual standard error	0.0267		

(continued)

64 A consistent estimator is one for which the probability of estimates close to the value of the population parameter increases as sample size increases.

65 This proposition does not generalize to regressions with more than one independent variable. Of course, we ignore serially-correlated errors in this example, but because the regression coefficients are inconsistent (due to measurement error), testing or correcting for serial correlation is not worthwhile.

Exhibit 23 (Continued)

	Coefficient	Standard Error	*t*-Statistic
Multiple R-squared	0.4356		
Observations	181		

Source: Federal Reserve Bank of Philadelphia, US Department of Commerce.

5.4 Other Types of Time-Series Misspecification

By far the most frequent source of misspecification in linear regressions that use time series from two or more different variables is nonstationarity. Very roughly, **nonstationarity** means that a variable's properties, such as mean and variance, are not constant through time. We will postpone our discussion about stationarity to the reading on time-series analysis, but we can list some examples in which we need to use stationarity tests before we use regression statistical inference.[66]

- Relations among time series with trends (for example, the relation between consumption and GDP).
- Relations among time series that may be **random walks** (time series for which the best predictor of next period's value is this period's value). Exchange rates are often random walks.

The time-series examples in this reading were carefully chosen such that nonstationarity was unlikely to be an issue for any of them. But nonstationarity can be a very severe problem for analyzing the relations among two or more time series in practice. Analysts must understand these issues before they apply linear regression to analyzing the relations among time series. Otherwise, they may rely on invalid statistical inference.

6 MODELS WITH QUALITATIVE DEPENDENT VARIABLES

Financial analysts often need to be able to explain the outcomes of a qualitative dependent variable. **Qualitative dependent variables** (also called **categorical dependent variables**) are dummy variables used as dependent variables instead of as independent variables.

For example, to predict whether or not a company will go bankrupt, we need to use a qualitative dependent variable (bankrupt or not) as the dependent variable and use data on the company's financial performance (e.g., return on equity, debt-to-equity ratio, or debt rating) as independent variables. Unfortunately, linear regression is not the best statistical method to use for estimating such a model. If we use the qualitative dependent variable Y = [bankrupt (= 1) or not bankrupt (= 0)] as the dependent variable in a regression with financial variables as the independent variables, then we are estimating a linear probability model:

$$\text{Prob}(Y_\text{i} = 1 | X_{1i}, X_{2i}, X_{3i}) = p_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \varepsilon_i \quad (10)$$

66 We include both unit root tests and tests for cointegration in the term "stationarity tests."

Unfortunately, the predicted value of the probability could be much greater than 1 or much lower than 0 depending on the estimated coefficients b_i and the value of observed X_i's. Of course, these results would be invalid. The probability of bankruptcy (or of anything, for that matter) cannot be greater than 1.0 or less than 0.0.

Instead of a linear regression to estimate the probability of bankruptcy, we should use **probit regression (probit model)**, **logistic regression (logit model)**, or discriminant analysis for this kind of estimation.

Probit and logit regression models are used to estimate the probability of a discrete outcome given the values of the independent variables used to explain that outcome. Logistic regression is widely used in machine learning where the objective is classification. Denote by "p" the probability that a condition is fulfilled or that an event happens. Logistic regression involves modeling a dependent variable that is the natural logarithm of a ratio of probabilities—the probability that the event of interest happens divided by the probability that it does not happen.

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \varepsilon \tag{11}$$

This quantity is called a log odds ratio or logit; the transformation tends to linearize the relationship between the dependent and independent variables. The estimated event probability can be derived as:

$$\hat{p} = \frac{\exp\left[\hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2 + \hat{b}_3X_3\right]}{1 + \exp\left[\hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2\right]} \tag{12}$$

The right side of Equation 12 is a non-linear, sigmoidal (i.e., roughly S-shaped) function that constrains probability estimates to be between 0 and 1.

Logistic regression assumes a logistic distribution for the error term; this distribution is similar in shape to the normal distribution but has heavier tails. Probit and logistic regressions are both estimated by maximum likelihood methods rather than by least squares.[67]

Another technique to handle qualitative dependent variables is **discriminant analysis**. In his Z-score and Zeta analysis, Altman (1968, 1977) reported on the results of discriminant analysis. Altman uses financial ratios to predict the qualitative dependent variable bankruptcy. Discriminant analysis yields a linear function, similar to a regression equation, which can then be used to create an overall score. Based on the score, an observation can be classified into the bankrupt or not bankrupt category.

Qualitative dependent variable models can be useful not only for portfolio management but also for business management. For example, we might want to predict whether a client is likely to continue investing in a company or to withdraw assets from the company. We might also want to explain how particular demographic characteristics might affect the probability that a potential investor will sign on as a new client or evaluate the effectiveness of a particular direct-mail advertising campaign based on the demographic characteristics of the target audience. These issues can be analyzed with either probit or logit models.

67 For more on logistic regression, including distributional assumptions, and for an explanation of maximum likelihood methods, see Greene (2018).

EXAMPLE 15

Explaining Analyst Coverage

Suppose we want to investigate what factors determine whether at least one analyst covers a company. We can employ probit regression to address the question. The sample consists of 4,619 observations on public companies in 2013.

The variables in the probit model are as follows:

ANALYSTS = the discrete dependent variable, which takes on a value of 1 if at least one analyst covers the company and a value of 0 if no analysts cover the company

LNVOLUME = the natural log of the company's trading volume in the last month of the year

LNMV = the natural log of the market value of the company's equity

MATURITY = the mix of the company's earned and contributed capital, i.e., retained earnings as a proportion of total equity (RE/TE)

DIVPAYER = a dummy independent variable that takes on a value of 1 if the company paid a dividend

In this attempt to explain analyst coverage, we are examining whether more liquid companies, as captured by the trading volume in their shares, and larger companies, as captured by their market values, are more likely to be followed by at least one analyst. We also examine whether more mature and well established firms, as reflected in their mix of earned and contributed capital and their status as dividend payers, are more likely to be followed by an analyst. Exhibit 24 shows the results of the probit estimation.

Exhibit 24 Explaining Analyst Coverage Using a Probit Model

	Coefficient	Standard Error	*t*-Statistic
Intercept	2.5066	0.1005	24.9413
LNVOLUME	–0.0221	0.0173	–1.2775
LNMV	0.2441	0.0177	13.7910
MATURITY	0.0011	0.0007	1.5714
DIVPAYER	0.2798	0.0468	5.9786
Percent correctly predicted		78.00	

Source: I/B/E/S from Thomson Reuters, Center for Research in Security Prices at the University of Chicago, and S&P Capital IQ/Compustat.

As Exhibit 24 shows, two coefficients (besides the intercept) have *t*-statistics with an absolute value greater than 2.0. The coefficient on LNMV has a *t*-statistic of 13.7910. That value is far above the critical value at the 0.05 level for the *t*-statistic (1.96), so we can reject at the 0.05 level of significance the null hypothesis that the coefficient on LNMV equals 0, in favor of the alternative hypothesis that the coefficient is not equal to 0. The second coefficient with an absolute value greater than 2 is DIVPAYER, which has a *t*-statistic of 5.9786.

We can also reject at the 0.05 level of significance the null hypothesis that the coefficient on DIVPAYER is equal to 0, in favor of the alternative hypothesis that the coefficient is not equal to 0.

Neither of the two remaining independent variables is statistically significant at the 0.05 level in this probit analysis. That is, neither one reaches the critical value of 1.96 needed to reject the null hypothesis (that the associated coefficient is significantly different from 0). This result shows that once we take into account a company's market value and whether it pays dividends, the other factors—trading volume and maturity—have no power to explain whether at least one analyst will cover the company.

SUMMARY

In this reading, we have presented the multiple linear regression model and discussed violations of regression assumptions, model specification and misspecification, and models with qualitative variables.

- The general form of a multiple linear regression model is $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + \varepsilon_i$
- We conduct hypothesis tests concerning the population values of regression coefficients using t-tests of the form

$$t = \frac{\hat{b}_j - b_j}{s_{\hat{b}_j}}$$

- The lower the p-value reported for a test, the more significant the result.
- The assumptions of classical normal multiple linear regression model are as follows:
  1. A linear relation exists between the dependent variable and the independent variables.
  2. The independent variables are not random. Also, no exact linear relation exists between two or more of the independent variables.
  3. The expected value of the error term, conditioned on the independent variables, is 0.
  4. The variance of the error term is the same for all observations.
  5. The error term is uncorrelated across observations.
  6. The error term is normally distributed.
- To make a prediction using a multiple linear regression model, we take the following three steps:
  1. Obtain estimates of the regression coefficients.
  2. Determine the assumed values of the independent variables.
  3. Compute the predicted value of the dependent variable.
- When predicting the dependent variable using a linear regression model, we encounter two types of uncertainty: uncertainty in the regression model itself, as reflected in the standard error of estimate, and uncertainty about the estimates of the regression coefficients.

- The F-test is reported in an ANOVA table. The F-statistic is used to test whether at least one of the slope coefficients on the independent variables is significantly different from 0.

  $$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n - (k + 1)\right]} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}}$$

  Under the null hypothesis that all the slope coefficients are jointly equal to 0, this test statistic has a distribution of $F_{k,n-(k+1)}$, where the regression has n observations and k independent variables. The F-test measures the overall significance of the regression.

- R^2 is nondecreasing in the number of independent variables, so it is less reliable as a measure of goodness of fit in a regression with more than one independent variable than in a one-independent-variable regression.

  Analysts often choose to use adjusted R^2 because it does not necessarily increase when one adds an independent variable.

- Dummy variables in a regression model can help analysts determine whether a particular qualitative independent variable explains the model's dependent variable. A dummy variable takes on the value of 0 or 1. If we need to distinguish among n categories, the regression should include $n - 1$ dummy variables. The intercept of the regression measures the average value of the dependent variable of the omitted category, and the coefficient on each dummy variable measures the average incremental effect of that dummy variable on the dependent variable.

- If a regression shows significant conditional heteroskedasticity, the standard errors and test statistics computed by regression programs will be incorrect unless they are adjusted for heteroskedasticity.

- One simple test for conditional heteroskedasticity is the Breusch–Pagan test. Breusch and Pagan showed that, under the null hypothesis of no conditional heteroskedasticity, nR^2 (from the regression of the squared residuals on the independent variables from the original regression) will be a χ^2 random variable with the number of degrees of freedom equal to the number of independent variables in the regression.

- The principal effect of serial correlation in a linear regression is that the standard errors and test statistics computed by regression programs will be incorrect unless adjusted for serial correlation. Positive serial correlation typically inflates the t-statistics of estimated regression coefficients as well as the F-statistic for the overall significance of the regression.

- The most commonly used test for serial correlation is based on the Durbin–Watson statistic. If the Durbin–Watson statistic differs sufficiently from 2, then the regression errors have significant serial correlation.

- Multicollinearity occurs when two or more independent variables (or combinations of independent variables) are highly (but not perfectly) correlated with each other. With multicollinearity, the regression coefficients may not be individually statistically significant even when the overall regression is significant as judged by the F-statistic.

- Model specification refers to the set of variables included in the regression and the regression equation's functional form. The following principles can guide model specification:
  - The model should be grounded in cogent economic reasoning.
  - The functional form chosen for the variables in the regression should be appropriate given the nature of the variables.

- The model should be parsimonious.
- The model should be examined for violations of regression assumptions before being accepted.
- The model should be tested and be found useful out of sample before being accepted.

- If a regression is misspecified, then statistical inference using OLS is invalid and the estimated regression coefficients may be inconsistent.
- Assuming that a model has the correct functional form, when in fact it does not, is one example of misspecification. There are several ways this assumption may be violated:
  - One or more important variables could be omitted from the regression.
  - One or more of the regression variables may need to be transformed before estimating the regression.
  - The regression model pools data from different samples that should not be pooled.
- Another type of misspecification occurs when independent variables are correlated with the error term. This is a violation of Regression Assumption 3, that the error term has a mean of 0, and causes the estimated regression coefficients to be biased and inconsistent. Three common problems that create this type of time-series misspecification are:
  - including lagged dependent variables as independent variables in regressions with serially correlated errors;
  - including a function of dependent variable as an independent variable, sometimes as a result of the incorrect dating of variables; and
  - independent variables that are measured with error.
- Probit and logit models estimate the probability of a discrete outcome (the value of a qualitative dependent variable, such as whether a company enters bankruptcy) given the values of the independent variables used to explain that outcome. The probit model, which is based on the normal distribution, estimates the probability that $Y = 1$ (a condition is fulfilled) given the values of the independent variables. The logit model is identical, except that it is based on the logistic distribution rather than the normal distribution.

REFERENCES

Altman, Edward I. 1968. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *Journal of Finance*, vol. 23: 589–609.

Altman, Edward I., R. Halderman, and P. Narayanan. 1977. "Zeta Analysis: A New Model to Identify Bankruptcy Risk of Corporations." *Journal of Banking & Finance*, vol. 1: 29–54.

Bhabra, Harjeet S., and Jiayin Huang. 2013. "An Empirical Investigation of Mergers and Acquisitions by Chinese Listed Companies, 1997–2007." *Journal of Multinational Financial Management*, vol. 23: 186–207.

Breusch, T., and A. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica*, vol. 47: 1287–1294.

Buetow, Gerald W., Jr, Robert R. Johnson, and David E. Runkle. 2000. "The Inconsistency of Return-Based Style Analysis." *Journal of Portfolio Management*, vol. 26, no. 3: 61–77.

Durbin, J., and G.S. Watson. 1951. "Testing for Serial Correlation in Least Squares Regression, II." *Biometrika*, vol. 38: 159–178.

Greene, William H. 2018. *Economic Analysis*, 8th edition. New York: Pearson Education.

Gujarati, Damodar N., Dawn C. Porter, and Sangeetha Gunasekar. 2011. *Basic Econometrics*, 5th edition. New York: McGraw-Hill Irwin.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, vol. 50, no. 4: 1029–1054.

Keane, Michael P., and David E. Runkle. 1998. "Are Financial Analysts' Forecasts of Corporate Profits Rational?" *Journal of Political Economy*, vol. 106, no. 4: 768–805.

Kmenta, Jan. 1986. *Elements of Econometrics*, 2nd edition. New York: Macmillan.

Kyaw, NyoNyo A., John Manley, and Anand Shetty. 2011. "Factors in Multinational Valuations: Transparency, Political Risk, and Diversification." *Journal of Multinational Financial Management*, vol. 21: 55–67.

MacKinlay, A. Craig, and Matthew P. Richardson. 1991. "Using Generalized Methods of Moments to Test Mean–Variance Efficiency." *Journal of Finance*, vol. 46, no. 2: 511–527.

Mankiw, N. Gregory. 2015. *Macroeconomics*, 9th edition. New York: Worth Publishers.

Mayer, Thomas. 1980. "Economics as a Hard Science: Realistic Goal or Wishful Thinking?" *Economic Inquiry*, vol. 18, no. 2: 165–178.

Mitchell, T. 1997. *Machine Learning*. McGraw-Hill.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, vol. 55, no. 3: 703–708.

Petersen, Mitchell A. 2009. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *Review of Financial Studies*, vol. 22, no. 1: 435–480.

Samuel, A.July 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3): 210–29.

Sharpe, William F. 1988. "Determining a Fund's Effective Asset Mix." *Investment Management Review*, November/December:59–69.

Siegel, Jeremy J. 2014. *Stocks for the Long Run*, 5th edition. New York: McGraw-Hill.

Stoll, Hans R. 1978. "The Pricing of Security Dealer Services: An Empirical Study of Nasdaq Stocks." *Journal of Finance*, vol. 33, no. 4: 1153–1172.

Theobald, O. 2017. *Machine Learning for Absolute Beginners*. Independently published.

Tibshirani, R. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B. Methodological* 58 (1): 267–88.

Treynor, Jack L., and Kay Mazuy. 1966. "Can Mutual Funds Outguess the Market?" *Harvard Business Review*, vol. 44: 131–136.

Vapnik, V. N. 1989. *Statistical Learning Theory*. Wiley-Interscience.

Wang, W., Y. Huang, Y. Wang, and L. Wang. 2014. "Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction." *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*: 490–497.

PRACTICE PROBLEMS

1 With many US companies operating globally, the effect of the US dollar's strength on a US company's returns has become an important investment issue. You would like to determine whether changes in the US dollar's value and overall US equity market returns affect an asset's returns. You decide to use the S&P 500 Index to represent the US equity market.

A Write a multiple regression equation to test whether changes in the value of the dollar and equity market returns affect an asset's returns. Use the notations below.

R_{it} = return on the asset in period t
R_{Mt} = return on the S&P 500 in period t
ΔX_t = change in period t in the log of a trade-weighted index of the foreign exchange value of US dollar against the currencies of a broad group of major US trading partners.

B You estimate the regression for Archer Daniels Midland Company (NYSE: ADM). You regress its monthly returns for the period January 1990 to December 2002 on S&P 500 Index returns and changes in the log of the trade-weighted exchange value of the US dollar. The table below shows the coefficient estimates and their standard errors.

Coefficient Estimates from Regressing ADM's Returns: Monthly Data, January 1990–December 2002

	Coefficient	Standard Error
Intercept	0.0045	0.0062
R_{Mt}	0.5373	0.1332
ΔX_t	−0.5768	0.5121
n = 156		

Source: FactSet, Federal Reserve Bank of Philadelphia.

Determine whether S&P 500 returns affect ADM's returns. Then determine whether changes in the value of the US dollar affect ADM's returns. Use a 0.05 significance level to make your decisions.

C Based on the estimated coefficient on R_{Mt}, is it correct to say that "for a 1 percentage point increase in the return on the S&P 500 in period t, we expect a 0.5373 percentage point increase in the return on ADM"?

2 One of the most important questions in financial economics is what factors determine the cross-sectional variation in an asset's returns. Some have argued that book-to-market ratio and size (market value of equity) play an important role.

A Write a multiple regression equation to test whether book-to-market ratio and size explain the cross-section of asset returns. Use the notations below.

$(\text{B/M})_i$ = book-to-market ratio for asset i

R_i = return on asset i in a particular month

Size_i = natural log of the market value of equity for asset i

B The table below shows the results of the linear regression for a cross-section of 66 companies. The size and book-to-market data for each company are for December 2001. The return data for each company are for January 2002.

Results from Regressing Returns on the Book-to-Market Ratio and Size

	Coefficient	Standard Error
Intercept	0.0825	0.1644
$(\text{B/M})_i$	−0.0541	0.0588
Size_i	−0.0164	0.0350
n = 66		

Source: FactSet.

Determine whether the book-to-market ratio and size are each useful for explaining the cross-section of asset returns. Use a 0.05 significance level to make your decision.

3 There is substantial cross-sectional variation in the number of financial analysts who follow a company. Suppose you hypothesize that a company's size (market cap) and financial risk (debt-to-equity ratios) influence the number of financial analysts who follow a company. You formulate the following regression model:

$$(\text{Analyst following})_i = b_0 + b_1\text{Size}_i + b_2(\text{D/E})_i + \varepsilon_i$$

where

$(\text{Analyst following})_i$ = the natural log of $(1 + n)$, where n_i is the number of analysts following company i

Size_i = the natural log of the market capitalization of company i in millions of dollars

$(\text{D/E})_i$ = the debt-to-equity ratio for company i

In the definition of Analyst following, 1 is added to the number of analysts following a company because some companies are not followed by any analysts, and the natural log of 0 is indeterminate. The following table gives the coefficient estimates of the above regression model for a randomly selected sample of 500 companies. The data are for the year 2002.

Coefficient Estimates from Regressing Analyst Following on Size and Debt-to-Equity Ratio

	Coefficient	Standard Error	*t*-Statistic
Intercept	−0.2845	0.1080	−2.6343
Size_i	0.3199	0.0152	21.0461

(Continued)

	Coefficient	Standard Error	*t*-Statistic
$(D/E)_i$	−0.1895	0.0620	−3.0565
$n = 500$			

Source: First Call/Thomson Financial, Compustat.

A Consider two companies, both of which have a debt-to-equity ratio of 0.75. The first company has a market capitalization of \$100 million, and the second company has a market capitalization of \$1 billion. Based on the above estimates, how many more analysts will follow the second company than the first company?

B Suppose the p-value reported for the estimated coefficient on $(D/E)_i$ is 0.00236. State the interpretation of 0.00236.

4 In early 2001, US equity marketplaces started trading all listed shares in minimal increments (ticks) of \$0.01 (decimalization). After decimalization, bid–ask spreads of stocks traded on the NASDAQ tended to decline. In response, spreads of NASDAQ stocks cross-listed on the Toronto Stock Exchange (TSE) tended to decline as well. Researchers Oppenheimer and Sabherwal (2003) hypothesized that the percentage decline in TSE spreads of cross-listed stocks was related to company size, the predecimalization ratio of spreads on NASDAQ to those on the TSE, and the percentage decline in NASDAQ spreads. The following table gives the regression coefficient estimates from estimating that relationship for a sample of 74 companies. Company size is measured by the natural logarithm of the book value of company's assets in thousands of Canadian dollars.

Coefficient Estimates from Regressing Percentage Decline in TSE Spreads on Company Size, Predecimalization Ratio of NASDAQ to TSE Spreads, and Percentage Decline in NASDAQ Spreads

	Coefficient	*t*-Statistic
Intercept	−0.45	−1.86
Size_i	0.05	2.56
$(\text{Ratio of spreads})_i$	−0.06	−3.77
$(\text{Decline in NASDAQ spreads})_i$	0.29	2.42
$n = 74$		

Source: Oppenheimer and Sabherwal (2003).

The average company in the sample has a book value of assets of C\$900 million and a predecimalization ratio of spreads equal to 1.3. Based on the above model, what is the predicted decline in spread on the TSE for a company with these average characteristics, given a 1 percentage point decline in NASDAQ spreads?

5 The "neglected-company effect" claims that companies that are followed by fewer analysts will earn higher returns on average than companies that are followed by many analysts. To test the neglected-company effect, you have collected data on 66 companies and the number of analysts providing earnings estimates for each company. You decide to also include size as an independent variable, measuring size as the log of the market value of the company's equity, to try to distinguish any small-company effect from a neglected-company effect. The small-company effect asserts that small-company stocks may earn average higher risk-adjusted returns than large-company stocks.

The table below shows the results from estimating the model $R_i = b_0 + b_1\text{Size}_i + b_2(\text{Number of analysts})_i + \varepsilon_i$ for a cross-section of 66 companies. The size and number of analysts for each company are for December 2001. The return data are for January 2002.

Results from Regressing Returns on Size and Number of Analysts

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0388	0.1556	0.2495
Size_i	−0.0153	0.0348	−0.4388
$(\text{Number of analysts})_i$	0.0014	0.0015	0.8995

ANOVA	df	SS	MSS
Regression	2	0.0094	0.0047
Residual	63	0.6739	0.0107
Total	65	0.6833	
Residual standard error	0.1034		
R-squared	0.0138		
Observations	66		

Source: First Call/Thomson Financial, FactSet.

A What test would you conduct to see whether the two independent variables are *jointly* statistically related to returns (H_0: $b_1 = b_2 = 0$)?

B What information do you need to conduct the appropriate test?

C Determine whether the two variables jointly are statistically related to returns at the 0.05 significance level.

D Explain the meaning of adjusted R^2 and state whether adjusted R^2 for the regression would be smaller than, equal to, or larger than 0.0138.

6 Some developing nations are hesitant to open their equity markets to foreign investment because they fear that rapid inflows and outflows of foreign funds will increase volatility. In July 1993, India implemented substantial equity market reforms, one of which allowed foreign institutional investors into the Indian equity markets. You want to test whether the volatility of returns of stocks traded on the Bombay Stock Exchange (BSE) increased after July 1993, when foreign institutional investors were first allowed to invest in India. You have collected monthly return data for the BSE from February 1990 to December 1997. Your dependent variable is a measure of return volatility of stocks traded on the BSE; your independent variable is a dummy variable that is coded 1 if foreign investment was allowed during the month and 0 otherwise.

You believe that market return volatility actually *decreases* with the opening up of equity markets. The table below shows the results from your regression.

Results from Dummy Regression for Foreign Investment in India with a Volatility Measure as the Dependent Variable

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0133	0.0020	6.5351
Dummy	−0.0075	0.0027	−2.7604
n = 95			

Source: FactSet.

A State null and alternative hypotheses for the slope coefficient of the dummy variable that are consistent with testing your stated belief about the effect of opening the equity markets on stock return volatility.

B Determine whether you can reject the null hypothesis at the 0.05 significance level (in a one-sided test of significance).

C According to the estimated regression equation, what is the level of return volatility before and after the market-opening event?

7 Both researchers and the popular press have discussed the question as to which of the two leading US political parties, Republicans or Democrats, is better for the stock market.

A Write a regression equation to test whether overall market returns, as measured by the annual returns on the S&P 500 Index, tend to be higher when the Republicans or the Democrats control the White House. Use the notations below.

R_{Mt} = return on the S&P 500 in period t

Party_t = the political party controlling the White House (1 for a Republican president; 0 for a Democratic president) in period t

B The table below shows the results of the linear regression from Part A using annual data for the S&P 500 and a dummy variable for the party that controlled the White House. The data are from 1926 to 2002.

Results from Regressing S&P 500 Returns on a Dummy Variable for the Party That Controlled the White House, 1926-2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.1494	0.0323	4.6270
Party_t	−0.0570	0.0466	−1.2242

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	0.0625	0.0625	1.4987	0.2247
Residual	75	3.1287	0.0417		
Total	76	3.1912			
Residual standard error		0.2042			

(continued)

(Continued)

ANOVA	df	SS	MSS	*F*	Significance *F*
R-squared		0.0196			
Observations		77			

Source: FactSet.

Based on the coefficient and standard error estimates, verify to two decimal places the *t*-statistic for the coefficient on the dummy variable reported in the table.

C Determine at the 0.05 significance level whether overall US equity market returns tend to differ depending on the political party controlling the White House.

8 Problem 3 addressed the cross-sectional variation in the number of financial analysts who follow a company. In that problem, company size and debt-to-equity ratios were the independent variables. You receive a suggestion that membership in the S&P 500 Index should be added to the model as a third independent variable; the hypothesis is that there is greater demand for analyst coverage for stocks included in the S&P 500 because of the widespread use of the S&P 500 as a benchmark.

A Write a multiple regression equation to test whether analyst following is systematically higher for companies included in the S&P 500 Index. Also include company size and debt-to-equity ratio in this equation. Use the notations below.

(Analyst following)$_i$ = natural log of (1 + Number of analysts following company i)

Size$_i$ = natural log of the market capitalization of company i in millions of dollars

(D/E)$_i$ = debt-to-equity ratio for company i

S&P$_i$ = inclusion of company i in the S&P 500 Index (1 if included, 0 if not included)

In the above specification for analyst following, 1 is added to the number of analysts following a company because some companies are not followed by any analyst, and the natural log of 0 is indeterminate.

B State the appropriate null hypothesis and alternative hypothesis in a two-sided test of significance of the dummy variable.

C The following table gives estimates of the coefficients of the above regression model for a randomly selected sample of 500 companies. The data are for the year 2002. Determine whether you can reject the null hypothesis at the 0.05 significance level (in a two-sided test of significance).

Coefficient Estimates from Regressing Analyst Following on Size, Debt-to-Equity Ratio, and S&P 500 Membership, 2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	−0.0075	0.1218	−0.0616
Size$_i$	0.2648	0.0191	13.8639

(Continued)

	Coefficient	Standard Error	*t*-Statistic
$(D/E)_i$	−0.1829	0.0608	−3.0082
$S\&P_i$	0.4218	0.0919	4.5898
n = 500			

Source: First Call/Thomson Financial, Compustat.

D Consider a company with a debt-to-equity ratio of 2/3 and a market capitalization of \$10 billion. According to the estimated regression equation, how many analysts would follow this company if it were not included in the S&P 500 Index, and how many would follow if it were included in the index?

E In Problem 3, using the sample, we estimated the coefficient on the size variable as 0.3199, versus 0.2648 in the above regression. Discuss whether there is an inconsistency in these results.

9 You believe there is a relationship between book-to-market ratios and subsequent returns. The output from a cross-sectional regression and a graph of the actual and predicted relationship between the book-to-market ratio and return are shown below.

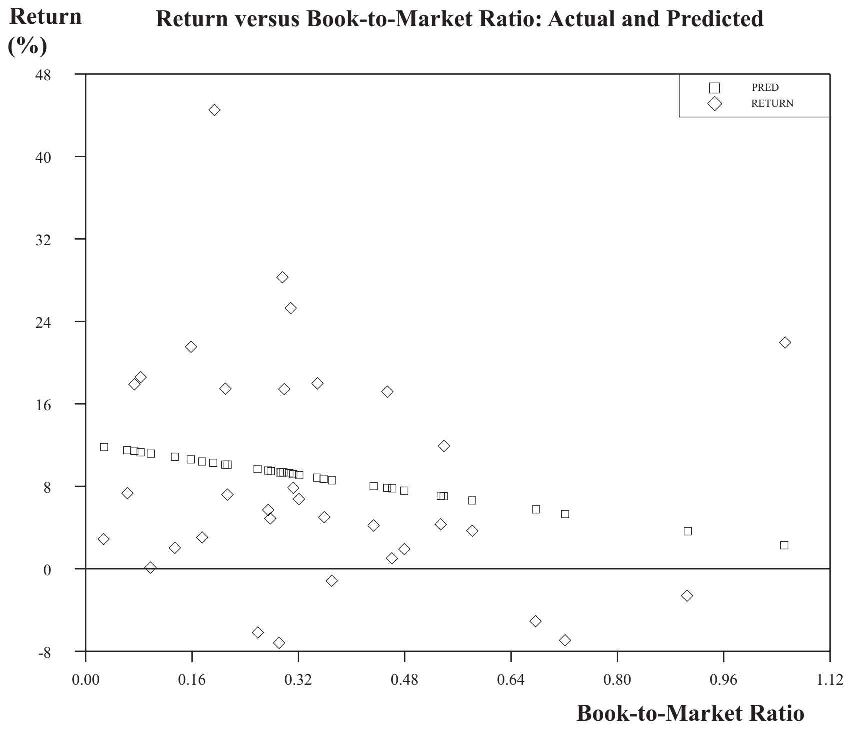
Results from Regressing Returns on the Book-to-Market Ratio

	Coefficient	Standard Error	*t*-Statistic
Intercept	12.0130	3.5464	3.3874
$\left(\frac{\text{Book value}}{\text{Market value}}\right)_i$	−9.2209	8.4454	−1.0918

ANOVA	df	SS	MSS	*F*	Significance *F*
Regression	1	154.9866	154.9866	1.1921	0.2831
Residual	32	4162.1895	130.0684		
Total	33	4317.1761			
Residual standard error		11.4048			
R-squared		0.0359			
Observations		34			

(continued)

(Continued)



A You are concerned with model specification problems and regression assumption violations. Focusing on assumption violations, discuss symptoms of conditional heteroskedasticity based on the graph of the actual and predicted relationship.

B Describe in detail how you could formally test for conditional heteroskedasticity in this regression.

C Describe a recommended method for correcting for conditional heteroskedasticity.

10 You are examining the effects of the January 2001 NYSE implementation of the trading of shares in minimal increments (ticks) of \$0.01 (decimalization). In particular, you are analyzing a sample of 52 Canadian companies cross-listed on both the NYSE and the Toronto Stock Exchange (TSE). You find that the bid–ask spreads of these shares decline on both exchanges after the NYSE decimalization. You run a linear regression analyzing the decline in spreads on the TSE, and find that the decline on the TSE is related to company size, pre-decimalization ratio of NYSE to TSE spreads, and decline in the NYSE spreads. The relationships are statistically significant. You want to be sure, however, that the results are not influenced by conditional heteroskedasticity. Therefore, you regress the squared residuals of the regression model on the three independent variables. The R^2 for this regression is 14.1 percent. Perform a statistical test to determine if conditional heteroskedasticity is present.

11 You are analyzing if institutional investors such as mutual funds and pension funds prefer to hold shares of companies with less volatile returns. You have the percentage of shares held by institutional investors at the end of 1998 for a random sample of 750 companies. For these companies, you compute the standard deviation of daily returns during that year. Then you regress the institutional holdings on the standard deviation of returns. You find that the regression is significant at the 0.01 level and the F-statistic is 12.98. The R^2 for this regression is 1.7 percent. As expected, the regression coefficient of the standard deviation of returns is negative. Its t-statistic is −3.60, which is also significant at the 0.01 level. Before concluding that institutions prefer to hold shares of less volatile stocks, however, you want to be sure that the regression results are not influenced by conditional heteroskedasticity. Therefore, you regress the squared residuals of the regression model on the standard deviation of returns. The R^2 for this regression is 0.6 percent.

A Perform a statistical test to determine if conditional heteroskedasticity is present at the 0.05 significance level.

B In view of your answer to Part A, what remedial action, if any, is appropriate?

12 In estimating a regression based on monthly observations from January 1987 to December 2002 inclusive, you find that the coefficient on the independent variable is positive and significant at the 0.05 level. You are concerned, however, that the t-statistic on the independent variable may be inflated because of serial correlation between the error terms. Therefore, you examine the Durbin–Watson statistic, which is 1.8953 for this regression.

A Based on the value of the Durbin–Watson statistic, what can you say about the serial correlation between the regression residuals? Are they positively correlated, negatively correlated, or not correlated at all?

B Compute the sample correlation between the regression residuals from one period and those from the previous period.

C Perform a statistical test to determine if serial correlation is present. Assume that the critical values for 192 observations when there is a single independent variable are about 0.09 above the critical values for 100 observations.

13 The book-to-market ratio and the size of a company's equity are two factors that have been asserted to be useful in explaining the cross-sectional variation in subsequent returns. Based on this assertion, you want to estimate the following regression model:

$$R_i = b_0 + b_1\left(\frac{\text{Book}}{\text{Market}}\right)_i + b_2\text{Size}_i + \varepsilon_i$$

where

R_i = Return of company i's shares (in the following period)

$\left(\frac{\text{Book}}{\text{Market}}\right)_i$ = company i's book-to-market ratio

Size_i = Market value of company i's equity

A colleague suggests that this regression specification may be erroneous, because he believes that the book-to-market ratio may be strongly related to (correlated with) company size.

A To what problem is your colleague referring, and what are its consequences for regression analysis?

B With respect to multicollinearity, critique the choice of variables in the regression model above.

Regression of Return on Book-to-Market and Size

	Coefficient	Standard Error	*t*-Statistic
Intercept	14.1062	4.220	3.3427
$\left(\frac{\text{Book}}{\text{Market}}\right)_i$	−12.1413	9.0406	−1.3430
Size_i	−0.00005502	0.00005977	−0.92047
R-squared	0.06156		
Observations	34		

Correlation Matrix

	Book-to-Market Ratio	Size
Book-to-Market Ratio	1.0000	
Size	−0.3509	1.0000

C State the classic symptom of multicollinearity and comment on that basis whether multicollinearity appears to be present, given the additional fact that the F-test for the above regression is not significant.

14 You are analyzing the variables that explain the returns on the stock of the Boeing Company. Because overall market returns are likely to explain a part of the returns on Boeing, you decide to include the returns on a value-weighted index of all the companies listed on the NYSE, AMEX, and NASDAQ as an independent variable. Further, because Boeing is a large company, you also decide to include the returns on the S&P 500 Index, which is a value-weighted index of the larger market-capitalization companies. Finally, you decide to include the changes in the US dollar's value. To conduct your test, you have collected the following data for the period 1990–2002.

R_t = monthly return on the stock of Boeing in month t
R_{ALLt} = monthly return on a value-weighted index of all the companies listed on the NYSE, AMEX, and NASDAQ in month t
R_{SPt} = monthly return on the S&P 500 Index in month t
ΔX_t = change in month t in the log of a trade-weighted index of the foreign exchange value of the US dollar against the currencies of a broad group of major US trading partners

The following table shows the output from regressing the monthly return on Boeing stock on the three independent variables.

Regression of Boeing Returns on Three Explanatory Variables: Monthly Data, January 1990–December 2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0026	0.0066	0.3939
R_{ALLt}	−0.1337	0.6219	−0.2150

(Continued)

	Coefficient	Standard Error	*t*-Statistic
R_{SPt}	0.8875	0.6357	1.3961
ΔX_t	0.2005	0.5399	0.3714

ANOVA	df	SS	MSS
Regression	3	0.1720	0.0573
Residual	152	0.8947	0.0059
Total	155	1.0667	
Residual standard error	0.0767		
R-squared	0.1610		
Observations	156		

Source: FactSet, Federal Reserve Bank of Philadelphia.

From the *t*-statistics, we see that none of the explanatory variables is statistically significant at the 5 percent level or better. You wish to test, however, if the three variables *jointly* are statistically related to the returns on Boeing.

A Your null hypothesis is that all three population slope coefficients equal 0—that the three variables *jointly* are statistically not related to the returns on Boeing. Conduct the appropriate test of that hypothesis.

B Examining the regression results, state the regression assumption that may be violated in this example. Explain your answer.

C State a possible way to remedy the violation of the regression assumption identified in Part B.

15 You are analyzing the cross-sectional variation in the number of financial analysts that follow a company (also the subject of Problems 3 and 8). You believe that there is less analyst following for companies with a greater debt-to-equity ratio and greater analyst following for companies included in the S&P 500 Index. Consistent with these beliefs, you estimate the following regression model.

$$(\text{Analysts following})_i = b_0 + b_1(\text{D/E})_i + b_2(\text{S\&P})_i + \varepsilon_i$$

where

(Analysts following)$_i$ = natural log of (1 + Number of analysts following company i)

(D/E)$_i$ = debt-to-equity ratio for company i

S&P$_i$ = inclusion of company i in the S&P 500 Index (1 if included; 0 if not included)

In the preceding specification, 1 is added to the number of analysts following a company because some companies are not followed by any analysts, and the natural log of 0 is indeterminate. The following table gives the coefficient estimates of the above regression model for a randomly selected sample of 500 companies. The data are for the year 2002.

Coefficient Estimates from Regressing Analyst Following on Debt-to-Equity Ratio and S&P 500 Membership, 2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	1.5367	0.0582	26.4038
$(D/E)_i$	−0.1043	0.0712	−1.4649
$S\&P_i$	1.2222	0.0841	14.5327
n = 500			

Source: First Call/Thomson Financial, Compustat.

You discuss your results with a colleague. She suggests that this regression specification may be erroneous, because analyst following is likely to be also related to the size of the company.

A What is this problem called, and what are its consequences for regression analysis?

B To investigate the issue raised by your colleague, you decide to collect data on company size also. You then estimate the model after including an additional variable, Size *i*, which is the natural log of the market capitalization of company *i* in millions of dollars. The following table gives the new coefficient estimates.

Coefficient Estimates from Regressing Analyst Following on Size, Debt-to-Equity Ratio, and S&P 500 Membership, 2002

	Coefficient	Standard Error	*t*-Statistic
Intercept	−0.0075	0.1218	−0.0616
$Size_i$	0.2648	0.0191	13.8639
$(D/E)_i$	−0.1829	0.0608	−3.0082
$S\&P_i$	0.4218	0.0919	4.5898
$n = 500$			

Source: First Call/Thomson Financial, Compustat.

What do you conclude about the existence of the problem mentioned by your colleague in the original regression model you had estimated?

16 You have noticed that hundreds of non-US companies are listed not only on a stock exchange in their home market but also on one of the exchanges in the United States. You have also noticed that hundreds of non-US companies are listed only in their home market and not in the United States. You are trying to predict whether or not a non-US company will choose to list on a US exchange. One of the factors that you think will affect whether or not a company lists in the United States is its size relative to the size of other companies in its home market.

A What kind of a dependent variable do you need to use in the model?

B What kind of a model should be used?

The following information relates to Questions 17–22

Gary Hansen is a securities analyst for a mutual fund specializing in small-capitalization growth stocks. The fund regularly invests in initial public offerings (IPOs). If the fund subscribes to an offer, it is allocated shares at the offer price. Hansen notes that IPOs frequently are underpriced, and the price rises when open market trading begins. The initial return for an IPO is calculated as the change in price on the first day of trading divided by the offer price. Hansen is developing a regression model to predict the initial return for IPOs. Based on past research, he selects the following independent variables to predict IPO initial returns:

Underwriter rank	=	1–10, where 10 is highest rank
Pre-offer price adjustment[a]	=	(Offer price – Initial filing price)/Initial filing price
Offer size ($ millions)	=	Shares sold × Offer price
Fraction retained[a]	=	Fraction of total company shares retained by insiders

[a]Expressed as a decimal

Hansen collects a sample of 1,725 recent IPOs for his regression model. Regression results appear in Exhibit 1, and ANOVA results appear in Exhibit 2.

Exhibit 1 Hansen's Regression Results Dependent Variable: IPO Initial Return (Expressed in Decimal Form, i.e., 1% = 0.01)

Variable	Coefficient (b_j)	Standard Error	*t*-Statistic
Intercept	0.0477	0.0019	25.11
Underwriter rank	0.0150	0.0049	3.06
Pre-offer price adjustment	0.4350	0.0202	21.53
Offer size	–0.0009	0.0011	–0.82
Fraction retained	0.0500	0.0260	1.92

Exhibit 2 Selected ANOVA Results for Hansen's Regression

	Degrees of Freedom (df)	Sum of Squares (SS)
Regression	4	51.433
Residual	1,720	91.436
Total	1,724	142.869
	Multiple R-squared = 0.36	

Hansen wants to use the regression results to predict the initial return for an upcoming IPO. The upcoming IPO has the following characteristics:

- underwriter rank = 6;
- pre-offer price adjustment = 0.04;

- offer size = $40 million;
- fraction retained = 0.70.

Because he notes that the pre-offer price adjustment appears to have an important effect on initial return, Hansen wants to construct a 95 percent confidence interval for the coefficient on this variable. He also believes that for each 1 percent increase in pre-offer price adjustment, the initial return will increase by less than 0.5 percent, holding other variables constant. Hansen wishes to test this hypothesis at the 0.05 level of significance.

Before applying his model, Hansen asks a colleague, Phil Chang, to review its specification and results. After examining the model, Chang concludes that the model suffers from two problems: 1) conditional heteroskedasticity, and 2) omitted variable bias. Chang makes the following statements:

Statement 1 "Conditional heteroskedasticity will result in consistent coefficient estimates, but both the *t*-statistics and *F*-statistic will be biased, resulting in false inferences."

Statement 2 "If an omitted variable is correlated with variables already included in the model, coefficient estimates will be biased and inconsistent and standard errors will also be inconsistent."

Selected values for the *t*-distribution and *F*-distribution appear in Exhibits 3 and 4, respectively.

Exhibit 3 Selected Values for the *t*-Distribution (df = ∞)

Area in Right Tail	*t*-Value
0.050	1.645
0.025	1.960
0.010	2.326
0.005	2.576

Exhibit 4 Selected Values for the *F*-Distribution (α = 0.01) (df1/df2: Numerator/Denominator Degrees of Freedom)

		df1	
		4	∞
df2	4	16.00	13.50
	∞	3.32	1.00

17 Based on Hansen's regression, the predicted initial return for the upcoming IPO is *closest* to:

A 0.0943.

B 0.1064.

C 0.1541.

18 The 95 percent confidence interval for the regression coefficient for the pre-offer price adjustment is *closest* to:

A 0.156 to 0.714.

B 0.395 to 0.475.

C 0.402 to 0.468.

19 The *most* appropriate null hypothesis and the *most* appropriate conclusion regarding Hansen's belief about the magnitude of the initial return relative to that of the pre-offer price adjustment (reflected by the coefficient b_j) are:

	Null Hypothesis	Conclusion about b_j (0.05 Level of Significance)
A	H_0: $b_j = 0.5$	Reject H_0
B	H_0: $b_j \geq 0.5$	Fail to reject H_0
C	H_0: $b_j \geq 0.5$	Reject H_0

20 The *most* appropriate interpretation of the multiple *R*-squared for Hansen's model is that:

A unexplained variation in the dependent variable is 36 percent of total variation.

B correlation between predicted and actual values of the dependent variable is 0.36.

C correlation between predicted and actual values of the dependent variable is 0.60.

21 Is Chang's Statement 1 correct?

A Yes.

B No, because the model's *F*-statistic will not be biased.

C No, because the model's *t*-statistics will not be biased.

22 Is Chang's Statement 2 correct?

A Yes.

B No, because the model's coefficient estimates will be unbiased.

C No, because the model's coefficient estimates will be consistent.

The following information relates to Questions 23–28

Adele Chiesa is a money manager for the Bianco Fund. She is interested in recent findings showing that certain business condition variables predict excess US stock market returns (one-month market return minus one-month T-bill return). She is also familiar with evidence showing how US stock market returns differ by the political party affiliation of the US President. Chiesa estimates a multiple regression model to predict monthly excess stock market returns accounting for business conditions and the political party affiliation of the US President:

$$\text{Excess stock market return}_t = a_0 + a_1\text{Default spread}_{t-1} + a_2\text{Term spread}_{t-1} + a_3\text{Pres party dummy}_{t-1} + e_t$$

Default spread is equal to the yield on Baa bonds minus the yield on Aaa bonds. Term spread is equal to the yield on a 10-year constant-maturity US Treasury index minus the yield on a 1-year constant-maturity US Treasury index. Pres party dummy is equal to 1 if the US President is a member of the Democratic Party and 0 if a member of the Republican Party.

Chiesa collects 432 months of data (all data are in percent form, i.e., 0.01 = 1 percent). The regression is estimated with 431 observations because the independent variables are lagged one month. The regression output is in Exhibit 1. Exhibits 2 through 5 contain critical values for selected test statistics.

Exhibit 1 Multiple Regression Output (the Dependent Variable Is the One-Month Market Return in Excess of the One-Month T-Bill Return)

	Coefficient	*t*-Statistic	*p*-Value
Intercept	−4.60	−4.36	<0.01
Default spread$_{t-1}$	3.04	4.52	<0.01
Term spread$_{t-1}$	0.84	3.41	<0.01
Pres party dummy$_{t-1}$	3.17	4.97	<0.01
Number of observations		431	
Test statistic from Breusch–Pagan (BP) test		7.35	
R^2		0.053	
Adjusted R^2		0.046	
Durbin–Watson (DW)		1.65	
Sum of squared errors (SSE)		19,048	
Regression sum of squares (SSR)		1,071	

An intern working for Chiesa has a number of questions about the results in Exhibit 1:

Question 1 How do you test to determine whether the overall regression model is significant?

Question 2 Does the estimated model conform to standard regression assumptions? For instance, is the error term serially correlated, or is there conditional heteroskedasticity?

Question 3 How do you interpret the coefficient for the Pres party dummy variable?

Question 4 Default spread appears to be quite important. Is there some way to assess the precision of its estimated coefficient? What is the economic interpretation of this variable?

After responding to her intern's questions, Chiesa concludes with the following statement: "Predictions from Exhibit 1 are subject to parameter estimate uncertainty, but not regression model uncertainty."

Exhibit 2 Critical Values for the Durbin–Watson Statistic (α = 0.05)

	K = 3	
N	d_l	d_u
420	1.825	1.854
430	1.827	1.855
440	1.829	1.857

Exhibit 3 Table of the Student's *t*-Distribution (One-Tailed Probabilities for df = ∞)

P	*t*
0.10	1.282
0.05	1.645
0.025	1.960
0.01	2.326
0.005	2.576

Exhibit 4 Values of χ^2

	Probability in Right Tail			
df	0.975	0.95	0.05	0.025
1	0.0001	0.0039	3.841	5.024
2	0.0506	0.1026	5.991	7.378
3	0.2158	0.3518	7.815	9.348
4	0.4840	0.7110	9.488	11.14

Exhibit 5 Table of the *F*-Distribution (Critical Values for Right-Hand Tail Area Equal to 0.05) Numerator: df1 and Denominator: df2

	df1				
df2	1	2	3	4	427
1	161	200	216	225	254
2	18.51	19.00	19.16	19.25	19.49
3	10.13	9.55	9.28	9.12	8.53

(continued)

Exhibit 5 (Continued)

	df1				
df2	**1**	**2**	**3**	**4**	**427**
4	7.71	6.94	6.59	6.39	5.64
427	3.86	3.02	2.63	2.39	1.17

23 Regarding the intern's Question 1, is the regression model as a whole significant at the 0.05 level?

A No, because the calculated *F*-statistic is less than the critical value for *F*.

B Yes, because the calculated *F*-statistic is greater than the critical value for *F*.

C Yes, because the calculated χ^2 statistic is greater than the critical value for χ^2.

24 Which of the following is Chiesa's *best* response to Question 2 regarding serial correlation in the error term? At a 0.05 level of significance, the test for serial correlation indicates that there is:

A no serial correlation in the error term.

B positive serial correlation in the error term.

C negative serial correlation in the error term.

25 Regarding Question 3, the Pres party dummy variable in the model indicates that the mean monthly value for the excess stock market return is:

A 1.43 percent larger during Democratic presidencies than Republican presidencies.

B 3.17 percent larger during Democratic presidencies than Republican presidencies.

C 3.17 percent larger during Republican presidencies than Democratic presidencies.

26 In response to Question 4, the 95 percent confidence interval for the regression coefficient for the default spread is *closest* to:

A 0.13 to 5.95.

B 1.72 to 4.36.

C 1.93 to 4.15.

27 With respect to the default spread, the estimated model indicates that when business conditions are:

A strong, expected excess returns will be higher.

B weak, expected excess returns will be lower.

C weak, expected excess returns will be higher.

28 Is Chiesa's concluding statement correct regarding parameter estimate uncertainty and regression model uncertainty?

A Yes.

B No, predictions are not subject to parameter estimate uncertainty.

C No, predictions are subject to regression model uncertainty and parameter estimate uncertainty.

The following information relates to Questions 29–36

Doris Honoré is a securities analyst with a large wealth management firm. She and her colleague Bill Smith are addressing three research topics: how investment fund characteristics affect fund total returns, whether a fund rating system helps predict fund returns, and whether stock and bond market returns explain the returns of a portfolio of utility shares run by the firm.

To explore the first topic, Honoré decides to study US mutual funds using a sample of 555 large-cap US equity funds. The sample includes funds in style classes of value, growth, and blend (i.e., combining value and growth characteristics). The dependent variable is the average annualized rate of return (in percent) over the past five years. The independent variables are fund expense ratio, portfolio turnover, the natural logarithm of fund size, fund age, and three dummy variables. The multiple manager dummy variable has a value of 1 if the fund has multiple managers (and a value of 0 if it has a single manager). The fund style is indicated by a growth dummy (value of 1 for growth funds and 0 otherwise) and a blend dummy (value of 1 for blend funds and 0 otherwise). If the growth and blend dummies are both zero, the fund is a value fund. The regression output is given in Exhibit 1.

Exhibit 1 Multiple Regression Output for Large-Cap Mutual Fund Sample

	Coefficient	Standard Error	*t*-Statistic
Intercept	10.9375	1.3578	8.0551
Expense ratio (%)	–1.4839	0.2282	–6.5039
Portfolio turnover (%)	0.0017	0.0016	1.0777
ln (fund size in $)	0.1467	0.0612	2.3976
Manager tenure (years)	–0.0098	0.0102	–0.9580
Multiple manager dummy	0.0628	0.1533	0.4100
Fund age (years)	–0.0123	0.0047	–2.6279
Growth dummy	2.4368	0.1886	12.9185
Blend dummy	0.5757	0.1881	3.0611

ANOVA	df	SS	MSS
Regression	8	714.169	89.2712
Residual	546	1583.113	2.8995
Total	554	2297.282	
Multiple *R*	0.5576		
R^2	0.3109		
Adjusted R^2	0.3008		
Standard error (%)	1.7028		
Observations	555		

Based on the results shown in Exhibit 1, Honoré wants to test the hypothesis that all of the regression coefficients are equal to zero. For the 555 fund sample, she also wants to compare the performance of growth funds with the value funds.

Honoré is concerned about the possible presence of multicollinearity in the regression. She states that adding a new independent variable that is highly correlated with one or more independent variables already in the regression model, has three potential consequences:

1 The R^2 is expected to decline.
2 The regression coefficient estimates can become imprecise and unreliable.
3 The standard errors for some or all of the regression coefficients will become inflated.

Another concern for the regression model (in Exhibit 1) is conditional heteroskedasticity. Honoré is concerned that the presence of heteroskedasticity can cause both the F-test for the overall significance of the regression and the t-tests for significance of individual regression coefficients to be unreliable. She runs a regression of the squared residuals from the model in Exhibit 1 on the eight independent variables, and finds the R^2 is 0.0669.

As a second research project, Honoré wants to test whether including Morningstar's rating system, which assigns a one- through five-star rating to a fund, as an independent variable will improve the predictive power of the regression model. To do this, she needs to examine whether values of the independent variables in a given period predict fund return in the next period. Smith suggests three different methods of adding the Morningstar ratings to the model:

- Method 1: Add an independent variable that has a value equal to the number of stars in the rating of each fund.
- Method 2: Add five dummy variables, one for each rating.
- Method 3: Add dummy variables for four of the five ratings.

As a third research project, Honoré wants to establish whether bond market returns (proxied by returns of long-term US Treasuries) and stock market returns (proxied by returns of the S&P 500 Index) explain the returns of a portfolio of utility stocks being recommended to clients. Exhibit 2 presents the results of a regression of 10 years of monthly percentage total returns for the utility portfolio on monthly total returns for US Treasuries and the S&P 500.

Exhibit 2 Regression Analysis of Utility Portfolio Returns

	Coefficient	Standard Error	t-Statistic	p-Value
Intercept	–0.0851	0.2829	–0.3008	0.7641
US Treasury	0.4194	0.0848	4.9474	<0.0001
S&P 500	0.6198	0.0666	9.3126	<0.0001

ANOVA	df	SS	MSS	F	Significance F
Regression	2	827.48	413.74	46.28	<0.0001
Residual	117	1045.93	8.94		
Total	119	1873.41			
Multiple R	0.6646				
R^2	0.4417				
Adjusted R^2	0.4322				

Exhibit 2 (Continued)

ANOVA	df	SS	MSS	F	Significance F
Standard error (%)	2.99				
Observations	120				

For the time-series model in Exhibit 2, Honoré says that positive serial correlation would not require that the estimated coefficients be adjusted, but that the standard errors of the regression coefficients would be underestimated. This issue would cause the t-statistics of the regression coefficients to be inflated. Honoré tests the null hypothesis that the there is no serial correlation in the regression residuals and finds that the Durbin–Watson statistic is equal to 1.81. The critical values at the 0.05 significance level for the Durbin–Watson statistic are $d_l = 1.63$ and $d_u = 1.72$.

Smith asks whether Honoré should have estimated the models in Exhibit 1 and Exhibit 2 using a probit or logit model instead of using a traditional regression analysis.

29 Considering Exhibit 1, the F-statistic is closest to:

A 3.22.

B 8.06.

C 30.79.

30 Based on Exhibit 1, the difference between the predicted annualized returns of a growth fund and an otherwise similar value fund is *closest* to:

A 1.86%.

B 2.44%.

C 3.01%.

31 Honoré describes three potential consequences of multicollinearity. Are all three consequences correct?

A Yes

B No, 1 is incorrect

C No, 2 is incorrect

32 Which of the three methods suggested by Smith would *best* capture the ability of the Morningstar rating system to predict mutual fund performance?

A Method 1

B Method 2

C Method 3

33 Honoré is concerned about the consequences of heteroskedasticity. Is she correct regarding the effect of heteroskedasticity on the reliability of the F-test and t-tests?

A Yes

B No, she is incorrect with regard to the F-test

C No, she is incorrect with regard to the t-tests

34 Is Honoré's description of the effects of positive serial correlation (in Exhibit 2) correct regarding the estimated coefficients and the standard errors?

A Yes

B No, she is incorrect about only the estimated coefficients

C No, she is incorrect about only the standard errors of the regression coefficients

35 Based on her estimated Durbin–Watson statistic, Honoré should:

A fail to reject the null hypothesis.

B reject the null hypothesis because there is significant positive serial correlation.

C reject the null hypothesis because there is significant negative serial correlation.

36 Should Honoré have estimated the models in Exhibit 1 and Exhibit 2 using probit or logit models instead of traditional regression analysis?

A Both should be estimated with probit or logit models.

B Neither should be estimated with probit or logit models.

C Only the analysis in Exhibit 1 should be done with probit or logit models.

The following information relates to Questions 37–45

Brad Varden, a junior analyst at an actively managed mutual fund, is responsible for research on a subset of the 500 large-cap equities the fund follows. Recently, the fund has been paying close attention to management turnover and to publicly available environmental, social, and governance (ESG) ratings. Varden is given the task of investigating whether any significant relationship exists between a company's profitability and either of these two characteristics. Colleen Quinni, a senior analyst at the fund, suggests that as an initial step in his investigation, Varden should perform a multiple regression analysis on the variables and report back to her.

Varden knows that Quinni is an expert at quantitative research, and she once told Varden that after you get an idea, you should formulate a hypothesis, test the hypothesis, and analyze the results. Varden expects to find that ESG rating is negatively related to ROE and CEO tenure is positively related to ROE. He considers a relationship meaningful when it is statistically significant at the 0.05 level. To begin, Varden collects values for ROE, CEO tenure, and ESG rating for a sample of 40 companies from the large-cap security universe. He performs a multiple regression with ROE (in percent) as the dependent variable and ESG rating and CEO tenure (in years) as the independent variables: $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \varepsilon_i$.

Exhibit 1 shows the regression results.

Exhibit 1 Regression Statistics

$$\hat{Y}_i = 9.442 + 0.069X_{1i} + 0.681X_{2i}$$

	Coefficient	Standard Error	*t*-Statistic	*p*-Value
Intercept	9.442	3.343	2.824	0.008
b_1 (ESG variable)	0.069	0.058	1.201	0.238
b_2 (Tenure variable)	0.681	0.295	2.308	0.027

Exhibit 1 (Continued)

ANOVA	df	SS	MSS	F	Significance F
Regression	2	240.410	120.205	4.161	0.023
Residual	37	1069.000	28.892		
Total	39	1309.410			
Multiple R	0.428				
R^2	0.183				
Adjusted R^2	0.139				
Standard error (%)	5.375				
Observations	40				

DF Associates is one of the companies Varden follows. He wants to predict its ROE using his regression model. DF Associates' corporate ESG rating is 55, and the company's CEO has been in that position for 10.5 years.

Varden also wants to check on the relationship between these variables and the dividend growth rate (divgr), so he completes the correlation matrix shown in Exhibit 2.

Exhibit 2 Correlation Matrix

	ROE	ESG	Tenure	Divgr
ROE	1.0			
ESG	0.446	1.0		
Tenure	0.369	0.091	1.0	
Divgr	0.117	0.046	0.028	1.0

Investigating further, Varden determines that dividend growth is not a linear combination of CEO tenure and ESG rating. He is unclear about how additional independent variables would affect the significance of the regression, so he asks Quinni, "Given this correlation matrix, will both R^2 and adjusted R^2 automatically increase if I add dividend growth as a third independent variable?"

The discussion continues, and Quinni asks two questions.

1 What does your F-statistic of 4.161 tell you about the regression?

2 In interpreting the overall significance of your regression model, which statistic do you believe is most relevant: R^2, adjusted R^2, or the F-statistic?

Varden answers both questions correctly and says he wants to check two more ideas. He believes the following:

1 ROE is less correlated with the dividend growth rate in firms whose CEO has been in office more than 15 years, and

2 CEO tenure is a normally distributed random variable.

Later, Varden includes the dividend growth rate as a third independent variable and runs the regression on the fund's entire group of 500 large-cap equities. He finds that the adjusted R^2 is much higher than the results in Exhibit 1. He reports this to Quinni and says, "Adding the dividend growth rate gives a model with a higher adjusted R^2. The three-variable model is clearly better." Quinni cautions, "I don't think you can conclude that yet."

37 Based on Exhibit 1 and given Varden's expectations, which is the *best* null hypothesis and conclusion regarding CEO tenure?

A $b_2 \leq 0$; reject the null hypothesis

B $b_2 = 0$; cannot reject the null hypothesis

C $b_2 \geq 0$; reject the null hypothesis

38 At a significance level of 1%, which of the following is the *best* interpretation of the regression coefficients with regard to explaining ROE?

A ESG is significant, but tenure is not.

B Tenure is significant, but ESG is not.

C Neither ESG nor tenure is significant.

39 Based on Exhibit 1, which independent variables in Varden's model are significant at the 0.05 level?

A ESG only

B Tenure only

C Neither ESG nor tenure

40 Based on Exhibit 1, the predicted ROE for DF Associates is *closest* to:

A 10.957%.

B 16.593%.

C 20.388%.

41 Based on Exhibit 2, Quinni's *best* answer to Varden's question about the effect of adding a third independent variable is:

A no for R^2 and no for adjusted R^2.

B yes for R^2 and no for adjusted R^2.

C yes for R^2 and yes for adjusted R^2.

42 Based on Exhibit 1, Varden's *best* answer to Quinni's question about the *F*-statistic is:

A both independent variables are significant at the 0.05 level.

B neither independent variable is significant at the 0.05 level.

C at least one independent variable is significant at the 0.05 level.

43 Varden's *best* answer to Quinni's question about overall significance is:

A R^2.

B adjusted R^2.

C the *F*-statistic.

44 If Varden's beliefs about ROE and CEO tenure are true, which of the following would violate the assumptions of multiple regression analysis?

A The assumption about CEO tenure distribution only

B The assumption about the ROE/dividend growth correlation only

C The assumptions about both the ROE/dividend growth correlation and CEO tenure distribution

45 The *best* rationale for Quinni's caution about the three-variable model is that the:

A dependent variable is defined differently.

B sample sizes are different in the two models.

C dividend growth rate is positively correlated with the other independent variables.

SOLUTIONS

1 A $R_{it} = b_0 + b_1 R_{Mt} + b_2 \Delta X_t + \varepsilon_{it}$

B We can test whether the coefficient on the S&P 500 Index returns is statistically significant. Our null hypothesis is that the coefficient is equal to 0 (H_0: $b_1 = 0$); our alternative hypothesis is that the coefficient is not equal to 0 (H_a: $b_1 \neq 0$). We construct the t-test of the null hypothesis as follows:

$$\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{0.5373 - 0}{0.1332} = 4.0338$$

where

$\hat{b}_1$ = regression estimate of b_1

b_1 = the hypothesized value of the coefficient (here, 0)

$s_{\hat{b}_1}$ = the estimated standard error of $\hat{b}_1$

Because this regression has 156 observations and three regression coefficients, the t-test has 156 − 3 = 153 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is between 1.98 and 1.97. The absolute value of the test statistic is 4.0338; therefore, we can reject the null hypothesis that $b_1 = 0$.

Similarly, we can test whether the coefficient on the change in the value of the US dollar is statistically significant in this regression. Our null hypothesis is that the coefficient is equal to 0 (H_0: $b_2 = 0$); our alternative hypothesis is that the coefficient is not equal to 0 (H_a: $b_2 \neq 0$). We construct the t-test as follows:

$$\frac{\hat{b}_2 - b_2}{s_{\hat{b}_2}} = \frac{-0.5768 - 0}{0.5121} = -1.1263$$

As before, the t-test has 153 degrees of freedom, and the critical value for the test statistic is between 1.98 and 1.97 at the 0.05 significance level. The absolute value of the test statistic is 1.1263; therefore, we cannot reject the null hypothesis that $b_2 = 0$.

Based on the above t-tests, we conclude that S&P 500 Index returns do affect ADM's returns but that changes in the value of the US dollar do not affect ADM's returns.

C The statement is not correct. To make it correct, we need to add the qualification "holding ΔX constant" to the end of the quoted statement.

2 A $R_i = b_0 + b_1(\text{B/M})_i + b_2 \text{Size}_i + \varepsilon_i$

B We can test whether the coefficients on the book-to-market ratio and size are individually statistically significant using t-tests. For the book-to-market ratio, our null hypothesis is that the coefficient is equal to 0 (H_0: $b_1 = 0$); our alternative hypothesis is that the coefficient is not equal to 0 (H_a: $b_1 \neq 0$). We can test the null hypothesis using a t-test constructed as follows:

$$\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{-0.0541 - 0}{0.0588} = -0.9201$$

where

$\hat{b}_1$ = regression estimate of b_1

b_1 = the hypothesized value of the coefficient (here, 0)

$s_{\hat{b}_1}$ = the estimated standard error of $\hat{b}_1$

This regression has 66 observations and three coefficients, so the t-test has 66 − 3 = 63 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is about 2.0. The absolute value of the test statistic is 0.9201; therefore, we cannot reject the null hypothesis that $b_1 = 0$. We can conclude that the book-to-market ratio is not useful in explaining the cross-sectional variation in returns for this sample.

We perform the same analysis to determine whether size (as measured by the log of the market value of equity) can help explain the cross-sectional variation in asset returns. Our null hypothesis is that the coefficient is equal to 0 (H_0: $b_2 = 0$); our alternative hypothesis is that the coefficient is not equal to 0 (H_a: $b_2 \neq 0$). We can test the null hypothesis using a t-test constructed as follows:

$$\frac{\hat{b}_2 - b_2}{s_{\hat{b}_2}} = \frac{-0.0164 - 0}{0.0350} = -0.4686$$

where

$\hat{b}_2$ = regression estimate of b_2

b_2 = the hypothesized value of the coefficient (here, 0)

$s_{\hat{b}_2}$ = the estimated standard error of $\hat{b}_2$

Again, because this regression has 66 observations and three coefficients, the t-test has 66 − 3 = 63 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is about 2.0. The absolute value of the test statistic is 0.4686; therefore, we cannot reject the null hypothesis that $b_2 = 0$. We can conclude that asset size is not useful in explaining the cross-sectional variation of asset returns in this sample.

3 A The estimated regression is (Analyst following)$_i$ = −0.2845 + 0.3199Size$_i$ − 0.1895(D/E)$_i$ + ε_i. Therefore, the prediction for the first company is

$$\begin{aligned}\text{(Analyst following)}_i &= -0.2845 + 0.3199(\ln 100) - 0.1895(0.75)\\ &= -0.2845 + 1.4732 - 0.1421 = 1.0466\end{aligned}$$

Recalling that (Analyst following)$_i$ is the natural log of $(1 + n_i)$, where n_i is the number of analysts following company i; it follows that $1 + n_1 = e^{1.0466}$ = 2.848, approximately. Therefore, n_1 = 2.848 − 1 = 1.848, or about two analysts. Similarly, the prediction for the second company is as follows:

$$\begin{aligned}\text{(Analyst following)}_i &= -0.2845 + 0.3199(\ln 1{,}000) - 0.1895(0.75)\\ &= -0.2845 + 2.2098 - 0.1421\\ &= 1.7832\end{aligned}$$

Thus, $1 + n_2 = e^{1.7832}$ = 5.949, approximately. Therefore, n_2 = 5.949 − 1 = 4.949, or about five analysts.

The model predicts that 5 − 2 = 3 more analysts will follow the second company than the first company.

B We would interpret the p-value of 0.00236 as the smallest level of significance at which we can reject a null hypothesis that the population value of the coefficient is 0, in a two-sided test. Clearly, in this regression the debt-to-equity ratio is a highly significant variable.

4 The estimated model is

Percentage decline in TSE spread of company i = –0.45 + 0.05Size$_i$ – 0.06(Ratio of spreads)$_i$ + 0.29(Decline in NASDAQ spreads)$_i$

Therefore, the prediction is

$$\begin{aligned}\text{Percentage decline in TSE spread} &= -0.45 + 0.05(\ln 900{,}000) - \\ &\quad\; 0.06(1.3) + 0.29(1) \\ &= -0.45 + 0.69 - 0.08 + 0.29 \\ &= 0.45\end{aligned}$$

The model predicts that for a company with average sample characteristics, the spread on the TSE declines by 0.45 percent for a 1 percent decline in NASDAQ spreads.

5 **A** To test the null hypothesis that all the slope coefficients in the regression model are equal to 0 (H_0: $b_1 = b_2 = 0$) against the alternative hypothesis that at least one slope coefficient is not equal to 0, we must use an F-test.

B To conduct the F-test, we need four inputs, all of which are found in the ANOVA section of the table in the statement of the problem:

i. total number of observations, n

ii. total number of regression coefficients to be estimated, $k + 1$

iii. sum of squared errors or residuals, $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ abbreviated SSE, and

iv. regression sum of squares, $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ abbreviated RSS

C The F-test formula is

$$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n - (k+1)\right]} = \frac{0.0094/2}{0.6739/\left[66 - (2+1)\right]} = 0.4394$$

The F-statistic has degrees of freedom $F\{k, [n - (k + 1)]\} = F(2, 63)$. From the F-test table, for the 0.05 significance level, the critical value for $F(2, 63)$ is about 3.15, so we cannot reject the hypothesis that the slope coefficients are both 0. The two independent variables are jointly statistically unrelated to returns.

D Adjusted R^2 is a measure of goodness of fit that takes into account the number of independent variables in the regression, in contrast to R^2. We can assert that adjusted R^2 is smaller than $R^2 = 0.0138$ without the need to perform any calculations. (However, adjusted R^2 can be shown to equal –0.0175 using an expression in the text on the relationship between adjusted R^2 and R^2.)

6 **A** You believe that opening markets actually reduces return volatility; if that belief is correct, then the slope coefficient would be negative, $b_1 < 0$. The null hypothesis is that the belief is not true: H_0: $b_1 \geq 0$. The alternative hypothesis is that the belief is true: H_a: $b_1 < 0$.

B The critical value for the t-statistic with 95 – 2 = 93 degrees of freedom at the 0.05 significance level in a one-sided test is about 1.66. For the one-sided test stated in Part A, we reject the null hypothesis if the t-statistic on the slope coefficient is less than −1.66. As the *t*-statistic of −2.7604 < −1.66, we reject the null. Because the dummy variable takes on a value of 1 when foreign investment is allowed, we can conclude that the volatility was lower with foreign investment.

C According to the estimated regression, average return volatility was 0.0133 (the estimated value of the intercept) before July 1993 and 0.0058 (= 0.0133 − 0.0075) after July 1993.

7 A The appropriate regression model is $R_{Mt} = b_0 + b_1 \text{Party}_t + \varepsilon_t$.

B The *t*-statistic reported in the table for the dummy variable tests whether the coefficient on Party_t is significantly different from 0. It is computed as follows:

$$\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{-0.0570 - 0}{0.0466} = -1.22$$

where

$\hat{b}_1$ = regression estimate of b_1

b_1 = the hypothesized value of the coefficient (here, 0)

$s_{\hat{b}_1}$ = the estimated standard error of $\hat{b}_1$

To two decimal places, this value is the same as the *t*-statistic reported in the table for the dummy variable, as expected. The problem specified two decimal places because the reported regression output reflects rounding; for this reason, we often cannot exactly reproduce reported *t*-statistics.

C Because the regression has 77 observations and two coefficients, the *t*-test has 77 − 2 = 75 degrees of freedom. At the 0.05 significance level, the critical value for the two-tailed test statistic is about 1.99. The absolute value of the test statistic is 1.2242; therefore, we do not reject the null hypothesis that b_1 = 0. We can conclude that the political party in the White House does not, on average, affect the annual returns of the overall market as measured by the S&P 500.

8 A The regression model is as follows:

$$(\text{Analyst following})_i = b_0 + b_1\text{Size}_i + b_2(\text{D/E})_i + b_3\text{S\&P}_i + \varepsilon_i$$

where $(\text{Analyst following})_i$ is the natural log of (1 + number of analysts following company *i*); Size_i is the natural log of the market capitalization of company *i* in millions of dollars; $(\text{D/E})_i$ is the debt-to-equity ratio for company *i*, and $\text{S\&P}_i$ is a dummy variable with a value of 1 if the company *i* belongs to the S&P 500 Index and 0 otherwise.

B The appropriate null and alternative hypotheses are H_0: $b_3 = 0$ and H_a: $b_3 \neq 0$, respectively.

C The *t*-statistic to test the null hypothesis can be computed as follows:

$$\frac{\hat{b}_3 - b_3}{s_{\hat{b}_3}} = \frac{0.4218 - 0}{0.0919} = 4.5898$$

This value is, of course, the same as the value reported in the table. The regression has 500 observations and 4 regression coefficients, so the *t*-test has 500 − 4 = 496 degrees of freedom. At the 0.05 significance level, the critical value for the test statistic is between 1.96 and 1.97. Because the value of the test statistic is 4.5898 we can reject the null hypothesis that $b_3 = 0$. Thus a company's membership in the S&P 500 appears to significantly influence the number of analysts who cover that company.

D The estimated model is

$$\left(\text{Analyst following}\right)_i = -0.0075 + 0.2648\text{Size}_i - 0.1829\left(\text{D/E}\right)_i + 0.4218\text{S\&P}_i + \varepsilon_i$$

Therefore the prediction for number of analysts following the indicated company that is not part of the S&P 500 Index is

$$\begin{aligned}(\text{Analyst following})_i &= -0.0075 + 0.2648(\ln 10{,}000) - 0.1829(2/3) + 0.4218(0)\\ &= -0.0075 + 2.4389 - 0.1219 + 0\\ &= 2.3095\end{aligned}$$

Recalling that (Analyst following)$_i$ is the natural log of $(1 + n_i)$, where n_i is the number of analysts following company i; it ensues (coding the company under consideration as 1) that $1 + n_1 = e^{2.3095} = 10.069$, approximately. Therefore, the prediction is that $n_1 = 10.069 - 1 = 9.069$, or about nine analysts.

Similarly, the prediction for the company that is included in the S&P 500 Index is

$$\begin{aligned}(\text{Analyst following})_i &= -0.0075 + 0.2648(\ln 10{,}000) - 0.1829(2/3) + 0.4218(1)\\ &= -0.0075 + 2.4389 - 0.1219 + 0.4218\\ &= 2.7313\end{aligned}$$

Coding the company that does belong to the S&P 500 as 2, $1 + n_2 = e^{2.7313} = 15.353$. Therefore, the prediction is that $n_2 = 15.353 - 1 = 14.353$, or about 14 analysts.

E There is no inconsistency in the coefficient on the size variable differing between the two regressions. The regression coefficient on an independent variable in a multiple regression model measures the expected net effect on the expected value of the dependent variable for a one-unit increase in that independent variable, after accounting for any effects of the other independent variables on the expected value of the dependent variable. The earlier regression had one fewer independent variable; after the effect of S&P 500 membership on the expected value of the dependent variable is taken into account, it is to be expected that the effect of the size variable on the dependent variable will change. What the regressions appear to indicate is that the net effect of the size variable on the expected analyst following diminishes when S&P 500 membership is taken into account.

9 **A** In a well-specified regression, the differences between the actual and predicted relationship should be random; the errors should not depend on the value of the independent variable. In this regression, the errors seem larger for smaller values of the book-to-market ratio. This finding indicates that we may have conditional heteroskedasticity in the errors, and consequently, the standard errors may be incorrect. We cannot proceed with hypothesis testing until we test for and, if necessary, correct for heteroskedasticity.

B A test for heteroskedasticity is to regress the squared residuals from the estimated regression equation on the independent variables in the regression. As seen in Section 4.1.2, Breusch and Pagan showed that, under the null hypothesis of no conditional heteroskedasticity, $n \times R^2$ (from the regression of the squared residuals on the independent variables from the original regression) will be a χ^2 random variable, with the number of degrees of freedom equal to the number of independent variables in the regression.

C One method to correct for heteroskedasticity is to use robust standard errors. This method uses the parameter estimates from the linear regression model but corrects the standard errors of the estimated parameters to account for the heteroskedasticity. Many statistical software packages can easily compute robust standard errors.

10 The test statistic is nR^2, where n is the number of observations and R^2 is the R^2 of the regression of squared residuals. So, the test statistic is 52 × 0.141 = 7.332. Under the null hypothesis of no conditional heteroskedasticity, this test statistic is a χ^2 random variable. There are three degrees of freedom, the number of independent variables in the regression. Appendix C, at the end of this volume, shows that for a one-tailed test, the test statistic critical value for a variable from a χ^2 distribution with 3 degrees of freedom at the 0.05 significance level is 7.815. The test statistic from the Breusch–Pagan test is 7.332. So, we cannot reject the hypothesis of no conditional heteroskedasticity at the 0.05 level. Therefore, we do not need to correct for conditional heteroskedasticity.

11 A The test statistic is nR^2, where n is the number of observations and R^2 is the R^2 of the regression of squared residuals. So, the test statistic is 750 × 0.006 = 4.5. Under the null hypothesis of no conditional heteroskedasticity, this test statistic is a χ^2 random variable. Because the regression has only one independent variable, the number of degrees of freedom is equal to 1. Appendix C, at the end of this volume, shows that for a one-tailed test, the test statistic critical value for a variable from a χ^2 distribution with one degree of freedom at the 0.05 significance level is 3.841. The test statistic is 4.5. So, we can reject the hypothesis of no conditional heteroskedasticity at the 0.05 level. Therefore, we need to correct for conditional heteroskedasticity.

B Two different methods can be used to correct for the effects of conditional heteroskedasticity in linear regression models. The first method involves computing robust standard errors. This method corrects the standard errors of the linear regression model's estimated parameters to account for the conditional heteroskedasticity. The second method is generalized least squares. This method modifies the original equation in an attempt to eliminate the heteroskedasticity. The new, modified regression equation is then estimated under the assumption that heteroskedasticity is no longer a problem.

Many statistical software packages can easily compute robust standard errors (the first method), and we recommend using them.

12 A Because the value of the Durbin–Watson statistic is less than 2, we can say that the regression residuals are positively correlated. Because this statistic is fairly close to 2, however, we cannot say without a statistical test if the serial correlation is statistically significant.

B From January 1987 through December 2002, there are 16 years, or 16 × 12 = 192 monthly returns. Thus the sample analyzed is quite large. Therefore, the Durbin–Watson statistic is approximately equal to $2(1 - r)$, where r is the sample correlation between the regression residuals from one period and those from the previous period.

$$\text{DW} = 1.8953 \approx 2(1 - r)$$

So, $r \approx 1 - \text{DW}/2 = 1 - 1.8953/2 = 0.0524$. Consistent with our answer to Part A, the correlation coefficient is positive.

C Appendix E indicates that the critical values d_l and d_u for 100 observations when there is one independent variable are 1.65 and 1.69, respectively. Based on the information given in the problem, the critical values d_l and d_u for about 200 observations when there is one independent variable are about 1.74 and 1.78, respectively. Because the DW statistic of 1.8953 for our regression is above d_u, we fail to reject the null hypothesis of no positive serial correlation. Therefore, we conclude that there is no evidence of positive serial correlation for the error term.

13 A This problem is known as multicollinearity. When some linear combinations of the independent variables in a regression model are highly correlated, the standard errors of the independent coefficient estimates become quite large, even though the regression equation may fit rather well.

B The choice of independent variables presents multicollinearity concerns because market value of equity appears in both variables.

C The classic symptom of multicollinearity is a high R^2 (and significant F-statistic) even though the t-statistics on the estimated slope coefficients are insignificant. Here a significant F-statistic does not accompany the insignificant t-statistics, so the classic symptom is not present.

14 A To test the null hypothesis that all of the regression coefficients except for the intercept in the multiple regression model are equal to 0 (H_0: $b_1 = b_2 = b_3 = 0$) against the alternative hypothesis that at least one slope coefficient is not equal to 0, we must use an F-test.

$$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n - (k + 1)\right]} = \frac{0.1720/3}{0.8947/\left[156 - (3 + 1)\right]} = 9.7403$$

The F-statistic has degrees of freedom $F\{k, [n - (k + 1)]\} = F(3, 152)$. From the F-test table, the critical value for $F(3, 120) = 2.68$ and $F(3, 152)$ will be less than $F(3, 120)$, so we can reject at the 0.05 significance level the null hypothesis that the slope coefficients are all 0. Changes in the three independent variables are jointly statistically related to returns.

B None of the t-statistics are significant, but the F-statistic is significant. This suggests the possibility of multicollinearity in the independent variables.

C The apparent multicollinearity is very likely related to the inclusion of *both* the returns on the S&P 500 Index *and* the returns on a value-weighted index of all the companies listed on the NYSE, AMEX, and NASDAQ as independent variables. The value-weighting of the latter index, giving relatively high weights to larger companies such as those included in the S&P 500, may make one return series an approximate linear function of the other. By dropping one or the other of these two variables, we might expect to eliminate the multicollinearity.

15 A Your colleague is indicating that you have omitted an important variable from the regression. This problem is called the omitted variable bias. If the omitted variable is correlated with an included variable, the estimated values of the regression coefficients would be biased and inconsistent. Moreover, the estimates of standard errors of those coefficients would also be inconsistent. So, we cannot use either the coefficient estimates or the estimates of their standard errors to perform statistical tests.

B A comparison of the new estimates with the original estimates clearly indicates that the original model suffered from the omitted variable bias due to the exclusion of company size from that model. As the *t*-statistics of the new model indicate, company size is statistically significant. Further, for the debt-to-equity ratio, the absolute value of the estimated coefficient substantially increases from 0.1043 to 0.1829, while its standard error declines. Consequently, it becomes significant in the new model, in contrast to the original model, in which it is not significant at the 5 percent level. The value of the estimated coefficient of the S&P 500 dummy substantially declines from 1.2222 to 0.4218. These changes imply that size should be included in the model.

16 **A** You need to use a qualitative dependent variable. You could give a value of 1 to this dummy variable for a listing in the United States and a value of 0 for not listing in the United States.

B Because you are using a qualitative dependent variable, linear regression is not the right technique to estimate the model. One possibility is to use either a probit or a logit model. Both models are identical, except that the logit model is based on logistic distribution while the probit model is based on normal distribution. Another possibility is to use discriminant analysis.

17 C is correct. The predicted initial return (IR) is:

$$\begin{aligned}IR &= 0.0477 + (0.0150 \times 6) + (0.435 \times 0.04) - (0.0009 \times 40) + (0.05 \times 0.70)\\ &= 0.1541\end{aligned}$$

18 B is correct. The 95% confidence interval is 0.435 ± (0.0202 × 1.96) = (0.395, 0.475).

19 C is correct. To test Hansen's belief about the direction and magnitude of the initial return, the test should be a one-tailed test. The alternative hypothesis is H_1: $b_j < 0.5$, and the null hypothesis is H_0: $b_j \geq 0.5$. The correct test statistic is: $t = (0.435 - 0.50)/0.0202 = -3.22$, and the critical value of the *t*-statistic for a one-tailed test at the 0.05 level is −1.645. The test statistic is significant, and the null hypothesis can be rejected at the 0.05 level of significance.

20 C is correct. The multiple *R*-squared for the regression is 0.36; thus, the model explains 36 percent of the variation in the dependent variable. The correlation between the predicted and actual values of the dependent variable is the square root of the *R*-squared or $\sqrt{0.36} = 0.60$.

21 A is correct. Chang is correct because the presence of conditional heteroskedasticity results in consistent parameter estimates, but biased (up or down) standard errors, *t*-statistics, and *F*-statistics.

22 A is correct. Chang is correct because a correlated omitted variable will result in biased and inconsistent parameter estimates and inconsistent standard errors.

23 B is correct.

The *F*-test is used to determine if the regression model as a whole is significant.

F = Mean square regression (MSR) ÷ Mean squared error (MSE)

MSE = SSE/[n − (k + 1)] = 19,048 ÷ 427 = 44.60

MSR = SSR/k = 1071 ÷ 3 = 357

F = 357 ÷ 44.60 = 8.004

The critical value for degrees of freedom of 3 and 427 with α = 0.05 (one-tail) is F = 2.63 from Exhibit 5. The calculated F is greater than the critical value, and Chiesa should reject the null hypothesis that all regression coefficients are equal to zero.

24 B is correct. The Durbin–Watson test used to test for serial correlation in the error term, and its value reported in Exhibit 1 is 1.65. For no serial correlation, DW is approximately equal to 2. If DW < d_l, *the error terms are positively serially correlated.* Because the DW = 1.65 is less than d_l = 1.827 for n = 431 (see Exhibit 2), Chiesa should reject the null hypothesis of no serial correlation and conclude that there is evidence of positive serial correlation among the error terms.

25 B is correct. The coefficient for the Pres party dummy variable (3.17) represents the increment in the mean value of the dependent variable related to the Democratic Party holding the presidency. In this case, the excess stock market return is 3.17 percent greater in Democratic presidencies than in Republican presidencies.

26 B is correct. The confidence interval is computed as $a_1 \pm s(a_1) \times t(95\%,\infty)$. From Exhibit 1, a_1 = 3.04 and $t(a_1)$ = 4.52, resulting in a standard error of $a_1 = s(a_1)$ = 3.04/4.52 = 0.673. The critical value for t from Exhibit 3 is 1.96 for p = 0.025. The confidence interval for a_1 is 3.04 ± 0.673 × 1.96 = 3.04 ± 1.31908 or from 1.72092 to 4.35908.

27 C is correct. The default spread is typically larger when business conditions are poor, i.e., a greater probability of default by the borrower. The positive sign for default spread (see Exhibit 1) indicates that expected returns are positively related to default spreads, meaning that excess returns are greater when business conditions are poor.

28 C is correct. Predictions in a multiple regression model are subject to both parameter estimate uncertainty and regression model uncertainty.

29 C is correct. The F-statistic is

$$F = \frac{\text{RSS}/k}{\text{SSE}/\left[n - (k+1)\right]} = \frac{714.169/8}{1583.113/546} = \frac{89.2712}{2.8995} = 30.79$$

Because F = 30.79 exceeds the critical F of 1.96, the null hypothesis that the regression coefficients are all 0 is rejected at the 0.05 significance level.

30 B is correct. The estimated coefficients for the dummy variables show the estimated difference between the returns on different types of funds. The growth dummy takes the value of 1 for growth funds and 0 for the value fund. Exhibit 1 shows a growth dummy coefficient of 2.4368. The estimated difference between the return of growth funds and value funds is thus 2.4368.

31 B is correct. The R^2 is expected to increase, not decline, with a new independent variable. The other two potential consequences Honoré describes are correct.

32 C is correct. Using dummy variables to distinguish among n categories would best capture the ability of the Morningstar rating system to predict mutual fund performance. We need n – 1 dummy variables to distinguish among n categories. In this case, there are five possible ratings and we need four dummy variables. Adding an independent variable that has a value equal to the number of stars in the rating of each fund is not appropriate because if the coefficient for this variable is positive, this method assumes that the extra return for a two-star fund is twice that of a one-star fund, the extra return for a three-star fund is three times that of a one-star fund, and so forth, which is not a reasonable assumption.

33 A is correct. Heteroskedasticity causes the *F*-test for the overall significance of the regression to be unreliable. It also causes the *t*-tests for the significance of individual regression coefficients to be unreliable because heteroskedasticity introduces bias into estimators of the standard error of regression coefficients.

34 A is correct. The model in Exhibit 2 does not have a lagged dependent variable. Positive serial correlation will, for such a model, not affect the consistency of the estimated coefficients. Thus, the coefficients will not need to be corrected for serial correlation. Positive serial correlation will, however, cause the standard errors of the regression coefficients to be understated; thus, the corresponding *t*-statistics will be inflated.

35 A is correct. The critical Durbin–Watson (D–W) values are d_l = 1.63 and d_u = 1.72. Because the estimated D–W value of 1.81 is greater than d_u = 1.73 (and less than 2), she fails to reject the null hypothesis of no serial correlation.

36 B is correct. Probit and logit models are used for models with qualitative dependent variables, such as models in which the dependent variable can have one of two discreet outcomes (i.e., 0 or 1). The analysis in the two exhibits are explaining security returns, which are continuous (not 0 or 1) variables.

37 A is correct. Varden expects to find that CEO tenure is positively related to the firm's ROE. If he is correct, the regression coefficient for tenure, b_2, will be greater than zero ($b_2 > 0$) and statistically significant. The null hypothesis supposes that the "suspected" condition is not true, so the null hypothesis should state the variable is less than or equal to zero. The *t*-statistic for tenure is 2.308, significant at the 0.027 level, meeting Varden's 0.05 significance requirement. Varden should reject the null hypothesis.

38 C is correct. The *t*-statistic for tenure is 2.308, indicating significance at the 0.027 level but not the 0.01 level. The *t*-statistic for ESG is 1.201, with a *p*-value of 0.238, which means we fail to reject the null hypothesis for ESG at the 0.01 significance level.

39 B is correct. The *t*-statistic for tenure is 2.308, which is significant at the 0.027 level. The *t*-statistic for ESG is 1.201, with a *p*-value of 0.238. This result is not significant at the 0.05 level.

40 C is correct. The regression equation is as follows:

$$\hat{Y}_i = 9.442 + 0.069X_{1i} + 0.681X_{2i}$$

$$\begin{aligned}\text{ROE} &= 9.442 + 0.069(\text{ESG}) + 0.681(\text{Tenure})\\ &= 9.442 + 0.069(55) + 0.681(10.5)\\ &= 9.442 + 3.795 + 7.151\\ &= 20.388.\end{aligned}$$

41 B is correct. When you add an additional independent variable to the regression model, the amount of unexplained variance will decrease, provided the new variable explains any of the previously unexplained variation. This result occurs as long as the new variable is even slightly correlated with the dependent variable. Exhibit 2 indicates the dividend growth rate is correlated with the dependent variable, ROE. Therefore, R^2 will increase.

Adjusted R^2, however, may not increase and may even decrease if the relationship is weak. This result occurs because in the formula for adjusted R^2, the new variable increases k (the number of independent variables) in the denominator, and the increase in R^2 may be insufficient to increase the value of the formula.

$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1 - R^2\right)$$

42 C is correct. Exhibit 1 indicates that the F-statistic of 4.161 is significant at the 0.05 level. A significant F-statistic means at least one of the independent variables is significant.

43 C is correct. In a multiple linear regression (as compared with simple regression), R^2 is less appropriate as a measure of whether a regression model fits the data well. A high adjusted R^2 does not necessarily indicate that the regression is well specified in the sense of including the correct set of variables. The F-test is an appropriate test of a regression's overall significance in either simple or multiple regressions.

44 C is correct. Multiple linear regression assumes that the relationship between the dependent variable and each of the independent variables is linear. Varden believes that this is not true for dividend growth because he believes the relationship may be different in firms with a long-standing CEO. Multiple linear regression also assumes that the independent variables are not random. Varden states that he believes CEO tenure is a random variable.

45 B is correct. If we use adjusted R^2 to compare regression models, it is important that the dependent variable be defined the same way in both models and that the sample sizes used to estimate the models are the same. Varden's first model was based on 40 observations, whereas the second model was based on 500.

READING

6

Time-Series Analysis

by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA

Richard A. DeFusco, PhD, CFA, is at the University of Nebraska-Lincoln (USA). Dennis W. McLeavey, DBA, CFA, is at the University of Rhode Island (USA). Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). David E. Runkle, PhD, CFA, is at Trilogy Global Advisors (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	**a.** calculate and evaluate the predicted trend value for a time series, modeled as either a linear trend or a log-linear trend, given the estimated trend coefficients;
☐	**b.** describe factors that determine whether a linear or a log-linear trend should be used with a particular time series and evaluate limitations of trend models;
☐	**c.** explain the requirement for a time series to be covariance stationary and describe the significance of a series that is not stationary;
☐	**d.** describe the structure of an autoregressive (AR) model of order *p* and calculate one- and two-period-ahead forecasts given the estimated coefficients;
☐	**e.** explain how autocorrelations of the residuals can be used to test whether the autoregressive model fits the time series;
☐	**f.** explain mean reversion and calculate a mean-reverting level;
☐	**g.** contrast in-sample and out-of-sample forecasts and compare the forecasting accuracy of different time-series models based on the root mean squared error criterion;
☐	**h.** explain the instability of coefficients of time-series models;
☐	**i.** describe characteristics of random walk processes and contrast them to covariance stationary processes;
☐	**j.** describe implications of unit roots for time-series analysis, explain when unit roots are likely to occur and how to test for them, and demonstrate how a time series with a unit root can be transformed so it can be analyzed with an AR model;

(continued)

LEARNING OUTCOMES

Mastery	The candidate should be able to:
☐	**k.** describe the steps of the unit root test for nonstationarity and explain the relation of the test to autoregressive time-series models;
☐	**l.** explain how to test and correct for seasonality in a time-series model and calculate and interpret a forecasted value using an AR model with a seasonal lag;
☐	**m.** explain autoregressive conditional heteroskedasticity (ARCH) and describe how ARCH models can be applied to predict the variance of a time series;
☐	**n.** explain how time-series variables should be analyzed for nonstationarity and/or cointegration before use in a linear regression;
☐	**o.** determine an appropriate time-series model to analyze a given investment problem and justify that choice.

1 INTRODUCTION TO TIME-SERIES ANALYSIS

As financial analysts, we often use time-series data to make investment decisions. A **time series** is a set of observations on a variable's outcomes in different time periods: the quarterly sales for a particular company during the past five years, for example, or the daily returns on a traded security. In this reading, we explore the two chief uses of time-series models: to explain the past and to predict the future of a time series. We also discuss how to estimate time-series models, and we examine how a model describing a particular time series can change over time. The following two examples illustrate the kinds of questions we might want to ask about time series.

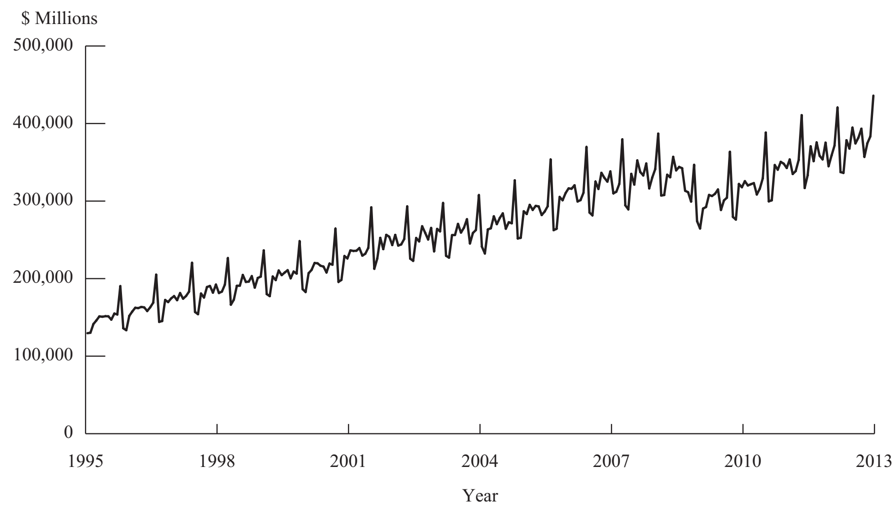
Suppose it is the beginning of 2014 and we are managing a US-based investment portfolio that includes Swiss stocks. Because the value of this portfolio would decrease if the Swiss franc depreciates with respect to the dollar, and vice-versa, holding all else constant, we are considering whether to hedge the portfolio's exposure to changes in the value of the franc. To help us in making this decision, we decide to model the time series of the franc/dollar exchange rate. Figure 1 shows monthly data on the franc/dollar exchange rate. (The data are monthly averages of daily exchange rates.) Has the exchange rate been more stable since 1987 than it was in previous years? Has the exchange rate shown a long-term trend? How can we best use past exchange rates to predict future exchange rates?

Figure 1 Swiss Franc/US Dollar Exchange Rate, Monthly Average of Daily Data

Source: Board of Governors of the Federal Reserve System.

As another example, suppose it is the beginning of 2014. We cover retail stores for a sell-side firm and want to predict retail sales for the coming year. Figure 2 shows monthly data on US retail sales. The data are not seasonally adjusted, hence the spikes around the holiday season at the turn of each year. Because the reported sales in the stores' financial statements are not seasonally adjusted, we model seasonally unadjusted retail sales. How can we model the trend in retail sales? How can we adjust for the extreme seasonality reflected in the peaks and troughs occurring at regular intervals? How can we best use past retail sales to predict future retail sales?

Figure 2 Monthly US Retail Sales



Source: US Department of Commerce, Census Bureau.

Some fundamental questions arise in time-series analysis: How do we model trends? How do we predict the future value of a time series based on its past values? How do we model seasonality? How do we choose among time-series models? And how do we model changes in the variance of time series over time? We address each of these issues in this reading.

The reading[1] is organized as follows. Section 2 describes typical challenges in applying the linear regression model to time series data. Section 3 presents linear and log-linear trend models, which describe, respectively, the value and the natural log of the value of a time series as a linear function of time. Section 4 presents autoregressive time series models—which explain the current value of a time series in terms of one or more lagged values of the series. Such models are among the most commonly used in investments, and the section addresses many related concepts and issues. Section 5 addresses time series that are random walks. Because such time series are not covariance stationary, they cannot be modeled using autoregressive models unless they can be transformed into stationary series. The section explores appropriate transformations and tests of stationarity. Section 6 addresses moving-average time-series models, and Section 7 discusses the problem of seasonality in time series and how to address it. Section 8 covers autoregressive moving-average models, a more complex alternative to autoregressive models. Section 9 addresses modeling changing variance of the error term in a time series. Section 10 examines the consequences of regression of one time series on another when one or both time series may not be covariance stationary.

2 CHALLENGES OF WORKING WITH TIME SERIES

Throughout the reading, our objective will be to apply linear regression to a given time series. Unfortunately, in working with time series we often find that the assumptions of the linear regression model are not satisfied. To apply time-series analysis, we need to assure ourselves that the linear regression model assumptions are met. When those assumptions are not satisfied, in many cases we can transform the time series, or specify the regression model differently, so that the assumptions of the linear regression model are met.

We can illustrate assumption difficulties in the context of a common time-series model, an autoregressive model. Informally, an autoregressive model is one in which the independent variable is a lagged (that is, past) value of the dependent variable, such as the model $x_t = b_0 + b_1x_{t-1} + \varepsilon_t$.[2] Specific problems that we often encounter in dealing with time series include the following:

- The residual errors are correlated instead of being uncorrelated. In the calculated regression, the difference between x_t and $b_0 + b_1x_{t-1}$ is called the residual error (ε_t). The linear regression assumes that this error term is not correlated across observations. The violation of that assumption is frequently more critical in terms of its consequences in the case of time-series models involving past values of the time series as independent variables than for other models (such as cross-sectional) in which the dependent and independent variables are distinct. As we discussed in the reading on multiple regression, in a regression in which the dependent and independent variables are distinct, serial correlation of the errors in this model does not affect the consistency of our estimates of intercept or slope coefficients. By contrast, in an autoregressive

1 Examples in this reading were updated in 2014 by Professor Sanjiv Sabherwal of the University of Texas, Arlington.

2 We could also write the equation as $y_t = b_0 + b_1y_{t-1} + \varepsilon_t$.

time-series regression such as $x_t = b_0 + b_1x_{t-1} + \varepsilon_t$, serial correlation in the error term causes estimates of the intercept (b_0) and slope coefficient (b_1) to be inconsistent.

- The mean and/or variance of the time series changes over time. Regression results are invalid if we estimate an autoregressive model for a time series with mean and/or variance that changes over time.

Before we try to use time series for forecasting, we may need to transform the time-series model so that it is well specified for linear regression. With this objective in mind, you will observe that time-series analysis is relatively straightforward and logical.

TREND MODELS

3

Estimating a trend in a time series and using that trend to predict future values of the time series is the simplest method of forecasting. For example, we saw in Figure 2 that monthly US retail sales show a long-term pattern of upward movement—that is, a **trend**. In this section, we examine two types of trends—linear trends and log-linear trends—and discuss how to choose between them.

3.1 Linear Trend Models

The simplest type of trend is a **linear trend**, one in which the dependent variable changes at a constant rate with time. If a time series, y_t, has a linear trend, then we can model the series using the following regression equation:

$$y_t = b_0 + b_1t + \varepsilon_t,\ t = 1, 2, \ldots, T \tag{1}$$

where

y_t = the value of the time series at time t (value of the dependent variable)
b_0 = the y-intercept term
b_1 = the slope coefficient
t = time, the independent or explanatory variable
ε_t = a random-error term

In Equation 1, the trend line, $b_0 + b_1t$, predicts the value of the time series at time t (where t takes on a value of 1 in the first period of the sample and increases by 1 in each subsequent period). Because the coefficient b_1 is the slope of the trend line, we refer to b_1 as the trend coefficient. We can estimate the two coefficients, b_0 and b_1, using ordinary least squares, denoting the estimated coefficients as $\hat{b}_0$ and $\hat{b}_1$.[3]

Now we demonstrate how to use these estimates to predict the value of the time series in a particular period. Recall that t takes on a value of 1 in Period 1. Therefore, the predicted or fitted value of y_t in Period 1 is $\hat{y}_1 = \hat{b}_0 + \hat{b}_1(1)$. Similarly, in a subsequent period, say the sixth period, the fitted value is $\hat{y}_6 = \hat{b}_0 + \hat{b}_1(6)$. Now suppose that we want to predict the value of the time series for a period outside the sample, say period $T + 1$. The predicted value of y_t for period $T + 1$ is $\hat{y}_{T+1} = \hat{b}_0 + \hat{b}_1(T + 1)$. For example, if $\hat{b}_0$ is 5.1 and $\hat{b}_1$ is 2, then at $t = 5$ the predicted value of y_5 is 15.1 and

3 Recall that ordinary least squares is an estimation method based on the criterion of minimizing the sum of a regression's squared residuals.

at t = 6 the predicted value of y_6 is 17.1. Note that each consecutive observation in this time series increases by $\hat{b}_1 = 2$ irrespective of the level of the series in the previous period.

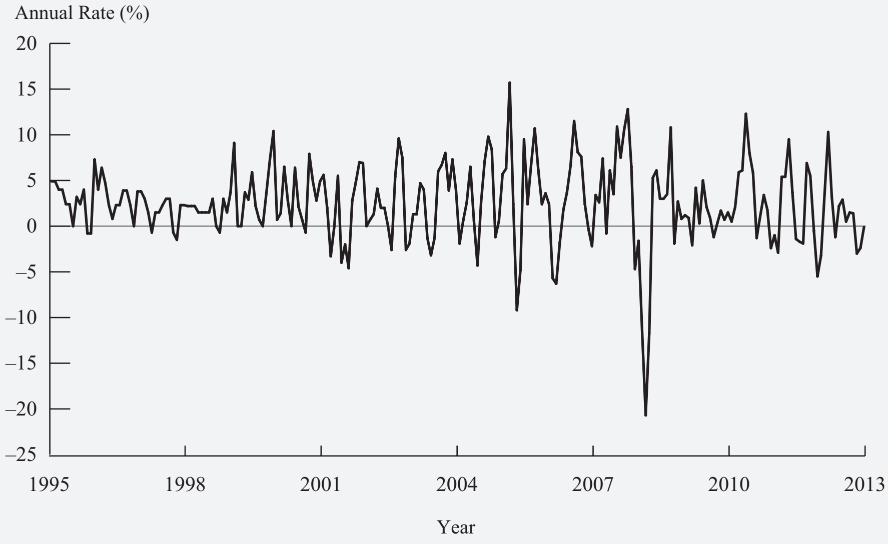
EXAMPLE 1

The Trend in the US Consumer Price Index

It is January 2014. As a fixed income analyst in the trust department of a bank, Lisette Miller is concerned about the future level of inflation and how it might affect portfolio value. Therefore, she wants to predict future inflation rates. For this purpose, she first needs to estimate the linear trend in inflation. To do so, she uses the monthly US Consumer Price Index (CPI) inflation data, expressed as an annual percentage rate,[4] shown in Figure 3. The data include 228 months from January 1995 through December 2013, and the model to be estimated is $y_t = b_0 + b_1 t + \varepsilon_t$, t = 1, 2, …, 228. Table 1 shows the results of estimating this equation. With 228 observations and two parameters, this model has 226 degrees of freedom. At the 0.05 significance level, the critical value for a t-statistic is 1.97. The intercept $\left(\hat{b}_0 = 2.8853\right)$ is statistically significant because the value of the t-statistic for the coefficient is well above the critical value. The trend coefficient is negative $\left(\hat{b}_1 = -0.0038\right)$, suggesting a declining trend in inflation during the sample time period. However, the trend is not statistically significant because the absolute value of the t-statistic for the coefficient is well below the critical value. The estimated regression equation can be written as

$$y_t = 2.8853 - 0.0038t$$

Figure 3 Monthly CPI Inflation, Not Seasonally Adjusted



Source: Bureau of Labor Statistics.

4 In these data, 1 percent is represented as 1.0.

Table 1 Estimating a Linear Trend in Inflation Monthly Observations, January 1995–December 2013

Regression Statistics	
R-squared	0.0033
Standard error	4.3297
Observations	228
Durbin–Watson	1.09

	Coefficient	Standard Error	*t*-Statistic
Intercept	2.8853	0.5754	5.0144
t (Trend)	−0.0038	0.0044	−0.8636

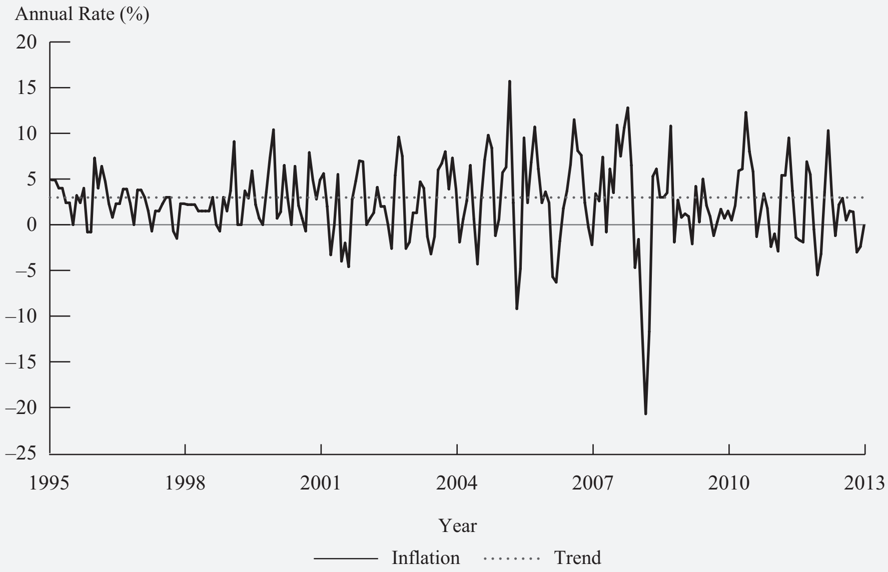
Source: US Bureau of Labor Statistics.

Because the trend line slope is estimated to be −0.0038, Miller concludes that the linear trend model's best estimate is that the annualized rate of inflation declined at a rate of about 38 basis points per month during the sample time period. The decline is not statistically significantly different from zero.

In January 1995, the first month of the sample, the predicted value of inflation is $\hat{y}_1 = 2.8853 - 0.0038(1) = 2.8815$ percent. In December 2013, the 228th or last month of the sample, the predicted value of inflation is $\hat{y}_{228} = 2.8853 - 0.0038(228) = 2.0189$ percent. Note, though, that these predicted values are for in-sample periods. A comparison of these values with the actual values indicates how well Miller's model fits the data; however, a main purpose of the estimated model is to predict the level of inflation for out-of-sample periods. For example, for December 2014 (12 months after the end of the sample), $t = 228 + 12 = 240$, and the predicted level of inflation is $\hat{y}_{240} = 2.8853 - 0.0038(240) = 1.9733$ percent.

Figure 4 shows the inflation data along with the fitted trend. Consistent with the negative but small and statistically insignificant trend coefficient, the fitted trend line is slightly downward sloping. Note that inflation does not appear to be above or below the trend line for a long period of time. No persistent differences exist between the trend and actual inflation. The residuals (actual minus trend values) appear to be unpredictable and uncorrelated in time. Therefore, using a linear trend line to model inflation rates from 1995 through 2013 does not appear to violate the assumptions of the linear regression model. Note also that the R^2 in this model is quite low, indicating great uncertainty in the inflation forecasts from this model. In fact, the estimated model explains only 0.33 percent of the variation in monthly inflation. Although linear trend models have their uses, they are often inappropriate for economic data. Most economic time series reflect trends with changing slopes and/or intercepts over time. The linear trend model identifies the slope and intercept that provides the best linear fit for all past data. The model's deviation from the actual data can be greatest near the end of a data series which can compromise forecasting accuracy. Later in this reading, we will examine whether we can build a better model of inflation than a model that uses only a trend line.

Figure 4 Monthly CPI Inflation with Trend



Source: US Bureau of Labor Statistics.

3.2 Log-Linear Trend Models

Sometimes a linear trend does not correctly model the growth of a time series. In those cases, we often find that fitting a linear trend to a time series leads to persistent rather than uncorrelated errors. If the residuals from a linear trend model are persistent, we then need to employ an alternative model satisfying the conditions of linear regression. For financial time series, an important alternative to a linear trend is a log-linear trend. Log-linear trends work well in fitting time series that have exponential growth.

Exponential growth means constant growth at a particular rate. For example, annual growth at a constant rate of 5 percent is exponential growth. How does exponential growth work? Suppose we describe a time series by the following equation:

$$y_t = e^{b_0 + b_1 t},\ t = 1,\ 2,\ \ldots,\ T \tag{2}$$

Exponential growth is growth at a constant rate $\left(e^{b_1} - 1\right)$ with continuous compounding. For instance, consider values of the time series in two consecutive periods. In Period 1, the time series has the value $y_1 = e^{b_0 + b_1(1)}$, and in Period 2, it has the value $y_2 = e^{b_0 + b_1(2)}$. The resulting ratio of the values of the time series in the first two periods is $y_2/y_1 = \left(e^{b_0 + b_1(2)}\right)\Big/\left(e^{b_0 + b_1(1)}\right) = e^{b_1(1)}$. Generally, in any period t, the time series has the value $y_t = e^{b_0 + b_1(t)}$. In period $t + 1$, the time series has the value $y_{t+1} = e^{b_0 + b_1(t+1)}$. The ratio of the values in the periods $(t + 1)$ and t is $y_{t+1}/y_t = e^{b_0 + b_1(t+1)}\Big/e^{b_0 + b_1(t)} = e^{b_1(1)}$. Thus, the proportional rate of growth in the time series over two consecutive periods is always the same: $(y_{t+1} - y_t)/y_t = y_{t+1}/y_t - 1 = e^{b_1} - 1$.[5] Therefore, exponential growth is growth at a constant rate. Continuous compounding is a mathematical convenience that allows us to restate the equation in a form that is easy to estimate.

If we take the natural log of both sides of Equation 2, the result is the following equation:

$$\ln y_t = b_0 + b_1 t,\ t = 1, 2, \ldots, T$$

Therefore, if a time series grows at an exponential rate, we can model the natural log of that series using a linear trend.[6] Of course, no time series grows exactly at a constant rate. Consequently, if we want to use a **log-linear model**, we must estimate the following equation:

$$\ln y_t = b_0 + b_1 t + \varepsilon_t,\ t = 1, 2, \ldots, T \tag{3}$$

Note that this equation is linear in the coefficients b_0 and b_1. In contrast to a linear trend model, in which the predicted trend value of y_t is $\hat{b}_0 + \hat{b}_1 t$, the predicted trend value of y_t in a log-linear trend model is $e^{\hat{b}_0 + \hat{b}_1 t}$ because $e^{\ln y_t} = y_t$.

Examining Equation 3, we see that a log-linear model predicts that $\ln y_t$ will increase by b_1 from one time period to the next. The model predicts a constant growth rate in y_t of $e^{b_1} - 1$. For example, if $b_1 = 0.05$, then the predicted growth rate of y_t in each period is $e^{0.05} - 1 = 0.051271$ or 5.13 percent. In contrast, the linear trend model (Equation 1) predicts that y_t grows by a constant amount from one period to the next.

Example 2 illustrates the problem of nonrandom residuals in a linear trend model, and Example 3 shows a log-linear regression fit to the same data.

EXAMPLE 2

A Linear Trend Regression for Quarterly Sales at Starbucks

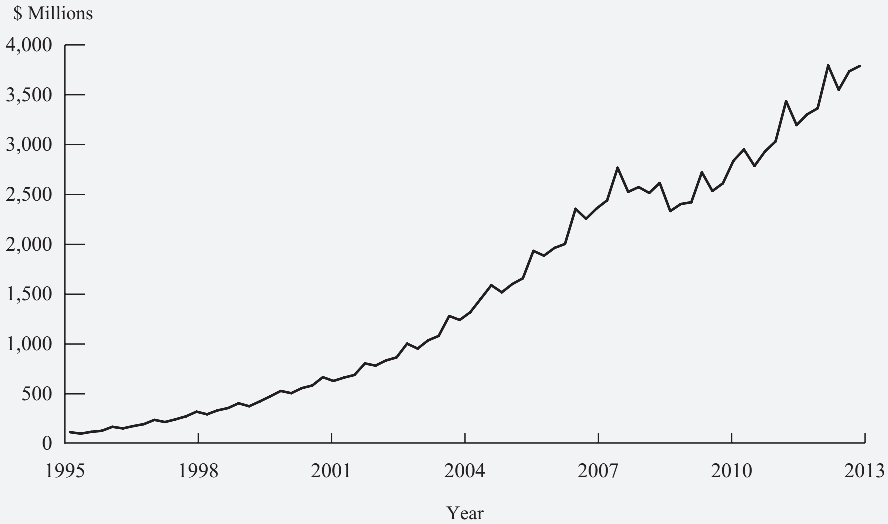
In October 2013, technology analyst Ray Benedict wants to use Equation 1 to fit the data on quarterly sales for Starbucks Corporation shown in Figure 5. Starbucks' fiscal year ends in September. Benedict uses 76 observations on Starbucks' sales from the first quarter of fiscal year 1995 (starting in October 1994) to the fourth quarter of fiscal year 2013 (ending in September 2013) to estimate the linear trend regression model $y_t = b_0 + b_1 t + \varepsilon_t$, $t = 1, 2, \ldots, 76$.[7] Table 2 shows the results of estimating this equation.

5 For example, if we use annual periods and $e^{b_1} = 1.04$ for a particular series, then that series grows by $1.04 - 1 = 0.04$, or 4 percent a year.

6 An exponential growth rate is a compound growth rate with continuous compounding.

7 In discussions of Starbucks' sales in this reading, year refers to Starbucks' fiscal year.

Figure 5 Starbucks Quarterly Sales by Fiscal Year



Source: Compustat.

Table 2 Estimating a Linear Trend in Starbucks Sales

Regression Statistics

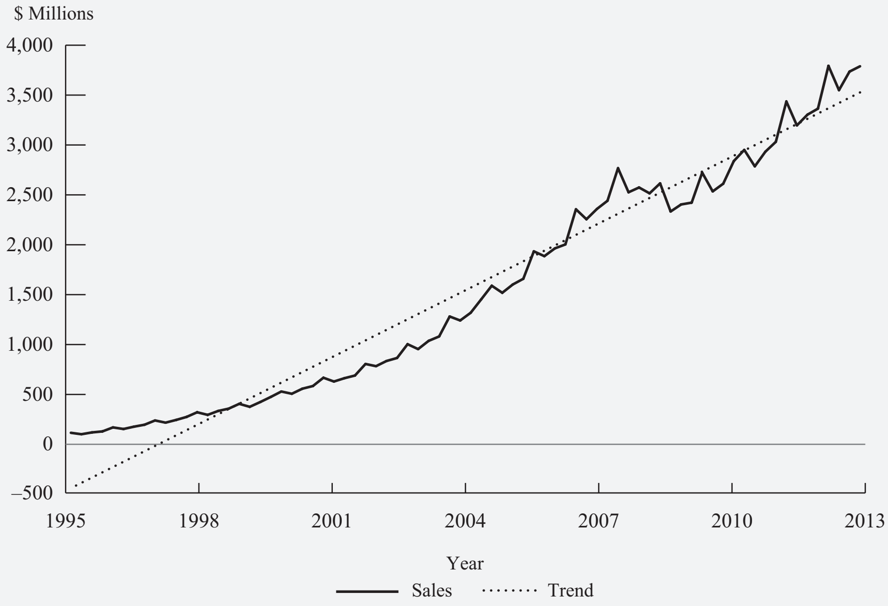
R-squared	0.9595
Standard error	233.21
Observations	76
Durbin–Watson	0.32

	Coefficient	Standard Error	t-Statistic
Intercept	−428.5380	54.0345	−7.9308
t (Trend)	51.0866	1.2194	41.8949

Source: Compustat.

At first glance, the results shown in Table 2 seem quite reasonable: Both the intercept and the trend coefficient are highly statistically significant. When Benedict plots the data on Starbucks' sales and the trend line, however, he sees a different picture. As Figure 6 shows, before 1998 the trend line is persistently below sales. Subsequently, until 2006, the trend line is persistently above sales and then varies somewhat thereafter.

Figure 6 Starbucks Quarterly Sales with Trend

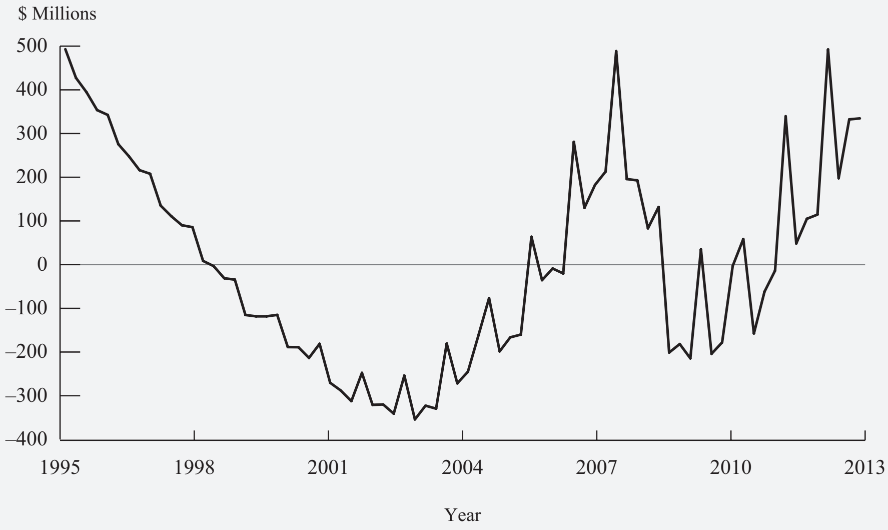


Source: Compustat.

Recall a key assumption underlying the regression model: that the regression errors are not correlated across observations. If a trend is persistently above or below the value of the time series, however, the residuals (the difference between the time series and the trend) are serially correlated. Figure 7 shows the residuals (the difference between sales and the trend) from estimating a linear trend model with the raw sales data. The figure shows that the residuals are persistent: they are consistently positive from 1995 to 1998, 2007 to 2009, and after 2012 and consistently negative from 1999 to 2006.

Because of this persistent serial correlation in the errors of the trend model, using a linear trend to fit sales at Starbucks would be inappropriate, even though the R^2 of the equation is high (0.96). The assumption of uncorrelated residual errors has been violated. Because the dependent and independent variables are not distinct, as in cross-sectional regressions, this assumption violation is serious and causes us to search for a better model.

Figure 7 Residual from Predicting Starbucks Sales with a Trend



Source: Compustat.

EXAMPLE 3

A Log-Linear Regression for Quarterly Sales at Starbucks

Having rejected a linear trend model in Example 2, technology analyst Benedict now tries a different model for the quarterly sales for Starbucks Corporation from the first quarter of 1995 to the fourth quarter of 2013. The curvature in the data plot shown in Figure 5 is a hint that an exponential curve may fit the data. Consequently, he estimates the following linear equation:

$$\ln y_t = b_0 + b_1 t + \varepsilon_t,\ t = 1, 2, \ldots, 76$$

This equation seems to fit the sales data well. As Table 3 shows, the R^2 for this equation is 0.95. An R^2 of 0.95 means that 95 percent of the variation in the natural log of Starbucks' sales is explained solely by a linear trend.

Table 3 Estimating a Linear Trend in Lognormal Starbucks Sales

Regression Statistics	
R-squared	0.9453
Standard error	0.2480
Observations	76
Durbin–Watson	0.12

Table 3 (Continued)

	Coefficient	Standard Error	*t*-Statistic
Intercept	5.1304	0.0575	89.2243
t (Trend)	0.0464	0.0013	35.6923

Source: Compustat.

Although both Equations 1 and 3 have a high R^2, Figure 8 shows how well a linear trend fits the natural log of Starbucks' sales (Equation 3). The natural logs of the sales data lie very close to the linear trend during the sample period, and log sales are not substantially above or below the trend for long periods of time. Thus, a log-linear trend model seems better suited for modeling Starbucks' sales than does a linear trend model.

1 Benedict wants to use the results of estimating Equation 3 to predict Starbucks' sales in the future. What is the predicted value of Starbucks' sales for the first quarter of 2014?

Solution to 1:

The estimated value $\hat{b}_0$ is 5.1304, and the estimated value $\hat{b}_1$ is 0.0464. Therefore, for the first quarter of 2014 (t = 77), the estimated model predicts that $\ln \hat{y}_{77}$ = 5.1304 + 0.0464(77) = 8.7032 and that predicted sales are $\hat{y} = e^{\ln \hat{y}_{77}} = e^{8.7032}$ = $6,022.15 million.[8]

8 Note that $\hat{b}_1$ = 0.0464 implies that the exponential growth rate per quarter in Starbucks' sales will be 4.75 percent ($e^{0.0464} - 1 = 0.0475$).

Figure 8 Natural Log of Starbucks Quarterly Sales

Source: Compustat.

2 How much different is the above forecast from the prediction of the linear trend model?

Solution to 2:

Table 2 showed that for the linear trend model, the estimated value of $\hat{b}_0$ is −428.5380 and the estimated value of $\hat{b}_1$ is 51.0866. Thus, if we predict Starbucks' sales for the first quarter of 2014 (t = 77) using the linear trend model, the forecast is $\hat{y}_{77}$ = −428.5380 + 51.0866(77) = $3,505.13 million. This forecast is far below the prediction made by the log-linear regression model. Later in this reading, we will examine whether we can build a better model of Starbucks' quarterly sales than a model that uses only a log-linear trend.

3.3 Trend Models and Testing for Correlated Errors

Both the linear trend model and the log-linear trend model are single-variable regression models. If they are to be correctly specified, the regression-model assumptions must be satisfied. In particular, the regression error for one period must be uncorrelated with the regression error for all other periods.[9] In Example 2 in the previous section, we could infer an obvious violation of that assumption from a visual inspection of a plot of residuals (Figure 7). The log-linear trend model of Example 3 appeared to fit the

9 Note that time-series observations, in contrast to cross-sectional observations, have a logical ordering: They must be processed in chronological order of the time periods involved. For example, we should not make a prediction of the inflation rate using a CPI series in which the order of the observations had been scrambled, because time patterns such as growth in the independent variables can negatively affect the statistical properties of the estimated regression coefficients.

data much better, but we still need to confirm that the uncorrelated errors assumption is satisfied. To address that question formally, we must carry out a Durbin–Watson test on the residuals.

In the reading on regression analysis, we showed how to test whether regression errors are serially correlated using the Durbin–Watson statistic. For example, if the trend models shown in Examples 1 and 3 really capture the time-series behavior of inflation and the log of Starbucks' sales, then the Durbin–Watson statistic for both of those models should not differ significantly from 2.0. Otherwise, the errors in the model are either positively or negatively serially correlated, and that correlation can be used to build a better forecasting model for those time series.

In Example 1, estimating a linear trend in the monthly CPI inflation yielded a Durbin–Watson statistic of 1.09. Is this result significantly different from 2.0? To find out, we need to test the null hypothesis of no positive serial correlation. For a sample with 228 observations and one independent variable, the critical value, d_l, for the Durbin–Watson test statistic at the 0.05 significance level is above 1.77. Because the value of the Durbin–Watson statistic (1.09) is below this critical value, we can reject the hypothesis of no positive serial correlation in the errors. We can conclude that a regression equation that uses a linear trend to model inflation has positive serial correlation in the errors.[10] We will need a different kind of regression model because this one violates the least-squares assumption of no serial correlation in the errors.

In Example 3, estimating a linear trend with the natural logarithm of sales for the Starbucks example yielded a Durbin–Watson statistic of 0.12. Suppose we wish to test the null hypothesis of no positive serial correlation. The critical value, d_l, is above 1.60 at the 0.05 significance level. The value of the Durbin–Watson statistic (0.12) is below this critical value, so we can reject the null hypothesis of no positive serial correlation in the errors. We can conclude that a regression equation that uses a trend to model the log of Starbucks' quarterly sales has positive serial correlation in the errors. So, for this series as well, we need to build a different kind of model.

Overall, we conclude that the trend models sometimes have the limitation that errors are serially correlated. Existence of serial correlation suggests that we can build better forecasting models for such time series than trend models.

AUTOREGRESSIVE (AR) TIME-SERIES MODELS

4

A key feature of the log-linear model's depiction of time series and a key feature of time series in general is that current-period values are related to previous-period values. For example, Starbucks' sales for the current period are related to its sales in the previous period. An **autoregressive model (AR)**, a time series regressed on its own past values, represents this relationship effectively. When we use this model, we can drop the normal notation of y as the dependent variable and x as the independent variable because we no longer have that distinction to make. Here we simply use x_t. For example, Equation 4 shows a first-order autoregression, AR(1), for the variable x_t:

$$x_t = b_0 + b_1 x_{t-1} + \varepsilon_t \tag{4}$$

10 Significantly small values of the Durbin–Watson statistic indicate positive serial correlation; significantly large values point to negative serial correlation. Here the DW statistic of 1.09 indicates positive serial correlation. For more information, see the readings on regression analysis.

Thus, in an AR(1) model, we use only the most recent past value of x_t to predict the current value of x_t. In general, a pth-order autoregression, AR(p), for the variable x_t is shown by

$$x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-2} + \ldots + b_p x_{t-p} + \varepsilon_t \quad (5)$$

In this equation, p past values of x_t are used to predict the current value of x_t. In the next section we discuss a key assumption of time-series models that include lagged values of the dependent variable as independent variables.

4.1 Covariance-Stationary Series

Note that the independent variable (x_{t-1}) in Equation 4 is a random variable. This fact may seem like a mathematical subtlety, but it is not. If we use ordinary least squares to estimate Equation 4 when we have a randomly distributed independent variable that is a lagged value of the dependent variable, our statistical inference may be invalid. To conduct valid statistical inference, we must make a key assumption in time-series analysis: We must assume that the time series we are modeling is **covariance stationary**.[11]

What does it mean for a time series to be covariance stationary? The basic idea is that a time series is covariance stationary if its properties, such as mean and variance, do not change over time. A covariance stationary series must satisfy three principal requirements.[12] First, the expected value of the time series must be constant and finite in all periods: $E(y_t) = \mu$ and $|\mu| < \infty$, $t = 1, 2, \ldots, T$. Second, the variance of the time series must be constant and finite in all periods. Third, the covariance of the time series with itself for a fixed number of periods in the past or future must be constant and finite in all periods. The second and third requirements can be summarized as follows:[13]

$$\text{Cov}(y_t, y_{t-s}) = \lambda_s,\ |\lambda_s| < \infty,\ t = 1, 2, \ldots, T;\ s = 0, \pm 1, \pm 2, \ldots, \pm T$$

where λ signifies a constant. What happens if a time series is not covariance stationary but we model it using Equation 4? The estimation results will have no economic meaning. For a non-covariance-stationary time series, estimating the regression in Equation 4 will yield spurious results. In particular, the estimate of b_1 will be biased, and any hypothesis tests will be invalid.

How can we tell if a time series is covariance stationary? We can often answer this question by looking at a plot of the time series. If the plot shows roughly the same mean and variance through time without any significant seasonality, then we may want to assume that the time series is covariance stationary.

Some of the time series we looked at in Figures 1 to 4 appear to be covariance stationary. For example, the inflation data shown in Figure 3 appear to have roughly the same mean and variance over the sample period. Many of the time series one encounters in business and investments, however, are not covariance stationary. For example, many time series appear to grow (or decline) steadily through time and so have a mean that is nonconstant, which implies that they are nonstationary. As an example, the time series of quarterly sales in Figure 5 clearly shows the mean

11 "Weakly stationary" is a synonym for covariance stationary. Note that the terms "stationary" or "stationarity" are often used to mean "covariance stationary" or "covariance stationarity," respectively. You may also encounter the more restrictive concept of "strictly" stationary, which has little practical application. For details, see Diebold (2008).

12 In the first requirement, we will use the absolute value to rule out the case in which the mean is negative without limit (minus infinity).

13 When s in this equation equals 0, then this equation imposes the condition that the variance of the time series is finite. This is so because the covariance of a random variable with itself is its variance: $\text{Cov}(y_t, y_t) = \text{Var}(y_t)$.

increasing as time passes. Thus Starbucks' quarterly sales are not covariance stationary.[14] Macroeconomic time series such as those relating to income and consumption are often strongly trending as well. A time series with seasonality (regular patterns of movement with the year) also has a nonconstant mean, as do other types of time series that we discuss later.[15]

Figure 2 showed that monthly retail sales (not seasonally adjusted) are also not covariance stationary. Sales in December are always much higher than sales in other months (these are the regular large peaks), and sales in January are always much lower (these are the regular large drops after the December peaks). On average, sales also increase over time, so the mean of sales is not constant.

Later in the reading, we will show that we can often transform a nonstationary time series into a stationary time series. But whether a stationary time series is original or transformed, a caution applies: Stationarity in the past does not guarantee stationarity in the future. There is always the possibility that a well-specified model will fail when the state of the world changes and yields a different underlying model that generates the time series.

4.2 Detecting Serially Correlated Errors in an Autoregressive Model

We can estimate an autoregressive model using ordinary least squares if the time series is covariance stationary and the errors are uncorrelated. Unfortunately, our previous test for serial correlation, the Durbin–Watson statistic, is invalid when the independent variables include past values of the dependent variable. Therefore, for most time-series models, we cannot use the Durbin–Watson statistic. Fortunately, we can use other tests to determine whether the errors in a time-series model are serially correlated. One such test reveals whether the autocorrelations of the error term are significantly different from 0. This test is a *t*-test involving a residual autocorrelation and the standard error of the residual autocorrelation. As background for the test, we next discuss autocorrelation in general before moving to residual autocorrelation.

The **autocorrelations** of a time series are the correlations of that series with its own past values. The order of the correlation is given by k where k represents the number of periods lagged. When $k = 1$, the autocorrelation shows the correlation of the variable in one period to its occurrence in the previous period. For example, the ***k*th order autocorrelation** (ρ_k) is

$$\rho_k = \frac{\mathrm{Cov}(x_t, x_{t-k})}{\sigma_x^2} = \frac{E\left[(x_t - \mu)(x_{t-k} - \mu)\right]}{\sigma_x^2}$$

where E stands for the expected value. Note that we have the relationship $\mathrm{Cov}(x_t, x_{t-k}) \le \sigma_x^2$ with equality holding when $k = 0$. This means that the absolute value of ρ_k is less than or equal to 1.

14 In general, any time series accurately described with a linear or log-linear trend model is not covariance stationary, although a transformation of the original series might be covariance stationary.

15 In particular, random walks are not covariance stationary.

Of course, we can never directly observe the autocorrelations, ρ_k. Instead, we must estimate them. Thus we replace the expected value of x_t, μ, with its estimated value, $\overline{x}$, to compute the estimated autocorrelations. The kth order estimated autocorrelation of the time series x_t, which we denote $\hat{\rho}_k$, is

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T}\left[(x_t - \overline{x})(x_{t-k} - \overline{x})\right]}{\sum_{t=1}^{T}(x_t - \overline{x})^2}$$

Analogous to the definition of autocorrelations for a time series, we can define the autocorrelations of the error term for a time-series model as[16]

$$\begin{aligned}\rho_{\varepsilon,k} &= \frac{\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-k})}{\sigma_\varepsilon^2} \\ &= \frac{E\left[(\varepsilon_t - 0)(\varepsilon_{t-k} - 0)\right]}{\sigma_\varepsilon^2} \\ &= \frac{E(\varepsilon_t \varepsilon_{t-k})}{\sigma_\varepsilon^2}\end{aligned}$$

We assume that the expected value of the error term in a time-series model is 0.[17]

We can determine whether we are using the correct time-series model by testing whether the autocorrelations of the error term (**error autocorrelations**) differ significantly from 0. If they do, the model is not specified correctly. We estimate the error autocorrelation using the sample autocorrelations of the residuals (**residual autocorrelations**) and their sample variance.

A test of the null hypothesis that an error autocorrelation at a specified lag equals 0 is based on the residual autocorrelation for that lag and the standard error of the residual correlation, which is equal to $1/\sqrt{T}$, where T is the number of observations in the time series.[18] Thus, if we have 100 observations in a time series, the standard error for each of the estimated autocorrelations is 0.1. We can compute the t-test of the null hypothesis that the error correlation at a particular lag equals 0, by dividing the residual autocorrelation at that lag by its standard error $\left(1/\sqrt{T}\right)$.

How can we use information about the error autocorrelations to determine whether an autoregressive time-series model is correctly specified? We can use a simple three-step method. First, estimate a particular autoregressive model, say an AR(1) model. Second, compute the autocorrelations of the residuals from the model.[19] Third, test to see whether the residual autocorrelations differ significantly from 0. If significance tests show that the residual autocorrelations differ significantly from 0, the model is not correctly specified; we may need to modify it in ways that we will discuss shortly.[20] We now present an example to demonstrate how this three-step method works.

---

16 Whenever we refer to autocorrelation without qualification, we mean autocorrelation of the time series itself rather than autocorrelation of the error term or residuals.

17 This assumption is similar to the one made in the previous two readings about the expected value of the error term.

18 This calculation is derived in Diebold (2008).

19 We can compute these residual autocorrelations easily with most statistical software packages. In Microsoft Excel, for example, to compute the first-order residual autocorrelation, we compute the correlation of the residuals from observations 1 through $T - 1$ with the residuals from observations 2 through T.

20 Often, econometricians use additional tests for the significance of residual autocorrelations. For example, the Box–Pierce Q-statistic is frequently used to test the joint hypothesis that all autocorrelations of the residuals are equal to 0. For further discussion, see Diebold (2008).

EXAMPLE 4

Predicting Gross Margins for Intel Corporation

Analyst Melissa Jones decides to use a time-series model to predict Intel Corporation's gross margin [(Sales − Cost of goods sold)/Sales] using quarterly data from the second quarter of 1999 through the fourth quarter of 2013. She does not know the best model for gross margin but believes that the current-period value will be related to the previous-period value. She decides to start out with a first-order autoregressive model, AR(1): Gross margin$_t$ = b_0 + b_1(Gross margin$_{t-1}$) + ε_t. Her observations on the dependent variable are 2Q:1999 through 4Q:2013. Table 4 shows the results of estimating this AR(1) model, along with the autocorrelations of the residuals from that model.

Table 4 Autoregression: AR(1) Model Gross Margin of Intel Quarterly Observations, April 1999–December 2013

Regression Statistics	
R-squared	0.5429
Standard error	0.0337
Observations	59
Durbin–Watson	2.0987

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.1795	0.0635	2.8268
Gross margin$_{t-1}$	0.7449	0.0905	8.2309

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	−0.0495	0.1302	−0.3802
2	−0.0392	0.1302	−0.3011
3	0.0524	0.1302	0.4025
4	0.1450	0.1302	1.1137

Source: Compustat.

The first thing to note about Table 4 is that both the intercept ($\hat{b}_0$ = 0.1795) and the coefficient on the first lag ($\hat{b}_1$ = 0.7449) of the gross margin are highly significant in the regression equation.[21] The *t*-statistic for the intercept is about 2.8, whereas the *t*-statistic for the first lag of the gross margin is more than 8. With 59 observations and two parameters, this model has 57 degrees of freedom. At the 0.05 significance level, the critical value for a *t*-statistic is about 2.0. Therefore, Jones must reject the null hypotheses that the intercept is equal to 0 (b_0 = 0) and the coefficient on the first lag is equal to 0 (b_1 = 0) in favor of the alternative hypothesis that the coefficients, individually, are not equal to 0. But are these statistics valid? Although the Durbin–Watson statistic is presented in

21 The first lag of a time series is the value of the time series in the previous period.

Table 4, it cannot be used to test serial correlation when the independent variables include past values of the dependent variable. The correct approach is to test whether the residuals from this model are serially correlated.

At the bottom of Table 4, the first four autocorrelations of the residual are displayed along with the standard error and the t-statistic for each of those autocorrelations.[22] The sample has 59 observations, so the standard error for each of the autocorrelations is $1/\sqrt{59} = 0.1302$. Table 4 shows that none of the first four autocorrelations has a t-statistic larger than 1.1137 in absolute value. Therefore, Jones can conclude that none of these autocorrelations differs significantly from 0. Consequently, she can assume that the residuals are not serially correlated and that the model is correctly specified, and she can validly use ordinary least squares to estimate the parameters and the parameters' standard errors in the autoregressive model.[23]

Now that Jones has concluded that this model is correctly specified, how can she use it to predict Intel's gross margin in the next period? The estimated equation is Gross margin$_t$ = 0.1795 + 0.7449(Gross margin$_{t-1}$) + ε_t. The expected value of the error term is 0 in any period. Therefore, this model predicts that gross margin in period $t + 1$ will be Gross margin$_{t+1}$ = 0.1795 + 0.7449(Gross margin$_t$). For example, if gross margin is 65 percent in this quarter (0.65), the model predicts that in the next quarter gross margin will increase to 0.1795 + 0.7449(0.65) = 0.6637 or 66.37 percent. On the other hand, if gross margin is currently 75 percent (0.75), the model predicts that in the next quarter, gross margin will fall to 0.1795 + 0.7449(0.75) = 0.7382 or 73.82 percent. As we show in the following section, the model predicts that gross margin will increase if it is below a certain level (70.36 percent) and decrease if it is above that level.

4.3 Mean Reversion

We say that a time series shows **mean reversion** if it tends to fall when its level is above its mean and rise when its level is below its mean. Much like the temperature in a room controlled by a thermostat, a mean-reverting time series tends to return to its long-term mean. How can we determine the value that the time series tends toward? If a time series is currently at its mean-reverting level, then the model predicts that the value of the time series will be the same in the next period. At its mean-reverting level, we have the relationship $x_{t+1} = x_t$. For an AR(1) model ($x_{t+1} = b_0 + b_1x_t$), the equality $x_{t+1} = x_t$ implies the level $x_t = b_0 + b_1x_t$, or that the mean-reverting level, x_t, is given by

$$x_t = \frac{b_0}{1 - b_1}$$

So the AR(1) model predicts that the time series will stay the same if its current value is $b_0/(1 - b_1)$, increase if its current value is below $b_0/(1 - b_1)$, and decrease if its current value is above $b_0/(1 - b_1)$.

In the case of gross margins for Intel, the mean-reverting level for the model shown in Table 4 is 0.1795/(1 – 0.7449) = 0.7036. If the current gross margin is above 0.7036, the model predicts that the gross margin will fall in the next period. If the current

22 For seasonally unadjusted data, analysts often compute the same number of autocorrelations as there are observations in a year (for example, four for quarterly data). The number of autocorrelations computed also often depends on sample size, as discussed in Diebold (2007).

23 Statisticians have many other tests for serial correlation of the residuals in a time-series model. For details, see Diebold (2008).

gross margin is below 0.7036, the model predicts that the gross margin will rise in the next period. As we will discuss later, all covariance-stationary time series have a finite mean-reverting level.

4.4 Multiperiod Forecasts and the Chain Rule of Forecasting

Often, financial analysts want to make forecasts for more than one period. For example, we might want to use a quarterly sales model to predict sales for a company for each of the next four quarters. To use a time-series model to make forecasts for more than one period, we must examine how to make multiperiod forecasts using an AR(1) model. The one-period-ahead forecast of x_t from an AR(1) model is as follows:

$$\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1 x_t \tag{6}$$

If we want to forecast x_{t+2} using an AR(1) model, our forecast will be based on

$$\hat{x}_{t+2} = \hat{b}_0 + \hat{b}_1 x_{t+1} \tag{7}$$

Unfortunately, we do not know x_{t+1} in period *t*, so we cannot use Equation 7 directly to make a two-period-ahead forecast. We can, however, use our forecast of x_{t+1} and the AR(1) model to make a prediction of x_{t+2}. The **chain rule of forecasting** is a process in which the next period's value, predicted by the forecasting equation, is substituted into the equation to give a predicted value two periods ahead. Using the chain rule of forecasting, we can substitute the predicted value of x_{t+1} into Equation 7 to get $\hat{x}_{t+2} = \hat{b}_0 + \hat{b}_1 \hat{x}_{t+1}$. We already know $\hat{x}_{t+1}$ from our one-period-ahead forecast in Equation 6. Now we have a simple way of predicting x_{t+2}.

Multiperiod forecasts are more uncertain than single-period forecasts because each forecast period has uncertainty. For example, in forecasting x_{t+2}, we first have the uncertainty associated with forecasting x_{t+1} using x_t, and then we have the uncertainty associated with forecasting x_{t+2} using the forecast of x_{t+1}. In general, the more periods a forecast has, the more uncertain it is.[24]

EXAMPLE 5

Multiperiod Prediction of Intel's Gross Margin

Suppose that at the beginning of 2014, we want to predict Intel's gross margin in two periods using the model shown in Table 4. Assume that Intel's gross margin in the current period is 65 percent. The one-period-ahead forecast of Intel's gross margin from this model is 0.6637 = 0.1795 + 0.7449(0.65). By substituting the one-period-ahead forecast, 0.6637, back into the regression equation, we can derive the following two-period-ahead forecast: 0.6739 = 0.1795 + 0.7449(0.6637). Therefore, if the current gross margin for Intel is 65 percent, the model predicts that Intel's gross margin in two quarters will be 67.39 percent.

24 If a forecasting model is well specified, the prediction errors from the model will not be serially correlated. If the prediction errors for each period are not serially correlated, then the variance of a multiperiod forecast will be higher than the variance of a single-period forecast.

EXAMPLE 6

Modeling US CPI Inflation

Analyst Lisette Miller has been directed to build a time-series model for monthly US inflation. Inflation and expectations about inflation, of course, have a significant effect on bond returns. For a 30-year period beginning with January 1984 and ending with December 2013, she selects as data the annualized monthly percentage change in the CPI. Which model should Miller use?

The process of model selection parallels that of Example 4 relating to Intel's gross margins. The first model Miller estimates is an AR(1) model, using the previous month's inflation rate as the independent variable: $\text{Inflation}_t = b_0 + b_1 \cdot \text{Inflation}_{t-1} + \varepsilon_t$, $t = 1, 2, \ldots, 359$. To estimate this model, she uses monthly CPI inflation data from January 1984 to December 2013 (t = 1 denotes February 1984). Table 5 shows the results of estimating this model.

Table 5 Monthly CPI Inflation at an Annual Rate: AR(1) Model Monthly Observations, February 1984–December 2013

Regression Statistics	
R-squared	0.2038
Standard error	3.4250
Observations	359
Durbin–Watson	1.8201

	Coefficient	Standard Error	t-Statistic
Intercept	1.5703	0.2266	6.9298
Inflation_{t-1}	0.4510	0.0472	9.5551

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	t-Statistic
1	0.0898	0.0528	1.7008
2	−0.1205	0.0528	−2.2822
3	−0.1571	0.0528	−2.9754
4	−0.0316	0.0528	−0.5985

Source: US Bureau of Labor Statistics.

As Table 5 shows, both the intercept ($\hat{b}_0$ = 1.5703) and the coefficient on the first lagged value of inflation ($\hat{b}_1$ = 0.4510) are highly statistically significant, with large t-statistics. With 359 observations and two parameters, this model has 357 degrees of freedom. The critical value for a t-statistic at the 0.05 significance level is about 1.97. Therefore, Miller can reject the individual null hypotheses that the intercept is equal to 0 (b_0 = 0) and the coefficient on the first lag is equal to 0 (b_1 = 0) in favor of the alternative hypothesis that the coefficients, individually, are not equal to 0.

Are these statistics valid? Miller will know when she tests whether the residuals from this model are serially correlated. With 359 observations in this sample, the standard error for each of the estimated autocorrelations is $1/\sqrt{359} = 0.0528$.

The critical value for the t-statistic is 1.97. Because both the second and the third estimated autocorrelation have t-statistics larger than 1.97 in absolute value, Miller concludes that the autocorrelations are significantly different from 0. This model is thus misspecified because the residuals are serially correlated.

If the residuals in an autoregressive model are serially correlated, Miller can eliminate the correlation by estimating an autoregressive model with more lags of the dependent variable as explanatory variables. Table 6 shows the result of estimating a second time-series model, an AR(2) model using the same data as in the analysis shown in Table 5.[25] With 358 observations and three parameters, this model has 355 degrees of freedom. Because the degrees of freedom are almost the same as those for the estimates shown in Table 5, the critical value of the t-statistic at the 0.05 significance level also is almost the same (1.97). If she estimates the equation with two lags, $\text{Inflation}_t = b_0 + b_1 \text{Inflation}_{t-1} + b_2 \text{Inflation}_{t-2} + \varepsilon_t$, Miller finds that all three of the coefficients in the regression model (an intercept and the coefficients on two lags of the dependent variable) differ significantly from 0. The bottom portion of Table 6 shows that none of the first four autocorrelations of the residual has a t-statistic greater in absolute value than the critical value of 1.97. Therefore, Miller fails to reject the hypothesis that the individual autocorrelations of the residual equal 0. She concludes that this model is correctly specified because she finds no evidence of serial correlation in the residuals.

Table 6 Monthly CPI Inflation at an Annual Rate: AR(2) Model Monthly Observations, March 1984–December 2013

Regression Statistics	
R-squared	0.2349
Standard error	3.3637
Observations	358
Durbin–Watson	2.0273

	Coefficient	Standard Error	t-Statistic
Intercept	1.8953	0.2379	7.9668
Inflation_{t-1}	0.5406	0.0520	10.3962
Inflation_{t-2}	−0.2015	0.0520	−3.8750

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	t-Statistic
1	−0.0140	0.0529	−0.2647
2	0.0335	0.0529	0.6333

(continued)

25 Note that Table 6 shows only 358 observations in the regression because the extra lag of inflation requires the estimation sample to start one month later than the regression in Table 5. (With two lags, inflation for January and February 1984 must be known in order to estimate the equation starting in March 1984.)

Table 6 (Continued)

Autocorrelations of the Residual			
Lag	Autocorrelation	Standard Error	*t*-Statistic
3	−0.0726	0.0529	−1.3724
4	−0.0056	0.0529	−0.1059

Source: US Bureau of Labor Statistics.

1 The analyst selected an AR(2) model because the residuals from the AR(1) model were serially correlated. Suppose that in a given month, inflation had been 4 percent at an annual rate in the previous month and 3 percent in the month before that. What would be the difference in the analyst forecast of the inflation for that month if she had used an AR(1) model instead of the AR(2) model?

Solution to 1:

The AR(1) model shown in Table 5 predicted that inflation in the next month would be 1.5703 + 0.4510(4) = 3.37 percent approximately, whereas the AR(2) model shown in Table 6 predicts that inflation in the next month will be 1.8953 + 0.5406(4) − 0.2015(3) = 3.45 percent approximately. If the analyst had used the incorrect AR(1) model, she would have predicted inflation to be 8 basis points lower (3.37 percent versus 3.45 percent) than using the AR(2) model. This incorrect forecast could have adversely affected the quality of her company's investment choices.

4.5 Comparing Forecast Model Performance

One way to compare the forecast performance of two models is to compare the variance of the forecast errors that the two models make. The model with the smaller forecast error variance will be the more accurate model, and it will also have the smaller standard error of the time-series regression. (This standard error usually is reported directly in the output for the time-series regression.)

In comparing forecast accuracy among models, we must distinguish between in-sample forecast errors and out-of-sample forecast errors. **In-sample forecast errors** are the residuals from a fitted time-series model. For example, when we estimated a linear trend with raw inflation data from January 1984 to December 2013, the in-sample forecast errors were the residuals from January 1984 to December 2013. If we use this model to predict inflation outside this period, the differences between actual and predicted inflation are **out-of-sample forecast errors**.

EXAMPLE 7

In-Sample Forecast Comparisons of US CPI Inflation

In Example 6, the analyst compared an AR(1) forecasting model of monthly US inflation with an AR(2) model of monthly US inflation and decided that the AR(2) model was preferable. Table 5 showed that the standard error from the AR(1) model of inflation is 3.4250, and Table 6 showed that the standard error from the AR(2) model is 3.3637. Therefore, the AR(2) model had a lower in-sample forecast error variance than the AR(1) model, which is consistent with our belief that the AR(2) model was preferable. Its standard error is 3.3637/3.4250 = 98.21 percent of the forecast error of the AR(1) model.

Often, we want to compare the forecasting accuracy of different models after the sample period for which they were estimated. We wish to compare the out-of-sample forecast accuracy of the models. Out-of-sample forecast accuracy is important because the future is always out of sample. Although professional forecasters distinguish between out-of-sample and in-sample forecasting performance, many articles that analysts read contain only in-sample forecast evaluations. Analysts should be aware that out-of-sample performance is critical for evaluating a forecasting model's real-world contribution.

Typically, we compare the out-of-sample forecasting performance of forecasting models by comparing their **root mean squared error (RMSE)**, which is the square root of the average squared error. The model with the smallest RMSE is judged most accurate. The following example illustrates the computation and use of RMSE in comparing forecasting models.

EXAMPLE 8

Out-of-Sample Forecast Comparisons of US CPI Inflation

Suppose we want to compare the forecasting accuracy of the AR(1) and AR(2) models of US inflation estimated over 1984 to 2013, using data on US inflation from January 2014 to September 2014.

Table 7 Out-of-Sample Forecast Error Comparisons: January 2014–September 2014 US CPI Inflation (Annualized)

Date	Infl(t)	Infl(t–1)	Infl(t–2)	AR(1) Error	Squared Error	AR(2) Error	Squared Error
2014							
January	4.5568	–0.1029	–2.4236	3.0329	9.1986	2.2288	4.9675
February	4.5289	4.5568	–0.1029	0.9034	0.8162	0.1494	0.0223
March	8.0077	4.5289	4.5568	4.3949	19.3154	4.5823	20.9978
April	4.0286	8.0077	4.5289	–1.1532	1.3298	–1.2831	1.6463
May	4.2726	4.0286	8.0077	0.8854	0.7839	1.8130	3.2869
June	2.2576	4.2726	4.0286	–1.2397	1.5368	–1.1357	1.2898
July	–0.4672	2.2576	4.2726	–3.0557	9.3373	–2.7221	7.4096
August	–1.9863	–0.4672	2.2576	–3.3459	11.1949	–3.1741	10.0750
September	0.9068	–1.9863	–0.4672	0.2324	0.0540	–0.0088	0.0001
				Average	5.9519	Average	5.5217
				RMSE	2.4396	RMSE	2.3498

Note: Any apparent discrepancies between error and squared error results are due to rounding.
Source: US Bureau of Labor Statistics.

For each month from January 2014 to September 2014, the first column of numbers in Table 7 shows the actual annualized inflation rate during the month. The second and third columns show the rate of inflation in the previous two months. The fourth column shows the out-of-sample errors (Actual – Forecast) from the AR(1) model shown in Table 5. The fifth column shows the squared errors from the AR(1) model. The sixth column shows the out-of-sample errors from the AR(2) model shown in Table 6. The final column shows the squared errors from the AR(2) model. The bottom of the table displays the average squared error and the RMSE. According to these measures, the AR(2) model was slightly more accurate than the AR(1) model in its out-of-sample forecasts of inflation from January 2014 to September 2014. The RMSE from the AR(2) model was only 2.3498/2.4396 = 96.32 percent as large as the RMSE from the AR(1) model. Therefore, the AR(2) model was more accurate both in-sample and out of sample. Of course, this was a small sample to use in evaluating out-of-sample forecasting performance. Sometimes, an analyst may have conflicting information about whether to choose an AR(1) or an AR(2) model. We must also consider regression coefficient stability. We will continue the comparison between these two models in the following section.

4.6 Instability of Regression Coefficients

One of the important issues an analyst faces in modeling a time series is the sample period to use. The estimates of regression coefficients of the time-series model can change substantially across different sample periods used for estimating the model. Often, the regression coefficient estimates of a time-series model estimated using an earlier sample period can be quite different from those of a model estimated using a later sample period. Similarly, the estimates can be different between models estimated using relatively shorter and longer sample periods. Further, the choice of model for a particular time series can also depend on the sample period. For example, an AR(1) model may be appropriate for the sales of a company in one particular sample period, but an AR(2) model may be necessary for an earlier or later sample period (or for a longer or shorter sample period). Thus the choice of a sample period is an important decision in modeling a financial time series.

Unfortunately, there is usually no clear-cut basis in economic or financial theory for determining whether to use data from a longer or shorter sample period to estimate a time-series model. We can get some guidance, however, if we remember that our models are valid only for covariance-stationary time series. For example, we should not combine data from a period when exchange rates were fixed with data from a period when exchange rates were floating. The exchange rates in these two periods would not likely have the same variance because exchange rates are usually much more volatile under a floating-rate regime than when rates are fixed. Similarly, many US analysts consider it inappropriate to model US inflation or interest-rate behavior since the 1960s as a part of one sample period, because the Federal Reserve had distinct policy regimes during this period. A simple way to determine appropriate samples for time-series estimation is to look at graphs of the data to see if the time series looks stationary before estimation begins. If we know that a government policy changed on a specific date, we might also test whether the time-series relation was the same before and after that date.

In the following example, we illustrate how the choice of a longer versus a shorter period can affect the decision of whether to use, for example, a first- or second-order time-series model. We then show how the choice of the time-series model (and the associated regression coefficients) affects our forecast. Finally, we discuss which sample period, and accordingly which model and corresponding forecast, is appropriate for the time series analyzed in the example.

EXAMPLE 9

Instability in Time-Series Models of US Inflation

In Example 6, analyst Lisette Miller concluded that US CPI inflation should be modeled as an AR(2) time series. A colleague examined her results and questioned estimating one time-series model for inflation in the United States since 1984, given that the Federal Reserve responded aggressively to the financial crisis that emerged in 2007. He argues that the inflation time series from 1984 to 2013 has two **regimes** or underlying models generating the time series: one running from 1984 through 2006, and another starting in 2007. Therefore, the colleague suggests that Miller estimate a new time-series model for US inflation starting in 2007. Because of his suggestion, Miller first estimates an AR(1) model for inflation using data for a shorter sample period from 2007 to 2013. Table 8 shows her AR(1) estimates.

Table 8 Autoregression: AR(1) Model Monthly CPI Inflation at an Annual Rate, February 2007–December 2013

Regression Statistics	
R-squared	0.3070
Standard error	4.4749
Observations	83
Durbin–Watson	1.8164

	Coefficient	Standard Error	t-Statistic
Intercept	0.9585	0.5337	1.7960
Inflation$_{t-1}$	0.5544	0.0926	5.9870

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	t-Statistic
1	0.0878	0.1098	0.7999
2	−0.0091	0.1098	−0.0829
3	−0.0355	0.1098	−0.3234
4	0.0020	0.1098	0.0182

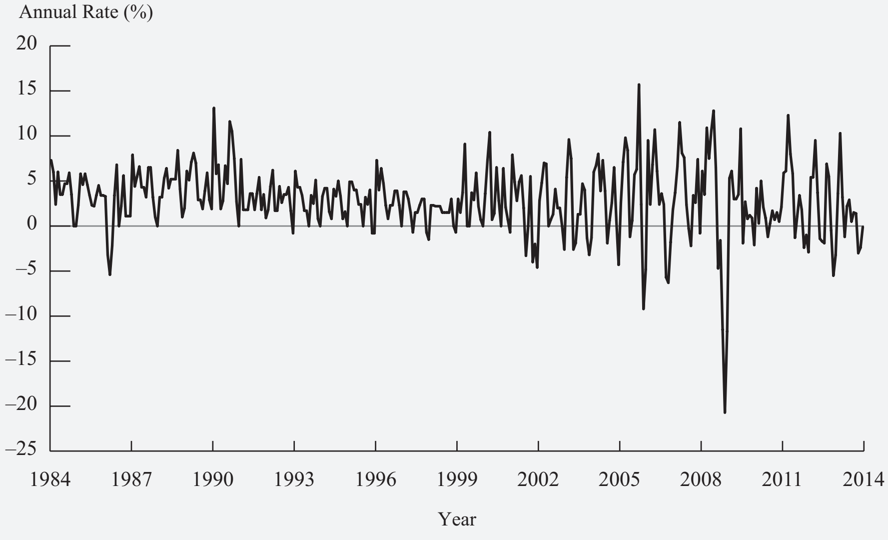
Source: US Bureau of Labor Statistics.

The bottom part of Table 8 shows that the first four autocorrelations of the residuals from the AR(1) model are quite small. None of these autocorrelations has a t-statistic larger than 1.99, the critical value for significance. Consequently, Miller cannot reject the null hypothesis that the residuals are serially uncorrelated. The AR(1) model is correctly specified for the sample period from 2007 to 2013, so there is no need to estimate the AR(2) model. This conclusion is very different from that reached in Example 6 using data from 1984 to 2013. In that example, Miller initially rejected the AR(1) model because its residuals exhibited serial correlation. When she used a larger sample, an AR(2) model initially appeared to fit the data much better than did an AR(1) model.

How deeply does our choice of sample period affect our forecast of future inflation? Suppose that in a given month, inflation was 4 percent at an annual rate, and the month before that it was 3 percent. The AR(1) model shown in Table 8 predicts that inflation in the next month will be 0.9585 + 0.5544(4) = approximately 3.17 percent. Therefore, the forecast of the next month's inflation using the 2007 to 2013 sample is 3.17 percent. Remember from the analysis following Example 6 that the AR(2) model for the 1984 to 2013 sample predicts inflation of 3.45 percent in the next month. Thus, using the correctly specified model for the shorter sample produces an inflation forecast 0.28 percentage points below the forecast made from the correctly specified model for the longer sample period. Such a difference might substantially affect a particular investment decision.

Which model is correct? Figure 9 suggests an answer. Monthly US inflation was so much more volatile during the latter part of the study period than in the earlier years that inflation is probably not a covariance-stationary time series from 1984 to 2013. Therefore, we can reasonably believe that the data have more than one regime and Miller should estimate a separate model for inflation from 2007 to 2013, as shown above. In fact, the standard deviation of annualized monthly inflation rates is just 3.24 percent for the period of 1984 to 2006 but 5.32 percent for the period of 2007 to 2013. As the example shows, experience (such as knowledge of government policy changes) and judgment play a vital role in determining how to model a time series. Simply relying on autocorrelations of the residuals from a time-series model cannot tell us the correct sample period for our analysis.

Figure 9 Monthly CPI Inflation



Source: US Bureau of Labor Statistics.

RANDOM WALKS AND UNIT ROOTS

5

So far, we have examined those time series in which the time series has a tendency to revert to its mean level as the change in a variable from one period to the next follows a mean-reverting pattern. In contrast, there are many financial time series in which the changes follow a random pattern. We discuss these "random walks" in the following section.

5.1 Random Walks

A random walk is one of the most widely studied time-series models for financial data. A **random walk** is a time series in which the value of the series in one period is the value of the series in the previous period plus an unpredictable random error. A random walk can be described by the following equation:

$$x_t = x_{t-1} + \varepsilon_t,\ E(\varepsilon_t) = 0,\ E\left(\varepsilon_t^2\right) = \sigma^2,\ \mathrm{Cov}(\varepsilon_t, \varepsilon_s) = E(\varepsilon_t \varepsilon_s) = 0 \text{ if } t \neq s \quad \textbf{(8)}$$

Equation 8 means that the time series x_t is in every period equal to its value in the previous period plus an error term, ε_t, that has constant variance and is uncorrelated with the error term in previous periods. Note two important points. First, this equation is a special case of an AR(1) model with $b_0 = 0$ and $b_1 = 1$.[26] Second, the expected value of ε_t is zero. Therefore, the best forecast of x_t that can be made in period $t - 1$ is x_{t-1}. In fact, in this model, x_{t-1} is the best forecast of x in every period after $t - 1$.

Random walks are quite common in financial time series. For example, many studies have tested and found that currency exchange rates follow a random walk. Consistent with the second point made above, some studies have found that sophisticated exchange rate forecasting models cannot outperform forecasts made using the random walk model, and that the best forecast of the future exchange rate is the current exchange rate.

Unfortunately, we cannot use the regression methods we have discussed so far to estimate an AR(1) model on a time series that is actually a random walk. To see why this is so, we must determine why a random walk has no finite mean-reverting level or finite variance. Recall that if x_t is at its mean-reverting level, then $x_t = b_0 + b_1 x_t$, or $x_t = b_0/(1 - b_1)$. In a random walk, however, $b_0 = 0$ and $b_1 = 1$, so $b_0/(1 - b_1) = 0/0$. Therefore, a random walk has an undefined mean-reverting level.

What is the variance of a random walk? Suppose that in Period 1, the value of x_1 is 0. Then we know that $x_2 = 0 + \varepsilon_2$. Therefore, the variance of $x_2 = \mathrm{Var}(\varepsilon_2) = \sigma^2$. Now $x_3 = x_2 + \varepsilon_3 = \varepsilon_2 + \varepsilon_3$. Because the error term in each period is assumed to be uncorrelated with the error terms in all other periods, the variance of $x_3 = \mathrm{Var}(\varepsilon_2) + \mathrm{Var}(\varepsilon_3) = 2\sigma^2$. By a similar argument, we can show that for any period t, the variance of $x_t = (t - 1)\sigma^2$. But this means that as t grows large, the variance of x_t grows without an upper bound: It approaches infinity. This lack of upper bound, in turn, means that a random walk is not a covariance-stationary time series, because a covariance-stationary time series must have a finite variance.

What is the practical implication of these issues? *We cannot use standard regression analysis on a time series that is a random walk.* We can, however, attempt to convert the data to a covariance-stationary time series if we suspect that the time series is a random walk. In statistical terms, we can difference it.

26 Equation 8 with a nonzero intercept added (as in Equation 9 given later) is sometimes referred to as a random walk with drift.

We difference a time series by creating a new time series, say y_t, that in each period is equal to the difference between x_t and x_{t-1}. This transformation is called **first-differencing** because it subtracts the value of the time series in the first prior period from the current value of the time series. Sometimes the first difference of x_t is written as $\Delta x_t = x_t - x_{t-1}$. Note that the first difference of the random walk in Equation 8 yields

$$y_t = x_t - x_{t-1} = \varepsilon_t,\ E(\varepsilon_t) = 0,\ E\left(\varepsilon_t^2\right) = \sigma^2,\ \mathrm{Cov}(\varepsilon_t,\varepsilon_s) = E(\varepsilon_t\varepsilon_s) = 0 \text{ for } t \neq s$$

The expected value of ε_t is 0. Therefore, the best forecast of y_t that can be made in period $t - 1$ is 0. This implies that the best forecast is that there will be no change in the value of the current time series, x_{t-1}.

The first-differenced variable, y_t, is covariance stationary. How is this so? First, note that this model ($y_t = \varepsilon_t$) is an AR(1) model with $b_0 = 0$ and $b_1 = 0$. We can compute the mean-reverting level of the first-differenced model as $b_0/(1 - b_1) = 0/1 = 0$. Therefore, a first-differenced random walk has a mean-reverting level of 0. Note also that the variance of y_t in each period is $\mathrm{Var}(\varepsilon_t) = \sigma^2$. Because the variance and the mean of y_t are constant and finite in each period, y_t is a covariance-stationary time series and we can model it using linear regression.[27] Of course, modeling the first-differenced series with an AR(1) model does not help us predict the future, as $b_0 = 0$ and $b_1 = 0$. We simply conclude that the original time series is, in fact, a random walk.

Had we tried to estimate an AR(1) model for a time series that was a random walk, our statistical conclusions would have been incorrect because AR models cannot be used to estimate random walks or any time series that is not covariance stationary. The following example illustrates this issue with exchange rates.

EXAMPLE 10

The Yen/US Dollar Exchange Rate

Financial analysts often assume that exchange rates are random walks. Consider an AR(1) model for the Japanese yen/US dollar exchange rate (JPY/USD). Table 9 shows the results of estimating the model using month-end observations from January 1980 through December 2013.

Table 9 Yen/US Dollar Exchange Rate: AR(1) Model Month-End Observations, January 1980–December 2013

Regression Statistics	
R-squared	0.9902
Standard error	4.9437
Observations	408
Durbin–Watson	1.8981

	Coefficient	Standard Error	t-Statistic
Intercept	0.9958	0.7125	1.3976
JPY/USD$_{t-1}$	0.9903	0.0049	202.1020

27 All the covariances are finite, for two reasons: The variance is finite, and the covariance of a time series with its own past value can be no greater than the variance of the series.

Table 9 (Continued)

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	0.0687	0.0495	1.3879
2	0.0384	0.0495	0.7758
3	0.0686	0.0495	1.3859
4	0.0407	0.0495	0.8222

Source: US Federal Reserve Board of Governors.

The results in Table 9 suggest that the yen/US dollar exchange rate is a random walk because the estimated intercept does not appear to be significantly different from 0 and the estimated coefficient on the first lag of the exchange rate is very close to 1. Can we use the t-statistics in Table 9 to test whether the exchange rate is a random walk? Unfortunately, no, because the standard errors in an AR model are invalid if the model is estimated using a data series that is a random walk (remember, a random walk is not covariance stationary). If the exchange rate is, in fact, a random walk, we might come to an incorrect conclusion based on faulty statistical tests and then invest incorrectly. We can use a test presented in the next section to test whether the time-series is a random walk.

Suppose the exchange rate is a random walk, as we now suspect. If so, the first-differenced series, $y_t = x_t - x_{t-1}$, will be covariance stationary. We present the results from estimating $y_t = b_0 + b_1 y_{t-1} + \varepsilon_t$ in Table 10. If the exchange rate is a random walk, then $b_0 = 0$ and $b_1 = 0$ and the error term will not be serially correlated.

Table 10 First-Differenced Yen/US Dollar Exchange Rate: AR(1) Model Month-End Observations, January 1980–December 2013

Regression Statistics

R-squared	0.0026
Standard error	4.9611
Observations	408
Durbin–Watson	2.0010

	Coefficient	Standard Error	*t*-Statistic
Intercept	−0.3128	0.2463	−1.2700
$\text{JPY/USD}_{t-1} - \text{JPY/USD}_{t-2}$	0.0506	0.0494	1.0243

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	0.0193	0.0495	0.3899
2	0.0345	0.0495	0.6970

(continued)

Table 10 (Continued)

Autocorrelations of the Residual			
Lag	**Autocorrelation**	**Standard Error**	***t*-Statistic**
3	0.0680	0.0495	1.3737
4	0.0399	0.0495	0.8061

Source: US Federal Reserve Board of Governors.

In Table 10, neither the intercept nor the coefficient on the first lag of the first-differenced exchange rate differs significantly from 0, and no residual autocorrelations differ significantly from 0.[28] These findings are consistent with the yen/US dollar exchange rate being a random walk.

We have concluded that the differenced regression is the model to choose. Now we can see that we would have been seriously misled if we had based our model choice on an R^2 comparison. In Table 9, the R^2 is 0.9902, whereas in Table 10 the R^2 is 0.0026. How can this be, if we just concluded that the model in Table 10 is the one that we should use? In Table 9, the R^2 measures how well the exchange rate in one period predicts the exchange rate in the next period. If the exchange rate is a random walk, its current value will be an extremely good predictor of its value in the next period, and thus the R^2 will be extremely high. At the same time, if the exchange rate is a random walk, then changes in the exchange rate should be completely unpredictable. Table 10 estimates whether changes in the exchange rate from one month to the next can be predicted by changes in the exchange rate over the previous month. If they cannot be predicted, the R^2 in Table 10 should be very low. In fact, it is low (0.0026). This comparison provides a good example of the general rule that we cannot necessarily choose which model is correct solely by comparing the R^2 from the two models.

The exchange rate is a random walk, and changes in a random walk are by definition unpredictable. Therefore, we cannot profit from an investment strategy that predicts changes in the exchange rate.

To this point, we have discussed only simple random walks; that is, random walks without drift. In a random walk without drift, the best predictor of the time series in the next period is its current value. A random walk with drift, however, should increase or decrease by a constant amount in each period. The equation describing a random walk with drift is a special case of the AR(1) model:

$$\begin{aligned} x_t &= b_0 + b_1 x_{t-1} + \varepsilon_t \\ b_1 &= 1,\ b_0 \neq 0, \text{ or} \\ x_t &= b_0 + x_{t-1} + \varepsilon_t,\ E(\varepsilon_t) = 0 \end{aligned} \tag{9}$$

A random walk with drift has $b_0 \neq 0$ compared to a simple random walk, which has $b_0 = 0$.

We have already seen that $b_1 = 1$ implies an undefined mean-reversion level and thus nonstationarity. Consequently, we cannot use an AR model to analyze a time series that is a random walk with drift until we transform the time series by taking first differences. If we first-difference Equation 9, the result is $y_t = x_t - x_{t-1}$, $y_t = b_0 + \varepsilon_t$, $b_0 \neq 0$.

28 See Greene (2018) for a test of the joint hypothesis that both regression coefficients are equal to 0.

5.2 The Unit Root Test of Nonstationarity

In this section, we discuss how to use random walk concepts to determine whether a time series is covariance stationary. This approach focuses on the slope coefficient in the random-walk-with-drift case of an AR(1) model in contrast with the traditional autocorrelation approach which we discuss first.

The examination of the autocorrelations of a time series at various lags is a well-known prescription for inferring whether or not a time series is stationary. Typically, for a stationary time series, either autocorrelations at all lags are statistically indistinguishable from zero, or the autocorrelations drop off rapidly to zero as the number of lags becomes large. Conversely, the autocorrelations of a nonstationary time series do not exhibit those characteristics. However, this approach is less definite than a currently more popular test for nonstationarity known as the Dickey–Fuller test for a unit root.

We can explain what is known as the unit root problem in the context of an AR(1) model. If a time series comes from an AR(1) model, then to be covariance stationary the absolute value of the lag coefficient, b_1, must be less than 1.0. We could not rely on the statistical results of an AR(1) model if the absolute value of the lag coefficient were greater than or equal to 1.0 because the time series would not be covariance stationary. If the lag coefficient is equal to 1.0, the time series has a **unit root**: it is a random walk and is not covariance stationary.[29] By definition, all random walks, with or without a drift term, have unit roots.

How do we test for unit roots in a time series? If we believed that a time series, x_t, was a random walk with drift, it would be tempting to estimate the parameters of the AR(1) model $x_t = b_0 + b_1x_{t-1} + \varepsilon_t$ using linear regression and conduct a t-test of the hypothesis that $b_1 = 1$. Unfortunately, if $b_1 = 1$, then x_t is not covariance stationary and the t-value of the estimated coefficient, $\hat{b}_1$, does not actually follow the t-distribution; consequently, a t-test would be invalid.

Dickey and Fuller (1979) developed a regression-based unit root test based on a transformed version of the AR(1) model $x_t = b_0 + b_1x_{t-1} + \varepsilon_t$. Subtracting x_{t-1} from both sides of the AR(1) model produces

$$x_t - x_{t-1} = b_0 + (b_1 - 1)x_{t-1} + \varepsilon_t$$

or

$$x_t - x_{t-1} = b_0 + g_1x_{t-1} + \varepsilon_t,\ E(\varepsilon_t) = 0 \qquad (10)$$

where $g_1 = (b_1 - 1)$. If $b_1 = 1$, then $g_1 = 0$ and thus a test of $g_1 = 0$ is a test of $b_1 = 1$. If there is a unit root in the AR(1) model, then g_1 will be 0 in a regression where the dependent variable is the first difference of the time series and the independent variable is the first lag of the time series. The null hypothesis of the Dickey–Fuller test is H_0: $g_1 = 0$—that is, that the time series has a unit root and is nonstationary—and the alternative hypothesis is H_a: $g_1 < 0$, that the time series does not have a unit root and is stationary.

To conduct the test, one calculates a t-statistic in the conventional manner for $\hat{g}_1$ but instead of using conventional critical values for a t-test, one uses a revised set of values computed by Dickey and Fuller; the revised set of critical values are larger in absolute value than the conventional critical values. A number of software packages incorporate Dickey–Fuller tests.[30]

---

29 When b_1 is greater than 1 in absolute value, we say that there is an explosive root. For details, see Diebold (2008).

30 Dickey and Fuller developed three separate tests of the hypothesis that $g_1 = 0$ assuming the following models: random walk, random walk with drift, or random walk with drift and trend. The critical values for the Dickey–Fuller tests for the three models are different. For more on this topic, see Greene (2018) or Tsay (2010).

EXAMPLE 11

AstraZeneca's Quarterly Sales (1)

In January 2012, equity analyst Aron Berglin is building a time-series model for the quarterly sales of AstraZeneca, a British-Swedish biopharmaceutical company headquartered in London, UK. He is using AstraZeneca's quarterly sales in US dollars during January 2000 to December 2011 and any lagged sales data that he may need prior to 2000 to build this model. He finds that a log-linear trend model seems better suited for modeling AstraZeneca's sales than does a linear trend model. However, the Durbin–Watson statistic from the log-linear regression is just 0.7064, which causes him to reject the hypothesis that the errors in the regression are serially uncorrelated. He concludes that he cannot model the log of AstraZeneca's quarterly sales using only a time-trend line. He decides to model the log of AstraZeneca's quarterly sales using an AR(1) model. He uses $\ln \text{Sales}_t = b_0 + b_1 \ln \text{Sales}_{t-1} + \varepsilon_t$.

Before he estimates this regression, the analyst should use the Dickey–Fuller test to determine whether there is a unit root in the log of AstraZeneca's quarterly sales. If he uses the sample of quarterly data on AstraZeneca's sales from the first quarter of 2000 through the fourth quarter of 2011, takes the natural log of each observation, and computes the Dickey–Fuller *t*-test statistic, the value of that statistic might cause him to fail to reject the null hypothesis that there is a unit root in the log of AstraZeneca's quarterly sales.

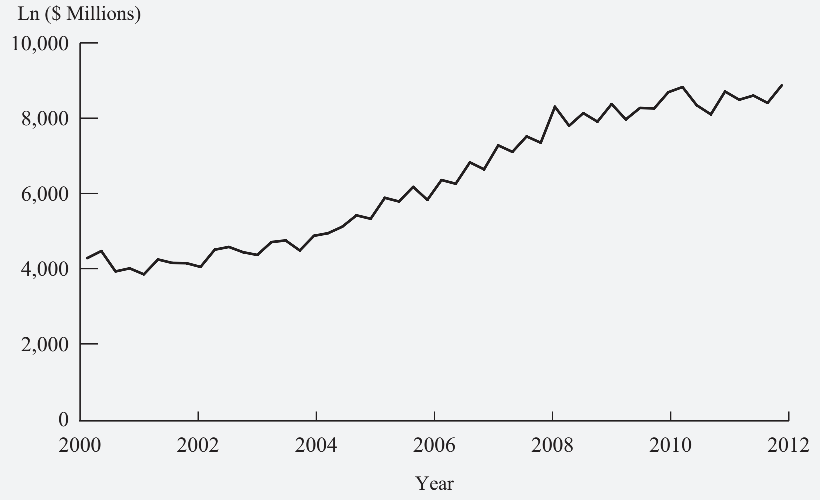
If a time series appears to have a unit root, how should we model it? One method that is often successful is to model the first-differenced series as an autoregressive time series. The following example demonstrates this method.

EXAMPLE 12

AstraZeneca's Quarterly Sales (2)

The plot of the log of AstraZeneca's quarterly sales is shown as Figure 10. By looking at the plot, Berglin is convinced that the log of quarterly sales is not covariance stationary (that it has a unit root).

Figure 10 Log of AstraZeneca's Quarterly Sales

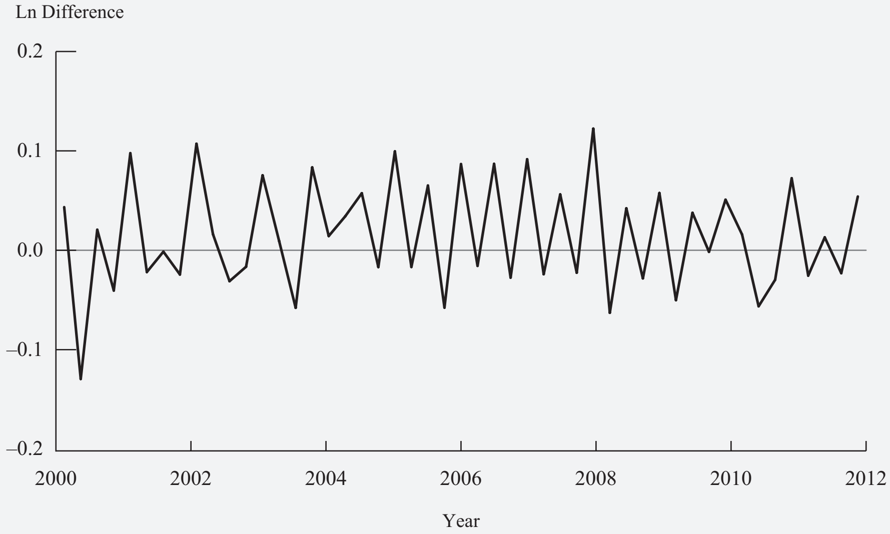


Source: Compustat.

So he creates a new series, y_t, that is the first difference of the log of AstraZeneca's quarterly sales. Figure 11 shows that series.

Berglin compares Figure 11 to Figure 10 and notices that first-differencing the log of AstraZeneca's quarterly sales eliminates the strong upward trend that was present in the log of AstraZeneca's sales. Because the first-differenced series has no strong trend, Berglin is better off assuming that the differenced series is covariance stationary rather than assuming that AstraZeneca's sales or the log of AstraZeneca's sales is a covariance-stationary time series.

Figure 11 Log Difference, AstraZeneca's Quarterly Sales



Source: Compustat.

Now suppose Berglin decides to model the new series using an AR(1) model. Berglin uses $\ln(\text{Sales}_t) - \ln(\text{Sales}_{t-1}) = b_0 + b_1[\ln(\text{Sales}_{t-1}) - \ln(\text{Sales}_{t-2})] + \varepsilon_t$. Table 11 shows the results of that regression.

Table 11 Log Differenced Sales: AR(1) Model of AstraZeneca Quarterly Observations, January 2000–December 2011

Regression Statistics	
R-squared	0.3005
Standard error	0.0475
Observations	48
Durbin–Watson	1.6874

	Coefficient	Standard Error	t-Statistic
Intercept	0.0222	0.0071	3.1268
$\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}$	−0.5493	0.1236	−4.4442

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	t-Statistic
1	0.2809	0.1443	1.9466
2	−0.0466	0.1443	−0.3229
3	0.0081	0.1443	0.0561
4	0.2647	0.1443	1.8344

Source: Compustat.

The lower part of Table 11 suggests that the first four autocorrelations of residuals in this model are not statistically significant. With 48 observations and two parameters, this model has 46 degrees of freedom. The critical value for a t-statistic in this model is above 2.0 at the 0.05 significance level. None of the t-statistics for these autocorrelations has an absolute value larger than 2.0. Therefore, we fail to reject the null hypotheses that each of these autocorrelations is equal to 0 and conclude instead that no significant autocorrelation is present in the residuals.

This result suggests that the model is well specified and that we could use the estimates. Both the intercept ($\hat{b}_0 = 0.0222$) and the coefficient ($\hat{b}_1 = -0.5493$) on the first lag of the new first-differenced series are statistically significant.

1 Explain how to interpret the estimated coefficients in the model.

Solution to 1:

The value of the intercept (0.0222) implies that if sales have not changed in the current quarter (y_t = ln Sales$_t$ − ln Sales$_{t-1}$ = 0), sales will grow by 2.22 percent next quarter.[31] If sales have changed during this quarter, however, the model predicts that sales will grow by 2.22 percent minus 0.5493 times the sales growth in this quarter.

2 AstraZenca's sales in the third and fourth quarters of 2011 were $8,405 million and $8,872 million, respectively. If we use the above model soon after the end of the fourth quarter of 2011, what will be the predicted value of AstraZenca's sales for the first quarter of 2012?

Solution to 2:

Let us say that t is the fourth quarter of 2011, so $t - 1$ is the third quarter of 2011 and $t + 1$ is the first quarter of 2012. Then we would have to compute $\hat{y}_{t+1} = 0.0222 - 0.5493y_t$. To compute $\hat{y}_{t+1}$, we need to know y_t = ln Sales$_t$ − ln Sales$_{t-1}$. In the third quarter of 2011, AstraZenca's sales were $8,405 million, so ln (Sales$_{t-1}$) = ln 8,405 = 9.0366. In the fourth quarter of 2011, AstraZenca's sales were $8,872 million, so ln (Sales$_t$) = ln 8,872 = 9.0907. Thus y_t = 9.0907 − 9.0366 = 0.0541. Therefore, $\hat{y}_{t+1}$ = 0.0222 − 0.5493(0.0541) = −0.0075. If $\hat{y}_{t+1}$ = −0.0075, then −0.0075 = ln (Sales$_{t+1}$) − ln (Sales$_t$) = ln (Sales$_{t+1}$/Sales$_t$). If we exponentiate both sides of this equation, the result is

$$e^{-0.0075} = \left(\frac{\text{Sales}_{t+1}}{\text{Sales}_t}\right)$$

$$\begin{aligned}\text{Sales}_{t+1} &= \text{Sales}_t e^{-0.0075} \\ &= \$8{,}872 \text{ million} \times 0.9925 \\ &= \$8{,}805 \text{ million}\end{aligned}$$

Thus, based on fourth quarter sales for 2011, this model would have predicted that AstraZenca's sales in the first quarter of 2012 would be $8,805 million. This sales forecast might have affected our decision to buy AstraZenca's stock at the time.

6 MOVING-AVERAGE TIME-SERIES MODELS

So far, many of the forecasting models we have used have been autoregressive models. Because most financial time series have the qualities of an autoregressive process, autoregressive time-series models are probably the most frequently used time-series models in financial forecasting. Some financial time series, however, seem to follow more closely another kind of time-series model called a moving-average model. For example, as we will see later, returns on the S&P BSE 100 Index can be better modeled as a moving-average process than as an autoregressive process.

31 Note that 2.22 percent is the exponential growth rate, not [(Current quarter sales/Previous quarter sales) − 1]. The difference between these two methods of computing growth is usually small.

In this section, we present the fundamentals of moving-average models so that you can ask the right questions when considering their use. We first discuss how to smooth past values with a moving average and then how to forecast a time series using a moving-average model. Even though both methods include the words "moving average" in the name, they are very different.

6.1 Smoothing Past Values with an *n*-Period Moving Average

Suppose you are analyzing the long-term trend in the past sales of a company. In order to focus on the trend, you may find it useful to remove short-term fluctuations or noise by smoothing out the time series of sales. One technique to smooth out period-to-period fluctuations in the value of a time series is an ***n*-period moving average**. An n-period moving average of the current and past $n - 1$ values of a time series, x_t, is calculated as

$$\frac{x_t + x_{t-1} + \cdots + x_{t-(n-1)}}{n} \tag{11}$$

The following example demonstrates how to compute a moving average of AstraZeneca's quarterly sales.

EXAMPLE 13

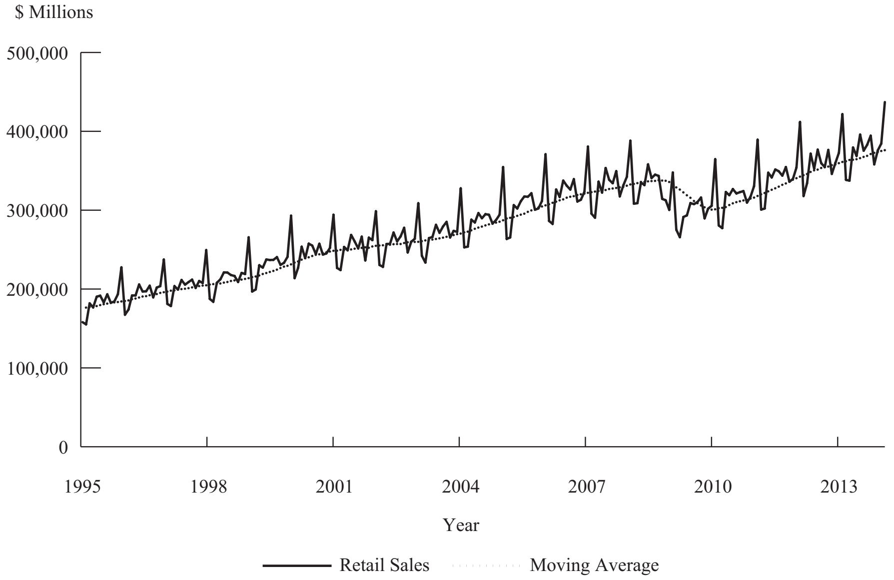
AstraZeneca's Quarterly Sales (3)

Suppose we want to compute the four-quarter moving average of AstraZeneca's sales as of the beginning of the first quarter of 2012. AstraZeneca's sales in the previous four quarters were 1Q:2011, \$8,490 million; 2Q:2011, \$8,601 million; 3Q:2011, \$8,405 million; and 4Q:2011, \$8,872 million. The four-quarter moving average of sales as of the beginning of the first quarter of 2012 is thus (8,490 + 8,601 + 8,405 + 8,872)/4 = \$8,592 million.

We often plot the moving average of a series with large fluctuations to help discern any patterns in the data. Figure 12 shows monthly retail sales for the United States from January 1995 to December 2013, along with a 12-month moving average of the data.[32]

32 A 12-month moving average is the average value of a time series over each of the last 12 months. Although the sample period starts in 1995, data from 1994 are used to compute the 12-month moving average for the months of 1994.

Figure 12 Monthly US Real Retail Sales and 12-Month Moving Average of Retail Sales

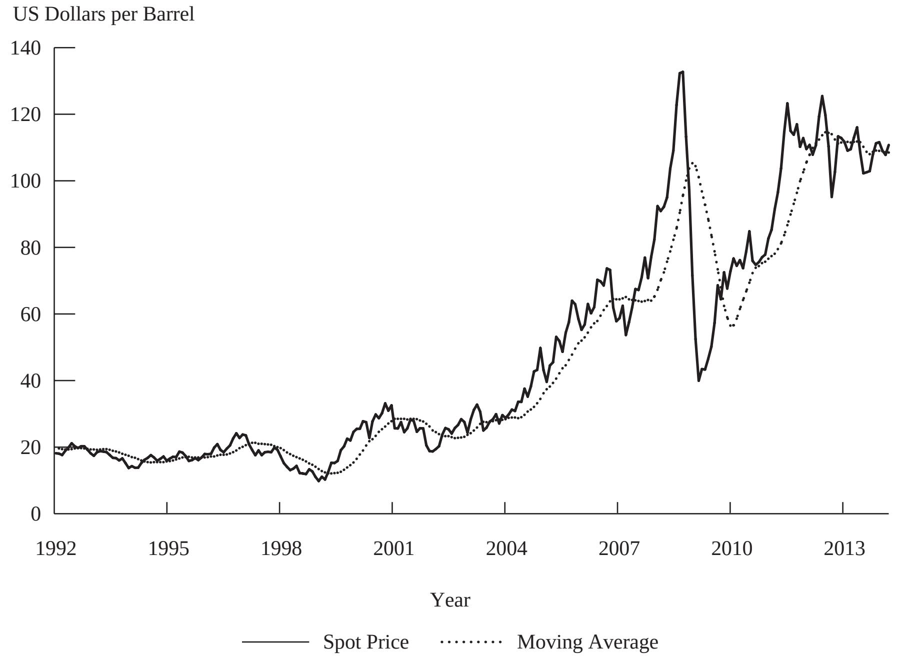


Source: US Department of Commerce, Census Bureau.

As Figure 12 shows, each year has a very strong peak in retail sales (December) followed by a sharp drop in sales (January). Because of the extreme seasonality in the data, a 12-month moving average can help us focus on the long-term movements in retail sales instead of seasonal fluctuations. Note that the moving average does not have the sharp seasonal fluctuations of the original retail sales data. Rather, the moving average of retail sales grows steadily, for example, from 1995 through the second half of 2008, then declines for about a year, and grows steadily thereafter. We can see that trend more easily by looking at a 12-month moving average than by looking at the time series itself.

Figure 13 shows monthly Europe Brent Crude Oil spot prices along with a 12-month moving average of oil prices. Although these data do not have the same sharp regular seasonality displayed in the retail sales data in Figure 12, the moving average smooths out the monthly fluctuations in oil prices to show the longer-term movements.

Figure 13 Monthly Europe Brent Crude Oil Price and 12-Month Moving Average of Prices



Source: US Energy Information Administration.

Figure 13 also shows one weakness with a moving average: It always lags large movements in the actual data. For example, when oil prices rose quickly in late 2007 and the first half of 2008, the moving average rose only gradually. When oil prices fell sharply toward the end of 2008, the moving average also lagged. Consequently, a simple moving average of the recent past, though often useful in smoothing out a time series, may not be the best predictor of the future. A main reason for this is that a simple moving average gives equal weight to all the periods in the moving average. In order to forecast the future values of a time series, it is often better to use a more sophisticated moving-average time-series model. We discuss such models below.

6.2 Moving-Average Time-Series Models for Forecasting

Suppose that a time series, x_t, is consistent with the following model:

$$\begin{aligned} x_t &= \varepsilon_t + \theta\varepsilon_{t-1},\ E(\varepsilon_t) = 0,\ E(\varepsilon_t^2) = \sigma^2, \\ \mathrm{Cov}(\varepsilon_t, \varepsilon_s) &= E(\varepsilon_t \varepsilon_s) = 0 \text{ for } t \neq s \end{aligned} \tag{12}$$

This equation is called a moving-average model of order 1, or simply an MA(1) model. Theta (θ) is the parameter of the MA(1) model.[33]

[33] Note that a moving-average time-series model is very different from a simple moving average, as discussed in Section 6.1. The simple moving average is based on observed values of a time series. In a moving-average time-series model, we never directly observe, ε_t or any other ε_{t-j}, but we can infer how a particular moving-average model will imply a particular pattern of serial correlation for a time series, as we discuss below.

Equation 12 is a moving-average model because in each period, x_t is a moving average of ε_t and ε_{t-1}, two uncorrelated random variables that each have an expected value of zero. Unlike the simple moving-average model of Equation 11, this moving-average model places different weights on the two terms in the moving average (1 on ε_t, and θ on ε_{t-1}).

We can see if a time series fits an MA(1) model by looking at its autocorrelations to determine whether x_t is correlated only with its preceding and following values. First, we examine the variance of x_t in Equation 12 and its first two autocorrelations. Because the expected value of x_t is 0 in all periods and ε_t is uncorrelated with its own past values, the first autocorrelation is not equal to 0, but the second and higher autocorrelations are equal to 0. Further analysis shows that all autocorrelations except for the first will be equal to 0 in an MA(1) model. Thus for an MA(1) process, any value x_t is correlated with x_{t-1} and x_{t+1} but with no other time-series values; we could say that an MA(1) model has a memory of one period.

Of course, an MA(1) model is not the most complex moving-average model. A qth order moving-average model, denoted MA(q) and with varying weights on lagged terms, can be written as

$$x_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q},\ E(\varepsilon_t) = 0,\ E\left(\varepsilon_t^2\right) = \sigma^2,$$
$$\operatorname{Cov}(\varepsilon_t, \varepsilon_s) = E(\varepsilon_t\varepsilon_s) = 0 \text{ for } t \neq s \tag{13}$$

How can we tell whether an MA(q) model fits a time series? We examine the autocorrelations. For an MA(q) model, the first q autocorrelations will be significantly different from 0, and all autocorrelations beyond that will be equal to 0; an MA(q) model has a memory of q periods. This result is critical for choosing the right value of q for an MA model. We discussed this result above for the specific case of $q = 1$ that all autocorrelations except for the first will be equal to 0 in an MA(1) model.

How can we distinguish an autoregressive time series from a moving-average time series? Once again, we do so by examining the autocorrelations of the time series itself. The autocorrelations of most autoregressive time series start large and decline gradually, whereas the autocorrelations of an MA(q) time series suddenly drop to 0 after the first q autocorrelations. We are unlikely to know in advance whether a time series is autoregressive or moving average. Therefore, the autocorrelations give us our best clue about how to model the time series. Most time series, however, are best modeled with an autoregressive model.

EXAMPLE 14

A Time-Series Model for Monthly Returns on the S&P BSE 100 Index

The S&P BSE 100 index is designed to reflect the performance of India's top 100 large-cap companies listed on the BSE Ltd. (formerly Bombay Stock Exchange). Are monthly returns on the S&P BSE 100 index autocorrelated? If so, we may be able to devise an investment strategy to exploit the autocorrelation. What is an appropriate time-series model for S&P BSE 100 monthly returns?

Table 12 shows the first six autocorrelations of returns to the S&P BSE 100 using monthly data from January 2000 through December 2013. Note that all of the autocorrelations are quite small. Do they reach significance? With 168 observations, the critical value for a t-statistic in this model is about 1.98 at the 0.05 significance level. None of the autocorrelations has a t-statistic larger in absolute value than the critical value of 1.98. Consequently, we fail to reject the null hypothesis that those autocorrelations, individually, do not differ significantly from 0.

Table 12 Annualized Monthly Returns to the S&P BSE 100 January 2000–December 2013

Autocorrelations			
Lag	**Autocorrelation**	**Standard Error**	***t*-Statistic**
1	0.1103	0.0772	1.4288
2	−0.0045	0.0772	−0.0583
3	0.0327	0.0772	0.4236
4	0.0370	0.0772	0.4793
5	−0.0218	0.0772	−0.2824
6	0.0191	0.0772	0.2474
Observations	168		

Source: BSE Ltd.

If returns on the S&P BSE 100 were an MA(q) time series, then the first q autocorrelations would differ significantly from 0. None of the autocorrelations is statistically significant, however, so returns to the S&P BSE 100 appear to come from an MA(0) time series. An MA(0) time series in which we allow the mean to be nonzero takes the following form:[34]

$$\begin{aligned} x_t &= \mu + \varepsilon_t,\ E(\varepsilon_t) = 0,\ E\left(\varepsilon_t^2\right) = \sigma^2, \\ \mathrm{Cov}(\varepsilon_t, \varepsilon_s) &= E(\varepsilon_t \varepsilon_s) = 0 \text{ for } t \neq s \end{aligned} \quad (14)$$

which means that the time series is not predictable. This result should not be too surprising, as most research suggests that short-term returns to stock indexes are difficult to predict.

We can see from this example how examining the autocorrelations allowed us to choose between the AR and MA models. If returns to the S&P BSE 100 had come from an AR(1) time series, the first autocorrelation would have differed significantly from 0 and the autocorrelations would have declined gradually. Not even the first autocorrelation is significantly different from 0, however. Therefore, we can be sure that returns to the S&P BSE 100 do not come from an AR(1) model—or from any higher-order AR model, for that matter. This finding is consistent with our conclusion that the S&P BSE 100 series is MA(0).

7 SEASONALITY IN TIME-SERIES MODELS

As we analyze the results of the time-series models in this reading, we encounter complications. One common complication is significant **seasonality**, a case in which the series shows regular patterns of movement within the year. At first glance,

34 On the basis of investment theory and evidence, we expect that the mean monthly return on the S&P BSE 100 is positive ($\mu > 0$). We can also generalize Equation 13 for an MA(q) time series by adding a constant term, μ. Including a constant term in a moving-average model does not change the expressions for the variance and autocovariances of the time series. A number of early studies of weak-form market efficiency used Equation 14 as the model for stock returns. See Garbade (1982).

seasonality might appear to rule out using autoregressive time-series models. After all, autocorrelations will differ by season. This problem can often be solved, however, by using seasonal lags in an autoregressive model.

A seasonal lag is usually the value of the time series one year before the current period, included as an extra term in an autoregressive model. Suppose, for example, that we model a particular quarterly time series using an AR(1) model, $x_t = b_0 + b_1x_{t-1} + \varepsilon_t$. If the time series had significant seasonality, this model would not be correctly specified. The seasonality would be easy to detect because the seasonal autocorrelation (in the case of quarterly data, the fourth autocorrelation) of the error term would differ significantly from 0. Suppose this quarterly model has significant seasonality. In this case, we might include a seasonal lag in the autoregressive model and estimate

$$x_t = b_0 + b_1x_{t-1} + b_2x_{t-4} + \varepsilon_t \quad (15)$$

to test whether including the seasonal lag would eliminate statistically significant autocorrelation in the error term.

In Examples 15 and 16, we illustrate how to test and adjust for seasonality in a time-series model. We also illustrate how to compute a forecast using an autoregressive model with a seasonal lag.

EXAMPLE 15

Seasonality in Sales at Starbucks

Earlier, we concluded that we could not model the log of Starbucks' quarterly sales using only a time-trend line (as shown in Example 3) because the Durbin–Watson statistic from the regression provided evidence of positive serial correlation in the error term. Based on methods presented in this reading, we might next investigate using the first difference of log sales to remove an exponential trend from the data to obtain a covariance stationary time series.

Using quarterly data from the first quarter of 1995 to the last quarter of 2012, we estimate the following AR(1) model using ordinary least squares: $(\ln \text{Sales}_t - \ln \text{Sales}_{t-1}) = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + \varepsilon_t$. Table 13 shows the results of the regression.

Table 13 Log Differenced Sales: AR(1) Model Starbucks, Quarterly Observations, 1995–2013

Regression Statistics	
R-squared	0.1548
Standard error	0.0762
Observations	74
Durbin–Watson	1.9165

	Coefficient	Standard Error	t-Statistic
Intercept	0.0669	0.0101	6.6238
$\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}$	−0.3813	0.1050	−3.6314

(continued)

Table 13 (Continued)

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	−0.0141	0.1162	−0.1213
2	−0.0390	0.1162	−0.3356
3	0.0294	0.1162	0.2530
4	0.7667	0.1162	6.5981

Source: Compustat.

The first thing to note in Table 13 is the strong seasonal autocorrelation of the residuals. The bottom portion of the table shows that the fourth autocorrelation has a value of 0.7667 and a t-statistic of 6.60. With 74 observations and two parameters, this model has 72 degrees of freedom.[35] The critical value for a t-statistic is about 1.99 at the 0.05 significance level. Given this value of the t-statistic, we must reject the null hypothesis that the fourth autocorrelation is equal to 0 because the t-statistic is larger than the critical value of 1.99.

In this model, the fourth autocorrelation is the seasonal autocorrelation because this AR(1) model is estimated with quarterly data. Table 13 shows the strong and statistically significant seasonal autocorrelation that occurs when a time series with strong seasonality is modeled without taking the seasonality into account. Therefore, the AR(1) model is misspecified, and we should not use it for forecasting.

Suppose we decide to use an autoregressive model with a seasonal lag because of the seasonal autocorrelation. We are modeling quarterly data, so we estimate Equation 15: $(\ln \text{Sales}_t - \ln \text{Sales}_{t-1}) = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + b_2(\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}) + \varepsilon_t$. Adding the seasonal difference $\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}$ is an attempt to remove a consistent quarterly pattern in the data and could also eliminate a seasonal non-stationarity if one existed. The estimates of this equation appear in Table 14.

Table 14 Log Differenced Sales: AR(1) Model with Seasonal Lag Starbucks, Quarterly Observations, 1995–2013

Regression Statistics

R-squared	0.8163
Standard error	0.03405
Observations	71
Durbin–Watson	2.0791

35 In this example, we restrict the start of the sample period to the beginning of 1995, and we do not use prior observations for the lags. Accordingly, the number of observations decreases with an increase in the number of lags. In Table 13, the first observation is for the third quarter of 1995 because we use up to two lags. In Table 14, the first observation is for the second quarter of 1996 because we use up to five lags.

Table 14 (Continued)

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0084	0.0059	1.4237
ln Sales_{t-1} – ln Sales_{t-2}	–0.0602	0.0540	–1.1148
ln Sales_{t-4} – ln Sales_{t-5}	0.8048	0.0524	15.3588

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	–0.0441	0.1187	–0.3715
2	0.0675	0.1187	0.5687
3	0.0749	0.1187	0.6310
4	–0.2091	0.1187	–1.7616

Source: Compustat.

Note the autocorrelations of the residual shown at the bottom of Table 14. When we include a seasonal lag in the regression, the coefficient on the first difference in log sales, with a *t*-statistic of –1.1148, is no longer statistically significant. However, none of the *t*-statistics on the first four autocorrelations is now significant. Because the overall regression is highly significant (an *F*-test is significant at the 0.01 level), we can take an AR(1) model with a seasonal lag as a reasonable working model for Starbucks sales. (A model having only a seasonal lag term was investigated and not found to improve on this model.)

How can we interpret the coefficients in this model? To predict the current quarter's sales growth at Starbucks, we need to know two things: sales growth in the previous quarter and sales growth four quarters ago. If sales remained constant in each of those two quarters, the model in Table 14 predicts that sales will grow by 0.0084 (0.84 percent) in the current quarter. If sales grew by 1 percent last quarter and by 2 percent four quarters ago, then the model predicts that sales growth this quarter will be 0.0084 – 0.0602(0.01) + 0.8048(0.02) = 0.0239 or 2.39 percent.[36] Notice also that the R^2 in the model with the seasonal lag (0.8163 in Table 14) was more than five times higher than the R^2 in the model without the seasonal lag (0.1548 in Table 13). Again, the seasonal lag model does a much better job of explaining the data.

EXAMPLE 16

Retail Sales Growth

We want to predict the growth in monthly retail sales of Canadian furniture and home furnishing stores so that we can decide whether to recommend the shares of these stores. We decide to use non-seasonally adjusted data on retail sales. To begin with, we estimate an AR(1) model with observations on the

36 Note that all of these growth rates are exponential growth rates.

annualized monthly growth in retail sales from January 1995 to December 2012. We estimate the following equation: Sales growth$_t$ = b_0 + b_1 Sales growth$_{t-1}$ + ε_t. Table 15 shows the results from this model.

The autocorrelations of the residuals from this model, shown at the bottom of Table 15, indicate that seasonality is extremely significant in this model. With 216 observations and two parameters, this model has 214 degrees of freedom. At the 0.05 significance level, the critical value for a t-statistic is about 1.97. The 12th-lag autocorrelation (the seasonal autocorrelation, because we are using monthly data) has a value of 0.7620 and a t-statistic of 11.21. The t-statistic on this autocorrelation is larger than the critical value (1.97) implying that we can reject the null hypothesis that the 12th autocorrelation is 0. Note also that many of the other t-statistics for autocorrelations shown in the table differ significantly from 0. Consequently, the model shown in Table 15 is misspecified, so we cannot rely on it to forecast sales growth.

Suppose we add the seasonal lag of sales growth (the 12th lag) to the AR(1) model to estimate the equation Sales growth$_t$ = b_0 + b_1(Sales growth$_{t-1}$) + b_2(Sales growth$_{t-12}$) + ε_t.[37] Table 16 presents the results of estimating this equation. The estimated value of the seasonal autocorrelation (the 12th autocorrelation) has fallen to −0.1168. None of the first 12 autocorrelations has a t-statistic with an absolute value greater than the critical value of 1.97 at the 0.05 significance level. We can conclude that there is no significant serial correlation in the residuals from this model. Because we can reasonably believe that the model is correctly specified, we can use it to predict retail sales growth. Note that the R^2 in Table 16 is 0.6724, much larger than the R^2 in Table 15 (computed by the model without the seasonal lag).

Table 15 Monthly Retail Sales Growth of Canadian Furniture and Home Furnishing Stores: AR(1) Model January 1995–December 2012

Regression Statistics	
R-squared	0.0509
Standard error	1.8198
Observations	216
Durbin–Watson	2.0956

	Coefficient	Standard Error	t-Statistic
Intercept	1.0518	0.1365	7.7055
Sales growth$_{t-1}$	−0.2252	0.0665	−3.3865

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	t-Statistic
1	−0.0109	0.0680	−0.1603
2	−0.1949	0.0680	−2.8662
3	0.1173	0.0680	1.7250

37 In this example, although we state that the sample period begins in 1995, we use prior observations for the lags. This results in the same number of observations irrespective of the number of lags.

Table 15 (Continued)

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
4	−0.0756	0.0680	−1.1118
5	−0.1270	0.0680	−1.8676
6	−0.1384	0.0680	−2.0353
7	−0.1374	0.0680	−2.0206
8	−0.0325	0.0680	−0.4779
9	0.1207	0.0680	1.7750
10	−0.2197	0.0680	−3.2309
11	−0.0342	0.0680	−0.5029
12	0.7620	0.0680	11.2059

Source: Statistics Canada (Government of Canada).

How can we interpret the coefficients in the model? To predict growth in retail sales in this month, we need to know last month's retail sales growth and retail sales growth 12 months ago. If retail sales remained constant both last month and 12 months ago, the model in Table 16 predicts that retail sales will grow at an annual rate of about 23.7 percent this month. If retail sales grew at an annual rate of 10 percent last month and at an annual rate of 5 percent 12 months ago, the model in Table 16 predicts that retail sales will grow in the current month at an annual rate of 0.2371 − 0.0792(0.10) + 0.7798(0.05) = 0.2682 or 26.8 percent.

Table 16 Monthly Retail Sales Growth of Canadian Furniture and Home Furnishing Stores: AR(1) Model with Seasonal Lag January 1995–December 2012

Regression Statistics

R-squared	0.6724
Standard error	1.0717
Observations	216
Durbin–Watson	2.1784

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.2371	0.0900	2.6344
Sales growth$_{t-1}$	−0.0792	0.0398	−1.9899
Sales growth$_{t-12}$	0.7798	0.0388	20.0979

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	−0.0770	0.0680	−1.1324
2	−0.0374	0.0680	−0.5500
3	0.0292	0.0680	0.4294

(continued)

Table 16 (Continued)

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
4	−0.0358	0.0680	−0.5265
5	−0.0399	0.0680	−0.5868
6	0.0227	0.0680	0.3338
7	−0.0967	0.0680	−1.4221
8	0.1241	0.0680	1.8250
9	0.0499	0.0680	0.7338
10	−0.0631	0.0680	−0.9279
11	0.0231	0.0680	0.3397
12	−0.1168	0.0680	−1.7176

Source: Statistics Canada (Government of Canada).

8 AUTOREGRESSIVE MOVING-AVERAGE MODELS

So far, we have presented autoregressive and moving-average models as alternatives for modeling a time series. The time series we have considered in examples have usually been explained quite well with a simple autoregressive model (with or without seasonal lags).[38] Some statisticians, however, have advocated using a more general model, the autoregressive moving-average (ARMA) model. The advocates of ARMA models argue that these models may fit the data better and provide better forecasts than do plain autoregressive (AR) models. However, as we discuss later in this section, there are severe limitations to estimating and using these models. Because you may encounter ARMA models, we provide a brief overview below.

An ARMA model combines both autoregressive lags of the dependent variable and moving-average errors. The equation for such a model with p autoregressive terms and q moving-average terms, denoted ARMA(p, q), is

$$\begin{aligned} x_t &= b_0 + b_1 x_{t-1} + \cdots + b_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \\ E(\varepsilon_t) &= 0,\ E(\varepsilon_t^2) = \sigma^2,\ \mathrm{Cov}(\varepsilon_t, \varepsilon_s) = E(\varepsilon_t \varepsilon_s) = 0 \text{ for } t \neq s \end{aligned} \tag{16}$$

where $b_1, b_2, \ldots, b_p$ are the autoregressive parameters and $\theta_1, \theta_2, \ldots, \theta_q$ are the moving-average parameters.

Estimating and using ARMA models has several limitations. First, the parameters in ARMA models can be very unstable. In particular, slight changes in the data sample or the initial guesses for the values of the ARMA parameters can result in very different final estimates of the ARMA parameters. Second, choosing the right ARMA model is more of an art than a science. The criteria for deciding on p and q for a particular time series are far from perfect. Moreover, even after a model is selected, that model may not forecast well.

38 For the returns on the S&P BSE 100 (see Example 14), we chose a moving-average model over an autoregressive model.

To reiterate, ARMA models can be very unstable, depending on the data sample used and the particular ARMA model estimated. Therefore, you should be skeptical of claims that a particular ARMA model provides much better forecasts of a time series than any other ARMA model. In fact, in most cases, you can use an AR model to produce forecasts that are just as accurate as those from ARMA models without nearly as much complexity. Even some of the strongest advocates of ARMA models admit that these models should not be used with fewer than 80 observations, and they do not recommend using ARMA models for predicting quarterly sales or gross margins for a company using even 15 years of quarterly data.

9 AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTICITY MODELS

Up to now, we have ignored any issues of heteroskedasticity in time-series models and have assumed homoskedasticity. **Heteroskedasticity** is the dependence of the error term variance on the independent variable; **homoskedasticity** is the independence of the error term variance from the independent variable. We have assumed that the error term's variance is constant and does not depend on the value of the time series itself or on the size of previous errors. At times, however, this assumption is violated and the variance of the error term is not constant. In such a situation, the standard errors of the regression coefficients in AR, MA, or ARMA models will be incorrect, and our hypothesis tests would be invalid. Consequently, we can make poor investment decisions based on those tests.

For example, suppose you are building an autoregressive model of a company's sales. If heteroskedasticity is present, then the standard errors of the regression coefficients of your model are incorrect. It is likely that due to heteroskedasticity, one or more of the lagged sales terms may appear statistically significant when in fact they are not. Therefore, if you use this model for your decision making, you may make some suboptimal decisions.

In work responsible in part for his shared Nobel Prize in Economics for 2003, Robert F. Engle in 1982 first suggested a way of testing whether the variance of the error in a particular time-series model in one period depends on the variance of the error in previous periods. He called this type of heteroskedasticity autoregressive conditional heteroskedasticity (ARCH).

As an example, consider the ARCH(1) model

$$\varepsilon_t \sim N\left(0, a_0 + a_1\varepsilon_{t-1}^2\right) \quad \textbf{(17)}$$

where the distribution of ε_t, conditional on its value in the previous period, ε_{t-1}, is normal with mean 0 and variance $a_0 + a_1\varepsilon_{t-1}^2$. If $a_1 = 0$, the variance of the error in every period is just a_0. The variance is constant over time and does not depend on past errors. Now suppose that $a_1 > 0$. Then the variance of the error in one period depends on how large the squared error was in the previous period. If a large error occurs in one period, the variance of the error in the next period will be even larger.

Engle shows that we can test whether a time series is ARCH(1) by regressing the squared residuals from a previously estimated time-series model (AR, MA, or ARMA) on a constant and one lag of the squared residuals. We can estimate the linear regression equation

$$\hat{\varepsilon}_t^2 = a_0 + a_1\hat{\varepsilon}_{t-1}^2 + u_t \quad \textbf{(18)}$$

where u_t is an error term. If the estimate of a_1 is statistically significantly different from zero, we conclude that the time series is ARCH(1). If a time-series model has ARCH(1) errors, then the variance of the errors in period $t + 1$ can be predicted in period t using the formula $\hat{\sigma}^2_{t+1} = \hat{a}_0 + \hat{a}_1 \hat{\varepsilon}^2_t$.

EXAMPLE 17

Testing for ARCH(1) in Monthly Inflation

Analyst Lisette Miller wants to test whether monthly data on CPI inflation contain autoregressive conditional heteroskedasticity. She could estimate Equation 18 using the residuals from the time-series model. Based on the analyses in Examples 6 through 9, she has concluded that if she modeled monthly CPI inflation from 1984 to 2013, there is not much difference in the performance of AR(1) and AR(2) models in forecasting inflation. The AR(1) model is clearly better for the period of 2007–2013. She decides to further explore the AR(1) model for the entire period of 1984 to 2013. Table 17 shows the results of testing whether the errors in that model are ARCH(1). Because the test involves the first lag of residuals of the estimated time series model, the number of observations in the test is one less than that in the model.

The t-statistic for the coefficient on the previous period's squared residuals is greater than 4.1. Therefore, Miller easily rejects the null hypothesis that the variance of the error does not depend on the variance of previous errors. Consequently, the test statistics she computed in Table 5 are not valid, and she should not use them in deciding her investment strategy.

Table 17 Test for ARCH(1) in an AR(1) Model Residuals from Monthly CPI Inflation at an Annual Rate February 1984–December 2013

Regression Statistics	
R-squared	0.0467
Standard error	24.3630
Observations	358
Durbin–Watson	2.1046

	Coefficient	Standard Error	t-Statistic
Intercept	9.1666	1.4228	6.4426
$\hat{\varepsilon}^2_{t-1}$	0.2161	0.0518	4.1718

Source: US Bureau of Labor Statistics.

It is possible Miller's conclusion—that the AR(1) model for monthly inflation has ARCH in the errors—may have been due to the sample period employed (1984 to 2013). In Example 9, she used a shorter sample period of 2007 to 2013 and concluded that monthly CPI inflation follows an AR(1) process. (These results were shown in Table 8.) Table 17 shows that errors for a time-series model of inflation for the entire sample (1984 to 2013) have ARCH errors. Do the errors estimated with a shorter sample period (2007 to 2013) also display

ARCH? For the shorter sample period, Miller estimated an AR(1) model using monthly inflation data. Now she tests to see whether the errors display ARCH. Table 18 shows the results.

In this sample, the coefficient on the previous period's squared residual is quite small and has a *t*-statistic of only 1.3713. Consequently, Miller fails to reject the null hypothesis that the errors in this regression have no autoregressive conditional heteroskedasticity. This is additional evidence that the AR(1) model for 2007 to 2013 is a good fit. The error variance appears to be homoskedastic, and Miller can rely on the *t*-statistics. This result again confirms that a single AR process for the entire 1984–2013 period is misspecified (it does not describe the data well).

Table 18 Test for ARCH(1) in an AR(1) Model Monthly CPI Inflation at an Annual Rate February 2007–December 2013

Regression Statistics	
R-squared	0.0230
Standard error	35.218
Observations	82
Durbin–Watson	2.0521

	Coefficient	Standard Error	*t*-Statistic
Intercept	16.6213	4.4632	3.7241
$\hat{\varepsilon}^2_{t-1}$	0.1518	0.1107	1.3713

Source: US Bureau of Labor Statistics.

Suppose a model contains ARCH(1) errors. What are the consequences of that fact? First, if ARCH exists, the standard errors for the regression parameters will not be correct. We will need to use generalized least squares[39] or other methods that correct for heteroskedasticity to correctly estimate the standard error of the parameters in the time-series model. Second, if ARCH exists and we have it modeled, for example as ARCH(1), we can predict the variance of the errors. Suppose, for instance, that we want to predict the variance of the error in inflation using the estimated parameters from Table 17: $\hat{\sigma}^2_t = 9.1666 + 0.2161\hat{\varepsilon}^2_{t-1}$. If the error in one period were 0 percent, the predicted variance of the error in the next period would be 9.1666 + 0.2161(0) = 9.1666. If the error in one period were 1 percent, the predicted variance of the error in the next period would be 9.1666 + 0.2161(1^2) = 9.3827.

Engle and other researchers have suggested many generalizations of the ARCH(1) model, including ARCH(p) and generalized autoregressive conditional heteroskedasticity (GARCH) models. In an ARCH(p) model, the variance of the error term in the current period depends linearly on the squared errors from the previous p periods: $\sigma^2_t = a_0 + a_1\varepsilon^2_{t-1} + \cdots + a_p\varepsilon^2_{t-p}$. GARCH models are similar to ARMA models

39 See Greene (2018).

of the error variance in a time series. Just like ARMA models, GARCH models can be finicky and unstable: Their results can depend greatly on the sample period and the initial guesses of the parameters in the GARCH model. Financial analysts who use GARCH models should be well aware of how delicate these models can be, and they should examine whether GARCH estimates are robust to changes in the sample and the initial guesses about the parameters.[40]

10 REGRESSIONS WITH MORE THAN ONE TIME SERIES

Up to now, we have discussed time-series models only for one time series. Although in the readings on correlation and regression and on multiple regression we used linear regression to analyze the relationship among different time series, in those readings we completely ignored unit roots. A time series that contains a unit root is not covariance stationary. If any time series in a linear regression contains a unit root, ordinary least squares estimates of regression test statistics may be invalid.

To determine whether we can use linear regression to model more than one time series, let us start with a single independent variable; that is, there are two time series, one corresponding to the dependent variable and one corresponding to the independent variable. We will then extend our discussion to multiple independent variables.

We first use a unit root test, such as the Dickey–Fuller test, for each of the two time series to determine whether either of them has a unit root.[41] There are several possible scenarios related to the outcome of these tests. One possible scenario is that we find that neither of the time series has a unit root. Then we can safely use linear regression to test the relations between the two time series. Otherwise, we may have to use additional tests, as we discuss later in this section.

EXAMPLE 18

Unit Roots and the Fisher Effect

In Example 8 in the reading on multiple regression, we examined the Fisher effect by estimating the regression relation between expected inflation and US Treasury bill (T-bill) returns. We used 181 quarterly observations on expected inflation rates and T-Bill returns from the sample period extending from the fourth quarter of 1968 through the fourth quarter of 2013. We used linear regression to analyze the relationship between the two time series. The results of this regression would be valid if both the time series are covariance stationary; that is, neither of the two time series has a unit root. So, if we compute the Dickey–Fuller *t*-test statistic of the hypothesis of a unit root separately for each time series and find that we can reject the null hypothesis that the T-bill return series has a unit root and the null hypothesis that the expected inflation time series has a unit root, then we can use linear regression to analyze the relation between the two series. In that case, the results of our analysis of the Fisher effect would be valid.

40 For more on ARCH, GARCH, and other models of time-series variance, see Hamilton (1994).

41 For theoretical details of unit root tests, see Greene (2018) or Tsay (2010). Unit root tests are available in some econometric software packages, such as EViews.

A second possible scenario is that we reject the hypothesis of a unit root for the independent variable but fail to reject the hypothesis of a unit root for the dependent variable. In this case, the error term in the regression would not be covariance stationary. Therefore, one or more of the following linear regression assumptions would be violated: 1) that the expected value of the error term is 0, 2) that the variance of the error term is constant for all observations, and 3) that the error term is uncorrelated across observations. Consequently, the estimated regression coefficients and standard errors would be inconsistent. The regression coefficients might appear significant, but those results would be spurious.[42] Thus we should not use linear regression to analyze the relation between the two time series in this scenario.

A third possible scenario is the reverse of the second scenario: We reject the hypothesis of a unit root for the dependent variable but fail to reject the hypothesis of a unit root for the independent variable. In this case also, like the second scenario, the error term in the regression would not be covariance stationary, and we cannot use linear regression to analyze the relation between the two time series.

EXAMPLE 19

Unit Roots and Predictability of Stock Market Returns by Price-to-Earnings Ratio

Johann de Vries is analyzing the performance of the South African stock market. He examines whether the percentage change in the Johannesburg Stock Exchange (JSE) All Share Index can be predicted by the price-to-earnings ratio (P/E) for the index. Using monthly data from January 1994 to December 2013, he runs a regression using $(P_t - P_{t-1})/P_{t-1}$ as the dependent variable and P_{t-1}/E_{t-2} as the independent variable, where P_t is the value of the JSE index at time t and E_t is the earnings on the index. De Vries finds that the regression coefficient is negative and statistically significant and the value of the R-squared for the regression is quite high. What additional analysis should he perform before accepting the regression as valid?

De Vries needs to perform unit root tests for each of the two time series. If one of the two time series has a unit root, implying that it is not stationary, the results of the linear regression are not meaningful and cannot be used to conclude that stock market returns are predictable by P/E.[43]

The next possibility is that both time series have a unit root. In this case, we need to establish whether the two time series are **cointegrated** before we can rely on regression analysis.[44] Two time series are cointegrated if a long-term financial or economic relationship exists between them such that they do not diverge from each other without bound in the long run. For example, two time series are cointegrated if they share a common trend.

In the fourth scenario, both time series have a unit root but are not cointegrated. In this scenario, as in the second and third scenarios above, the error term in the linear regression will not be covariance stationary, some regression assumptions will be violated, the regression coefficients and standard errors will not be consistent, and we cannot use them for hypothesis tests. Consequently, linear regression of one variable on the other would be meaningless.

42 The problem of spurious regression for nonstationary time series was first discussed by Granger and Newbold (1974).

43 Barr and Kantor (1999) contains evidence that the P/E time series is nonstationary.

44 Engle and Granger (1987) first discussed cointegration.

Finally, the fifth possible scenario is that both time series have a unit root, but they are cointegrated. In this case, the error term in the linear regression of one time series on the other will be covariance stationary. Accordingly, the regression coefficients and standard errors will be consistent, and we can use them for hypothesis tests. However, we should be very cautious in interpreting the results of a regression with cointegrated variables. The cointegrated regression estimates the long-term relation between the two series but may not be the best model of the short-term relation between the two series. Short-term models of cointegrated series (error correction models) are discussed in Engle and Granger (1987) and Tsay (2010), but these are specialist topics.

Now let us look at how we can test for cointegration between two time series that each have a unit root as in the last two scenarios above.[45] Engle and Granger suggest this test: If y_t and x_t are both time series with a unit root, we should do the following:

1 Estimate the regression $y_t = b_0 + b_1x_t + \varepsilon_t$.

2 Test whether the error term from the regression in Step 1 has a unit root using a Dickey–Fuller test. Because the residuals are based on the estimated coefficients of the regression, we cannot use the standard critical values for the Dickey–Fuller test. Instead, we must use the critical values computed by Engle and Granger, which take into account the effect of uncertainty about the regression parameters on the distribution of the Dickey–Fuller test.

3 If the (Engle–Granger) Dickey–Fuller test fails to reject the null hypothesis that the error term has a unit root, then we conclude that the error term in the regression is not covariance stationary. Therefore, the two time series are not cointegrated. In this case any regression relation between the two series is spurious.

4 If the (Engle–Granger) Dickey–Fuller test rejects the null hypothesis that the error term has a unit root, then we may assume that the error term in the regression is covariance stationary and that the two time series are cointegrated. The parameters and standard errors from linear regression will be consistent and will let us test hypotheses about the long-term relation between the two series.

45 Consider a time series, x_t, that has a unit root. For many such financial and economic time series, the first difference of the series, $x_t - x_{t-1}$, is stationary. We say that such a series, whose first difference is stationary, has a *single* unit root. However, for some time series, even the first difference may not be stationary and further differencing may be needed to achieve stationarity. Such a time series is said to have *multiple* unit roots. In this section, we consider only the case in which each nonstationary series has a single unit root (which is quite common).

EXAMPLE 20

Testing for Cointegration between Intel Sales and Nominal GDP

Suppose we want to test whether the natural log of Intel's sales and the natural log of GDP are cointegrated (that is, whether there is a long-term relation between GDP and Intel sales). We want to test this hypothesis using quarterly data from the first quarter of 1995 through the fourth quarter of 2013. Here are the steps:

1. Test whether the two series each have a unit root. If we cannot reject the null hypothesis of a unit root for both series, implying that both series are nonstationary, we must then test whether the two series are cointegrated.
2. Having established that each series has a unit root, we estimate the regression $\ln(\text{Intel Sales}_t) = b_0 + b_1 \ln \text{GDP}_t + \varepsilon_t$, then conduct the (Engle–Granger) Dickey–Fuller test of the hypothesis that there is a unit root in the error term of this regression using the residuals from the estimated regression. If we reject the null hypothesis of a unit root in the error term of the regression, we reject the null hypothesis of no cointegration. That is, the two series would be cointegrated. If the two series are cointegrated, we can use linear regression to estimate the long-term relation between the natural log of Intel Sales and the natural log of GDP.

We have so far discussed models with a single independent variable. We now extend the discussion to a model with two or more independent variables, so that there are three or more time series. The simplest possibility is that none of the time series in the model has a unit root. Then, we can safely use multiple regression to test the relation among the time series.

EXAMPLE 21

Unit Roots and Returns to the Fidelity Select Technology Fund

In Example 3 in the reading on multiple regression, we used multiple linear regression to examine whether returns to either the S&P 500 Growth Index or the S&P 500 Value Index explain returns to the Fidelity Select Technology Portfolio using 60 monthly observations between January 2009 and December 2013. Of course, if any of the three time series has a unit root, then the results of our regression analysis may be invalid. Therefore, we could use a Dickey–Fuller test to determine whether any of these series has a unit root.

If we reject the hypothesis of unit roots for all three series, we can use linear regression to analyze the relation among the series. In that case the results of our analysis of the factors affecting returns to the Fidelity Select Technology Portfolio would be valid.

If at least one time series (the dependent variable or one of the independent variables) has a unit root while at least one time series (the dependent variable or one of the independent variables) does not, the error term in the regression cannot be covariance stationary. Consequently, we should not use multiple linear regression to analyze the relation among the time series in this scenario.

Another possibility is that each time series, including the dependent variable and each of the independent variables, has a unit root. If this is the case, we need to establish whether the time series are cointegrated. To test for cointegration, the procedure is similar to that for a model with a single independent variable. First, estimate the regression $y_t = b_0 + b_1x_{1t} + b_2x_{2t} + \ldots + b_kx_{kt} + \varepsilon_t$. Then conduct the (Engle–Granger) Dickey–Fuller test of the hypothesis that there is a unit root in the errors of this regression using the residuals from the estimated regression.

If we cannot reject the null hypothesis of a unit root in the error term of the regression, we cannot reject the null hypothesis of no cointegration. In this scenario, the error term in the multiple regression will not be covariance stationary, so we cannot use multiple regression to analyze the relationship among the time series.

If we can reject the null hypothesis of a unit root in the error term of the regression, we can reject the null hypothesis of no cointegration. However, modeling three or more time series that are cointegrated may be difficult. For example, an analyst may want to predict a retirement services company's sales based on the country's GDP and the total population over age 65. Although the company's sales, GDP, and the population over 65 may each have a unit root and be cointegrated, modeling the cointegration of the three series may be difficult, and doing so is beyond the scope of this volume. Analysts who have not mastered all these complex issues should avoid forecasting models with multiple time series that have unit roots: The regression coefficients may be inconsistent and may produce incorrect forecasts.

11 OTHER ISSUES IN TIME SERIES

Time-series analysis is an extensive topic and includes many highly complex issues. Our objective in this reading has been to present those issues in time series that are the most important for financial analysts and can also be handled with relative ease. In this section, we briefly discuss some of the issues that we have not covered but could be useful for analysts.

In this reading, we have shown how to use time-series models to make forecasts. We have also introduced the RMSE as a criterion for comparing forecasting models. However, we have not discussed measuring the uncertainty associated with forecasts made using time-series models. The uncertainty of these forecasts can be very large, and should be taken into account when making investment decisions. Fortunately, the same techniques apply to evaluating the uncertainty of time-series forecasts as apply to evaluating the uncertainty about forecasts from linear regression models. To accurately evaluate forecast uncertainty, we need to consider both the uncertainty about the error term and the uncertainty about the estimated parameters in the time-series model. Evaluating this uncertainty is fairly complicated when using regressions with more than one independent variable.

In this reading, we used the US CPI inflation series to illustrate some of the practical challenges analysts face in using time-series models. We used information on US Federal Reserve policy to explore the consequences of splitting the inflation series in two. In financial time-series work, we may suspect that a time series has more than one regime but lack the information to attempt to sort the data into different regimes. If you face such a problem, you may want to investigate other methods, especially switching regression models, to identify multiple regimes using only the time series itself.

If you are interested in these and other advanced time-series topics, you can learn more in Diebold (2008) and Tsay (2010).

SUGGESTED STEPS IN TIME-SERIES FORECASTING

12

The following is a step-by-step guide to building a model to predict a time series.

1 Understand the investment problem you have, and make an initial choice of model. One alternative is a regression model that predicts the future behavior of a variable based on hypothesized causal relationships with other variables. Another is a time-series model that attempts to predict the future behavior of a variable based on the past behavior of the same variable.

2 If you have decided to use a time-series model, compile the time series and plot it to see whether it looks covariance stationary. The plot might show important deviations from covariance stationarity, including the following:

- a linear trend;
- an exponential trend;
- seasonality; or
- a significant shift in the time series during the sample period (for example, a change in mean or variance).

3 If you find no significant seasonality or shift in the time series, then perhaps either a linear trend or an exponential trend will be sufficient to model the time series. In that case, take the following steps:

- Determine whether a linear or exponential trend seems most reasonable (usually by plotting the series).
- Estimate the trend.
- Compute the residuals.
- Use the Durbin–Watson statistic to determine whether the residuals have significant serial correlation. If you find no significant serial correlation in the residuals, then the trend model is sufficient to capture the dynamics of the time series and you can use that model for forecasting.

4 If you find significant serial correlation in the residuals from the trend model, use a more complex model, such as an autoregressive model. First, however, reexamine whether the time series is covariance stationary. Following is a list of violations of stationarity, along with potential methods to adjust the time series to make it covariance stationary:

- If the time series has a linear trend, first-difference the time series.
- If the time series has an exponential trend, take the natural log of the time series and then first-difference it.
- If the time series shifts significantly during the sample period, estimate different time-series models before and after the shift.
- If the time series has significant seasonality, include seasonal lags (discussed in Step 7).

5 After you have successfully transformed a raw time series into a covariance-stationary time series, you can usually model the transformed series with a short autoregression.[46] To decide which autoregressive model to use, take the following steps:

46 Most financial time series can be modeled using an autoregressive process. For a few time series, a moving-average model may fit better. To see if this is the case, examine the first five or six autocorrelations of the time series. If the autocorrelations suddenly drop to 0 after the first q autocorrelations, a moving-average model (of order q) is appropriate. If the autocorrelations start large and decline gradually, an autoregressive model is appropriate.

- Estimate an AR(1) model.
- Test to see whether the residuals from this model have significant serial correlation.
- If you find no significant serial correlation in the residuals, you can use the AR(1) model to forecast.

6 If you find significant serial correlation in the residuals, use an AR(2) model and test for significant serial correlation of the residuals of the AR(2) model.
- If you find no significant serial correlation, use the AR(2) model.
- If you find significant serial correlation of the residuals, keep increasing the order of the AR model until the residual serial correlation is no longer significant.

7 Your next move is to check for seasonality. You can use one of two approaches:
- Graph the data and check for regular seasonal patterns.
- Examine the data to see whether the seasonal autocorrelations of the residuals from an AR model are significant (for example, the fourth autocorrelation for quarterly data) and whether the autocorrelations before and after the seasonal autocorrelations are significant. To correct for seasonality, add seasonal lags to your AR model. For example, if you are using quarterly data, you might add the fourth lag of a time series as an additional variable in an AR(1) or an AR(2) model.

8 Next, test whether the residuals have autoregressive conditional heteroskedasticity. To test for ARCH(1), for example, do the following:
- Regress the squared residual from your time-series model on a lagged value of the squared residual.
- Test whether the coefficient on the squared lagged residual differs significantly from 0.
- If the coefficient on the squared lagged residual does not differ significantly from 0, the residuals do not display ARCH and you can rely on the standard errors from your time-series estimates.
- If the coefficient on the squared lagged residual does differ significantly from 0, use generalized least squares or other methods to correct for ARCH.

9 Finally, you may also want to perform tests of the model's out-of-sample forecasting performance to see how the model's out-of-sample performance compares to its in-sample performance.

Using these steps in sequence, you can be reasonably sure that your model is correctly specified.

SUMMARY

- The predicted trend value of a time series in period t is $\hat{b}_0 + \hat{b}_1 t$ in a linear trend model; the predicted trend value of a time series in a log-linear trend model is $e^{\hat{b}_0 + \hat{b}_1 t}$.
- Time series that tend to grow by a constant amount from period to period should be modeled by linear trend models, whereas time series that tend to grow at a constant rate should be modeled by log-linear trend models.

- Trend models often do not completely capture the behavior of a time series, as indicated by serial correlation of the error term. If the Durbin–Watson statistic from a trend model differs significantly from 2, indicating serial correlation, we need to build a different kind of model.
- An autoregressive model of order *p,* denoted AR(*p*), uses *p* lags of a time series to predict its current value: $x_t = b_0 + b_1x_{t-1} + b_2x_{t-2} + \ldots + b_px_{t-p} + \varepsilon_t$.
- A time series is covariance stationary if the following three conditions are satisfied: First, the expected value of the time series must be constant and finite in all periods. Second, the variance of the time series must be constant and finite in all periods. Third, the covariance of the time series with itself for a fixed number of periods in the past or future must be constant and finite in all periods. Inspection of a nonstationary time-series plot may reveal an upward or downward trend (nonconstant mean) and/or nonconstant variance. The use of linear regression to estimate an autoregressive time-series model is not valid unless the time series is covariance stationary.
- For a specific autoregressive model to be a good fit to the data, the autocorrelations of the error term should be 0 at all lags.
- A time series is mean reverting if it tends to fall when its level is above its long-run mean and rise when its level is below its long-run mean. If a time series is covariance stationary, then it will be mean reverting.
- The one-period-ahead forecast of a variable x_t from an AR(1) model made in period *t* for period *t* + 1 is $\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1x_t$. This forecast can be used to create the two-period ahead forecast from the model made in period *t*, $\hat{x}_{t+2} = \hat{b}_0 + \hat{b}_1x_{t+1}$. Similar results hold for AR(*p*) models.
- In-sample forecasts are the in-sample predicted values from the estimated time-series model. Out-of-sample forecasts are the forecasts made from the estimated time-series model for a time period different from the one for which the model was estimated. Out-of-sample forecasts are usually more valuable in evaluating the forecasting performance of a time-series model than are in-sample forecasts. The root mean squared error (RMSE), defined as the square root of the average squared forecast error, is a criterion for comparing the forecast accuracy of different time-series models; a smaller RMSE implies greater forecast accuracy.
- Just as in regression models, the coefficients in time-series models are often unstable across different sample periods. In selecting a sample period for estimating a time-series model, we should seek to assure ourselves that the time series was stationary in the sample period.
- A random walk is a time series in which the value of the series in one period is the value of the series in the previous period plus an unpredictable random error. If the time series is a random walk, it is not covariance stationary. A random walk with drift is a random walk with a nonzero intercept term. All random walks have unit roots. If a time series has a unit root, then it will not be covariance stationary.
- If a time series has a unit root, we can sometimes transform the time series into one that is covariance stationary by first-differencing the time series; we may then be able to estimate an autoregressive model for the first-differenced series.
- An *n*-period moving average of the current and past ($n - 1$) values of a time series, x_t, is calculated as $[x_t + x_{t-1} + \ldots + x_{t-(n-1)}]/n$.
- A moving-average model of order *q*, denoted MA(*q*), uses *q* lags of a random error term to predict its current value.

- The order q of a moving average model can be determined using the fact that if a time series is a moving-average time series of order q, its first q autocorrelations are nonzero while autocorrelations beyond the first q are zero.
- The autocorrelations of most autoregressive time series start large and decline gradually, whereas the autocorrelations of an MA(q) time series suddenly drop to 0 after the first q autocorrelations. This helps in distinguishing between autoregressive and moving-average time series.
- If the error term of a time-series model shows significant serial correlation at seasonal lags, the time series has significant seasonality. This seasonality can often be modeled by including a seasonal lag in the model, such as adding a term lagged four quarters to an AR(1) model on quarterly observations.
- The forecast made in time t for time t + 1 using a quarterly AR(1) model with a seasonal lag would be $x_{t+1} = \hat{b}_0 + \hat{b}_1 x_t + \hat{b}_2 x_{t-3}$.
- ARMA models have several limitations: the parameters in ARMA models can be very unstable; determining the AR and MA order of the model can be difficult; and even with their additional complexity, ARMA models may not forecast well.
- The variance of the error in a time-series model sometimes depends on the variance of previous errors, representing autoregressive conditional heteroskedasticity (ARCH). Analysts can test for first-order ARCH in a time-series model by regressing the squared residual on the squared residual from the previous period. If the coefficient on the squared residual is statistically significant, the time-series model has ARCH(1) errors.
- If a time-series model has ARCH(1) errors, then the variance of the errors in period t + 1 can be predicted in period t using the formula $\hat{\sigma}_{t+1}^2 = \hat{a}_0 + \hat{a}_1 \hat{\varepsilon}_t^2$.
- If linear regression is used to model the relationship between two time series, a test should be performed to determine whether either time series has a unit root:
  - If neither of the time series has a unit root, then we can safely use linear regression.
  - If one of the two time series has a unit root, then we should not use linear regression.
  - If both time series have a unit root and the time series are cointegrated, we may safely use linear regression; however, if they are not cointegrated, we should not use linear regression. The (Engle–Granger) Dickey–Fuller test can be used to determine if time series are cointegrated.

REFERENCES

Barr, G. D. I., and B. S. Kantor. 1999. “Price–Earnings Ratios on the Johannesburg Stock Exchange—Are They a Good Value?” *SA Journal of Accounting Research* 13 (1): 1–23.

Dickey, David A., and Wayne A. Fuller. 1979. “Distribution of the Estimators for Autoregressive Time Series with a Unit Root.” *Journal of the American Statistical Association*, 74: 427–431.

Diebold, Francis X. 2008. *Elements of Forecasting*. 4th ed. Cincinnati, OH: South-Western.

Engle, Robert F., and Clive W.J. Granger. 1987. “Co-Integration and Error Correction: Representation, Estimation, and Testing.” *Econometrica* 55 (2): 251–76.

Garbade, Kenneth. 1982. *Securities Markets*. New York: McGraw-Hill.

Granger, Clive W.J., and Paul Newbold. 1974. “Spurious Regressions in Econometrics.” *Journal of Econometrics* 2:111–20.

Hamilton, James D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Tsay, Ruey S. 2010. *Analysis of Financial Time Series*. 3rd ed. New York: John Wiley.

PRACTICE PROBLEMS

Note: In the Problems and Solutions for this reading, we use the hat (ˆ) to indicate an estimate if we are trying to differentiate between an estimated and an actual value. However, we suppress the hat when we are clearly showing regression output.

1 The civilian unemployment rate (UER) is an important component of many economic models. Table 1 gives regression statistics from estimating a linear trend model of the unemployment rate: $UER_t = b_0 + b_1 t + \varepsilon_t$.

Table 1 Estimating a Linear Trend in the Civilian Unemployment Rate Monthly Observations, January 1996–December 2000

Regression Statistics	
R-squared	0.9314
Standard error	0.1405
Observations	60
Durbin–Watson	0.9099

	Coefficient	Standard Error	*t*-Statistic
Intercept	5.5098	0.0367	150.0363
Trend	–0.0294	0.0010	–28.0715

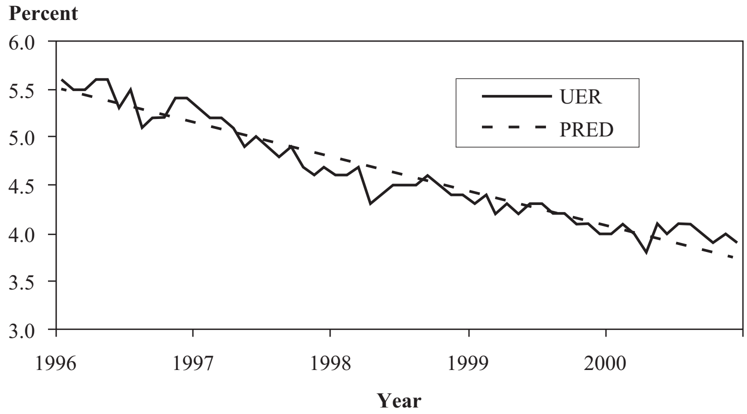
A Using the regression output in the above table, what is the model's prediction of the unemployment rate for July 1996?

B How should we interpret the Durbin–Watson (DW) statistic for this regression? What does the value of the DW statistic say about the validity of a *t*-test on the coefficient estimates?

2 Figure 1 compares the predicted civilian unemployment rate (PRED) with the actual civilian unemployment rate (UER) from January 1996 to December 2000. The predicted results come from estimating the linear time trend model $UER_t = b_0 + b_1 t + \varepsilon_t$.

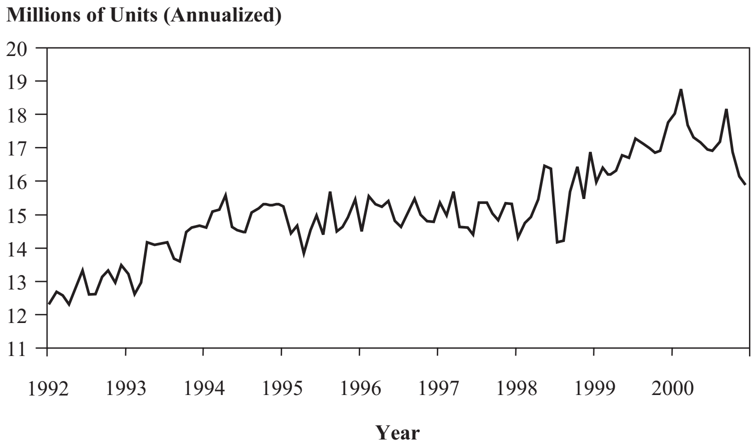
What can we conclude about the appropriateness of this model?

Figure 1 Predicted and Actual Civilian Unemployment Rates



3 You have been assigned to analyze automobile manufacturers and as a first step in your analysis, you decide to model monthly sales of lightweight vehicles to determine sales growth in that part of the industry. Figure 2 gives lightweight vehicle monthly sales (annualized) from January 1992 to December 2000.

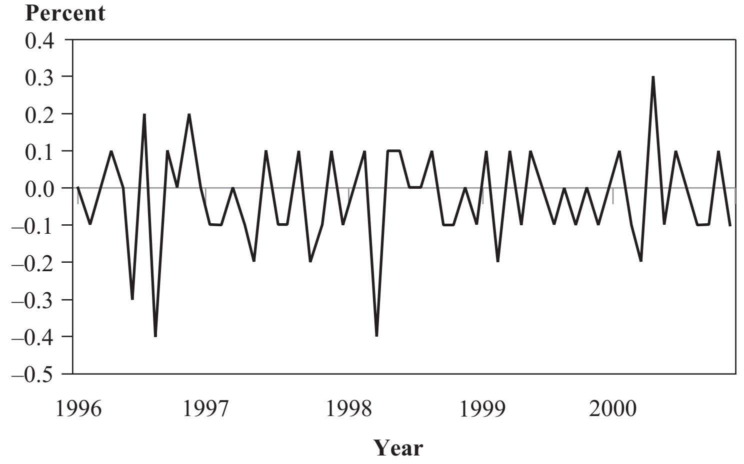
Figure 2 Lightweight Vehicle Sales



Monthly sales in the lightweight vehicle sector, Sales_t, have been increasing over time, but you suspect that the growth rate of monthly sales is relatively constant. Write the simplest time-series model for Sales_t that is consistent with your perception.

4 Figure 3 shows a plot of the first differences in the civilian unemployment rate (UER) between January 1996 and December 2000, $\Delta\text{UER}_t = \text{UER}_t - \text{UER}_{t-1}$.

Figure 3 Change in Civilian Unemployment Rate



A Has differencing the data made the new series, ΔUER_t, covariance stationary? Explain your answer.

B Given the graph of the change in the unemployment rate shown in the figure, describe the steps we should take to determine the appropriate autoregressive time-series model specification for the series ΔUER_t.

5 Table 2 gives the regression output of an AR(1) model on first differences in the unemployment rate. Describe how to interpret the DW statistic for this regression.

Table 2 Estimating an AR(1) Model of Changes in the Civilian Unemployment Rate Monthly Observations, March 1996–December 2000

Regression Statistics	
R-squared	0.2184
Standard error	0.1202
Observations	58
Durbin–Watson	2.1852

	Coefficient	Standard Error	t-Statistic
Intercept	–0.0405	0.0161	–2.5110
ΔUER_{t-1}	–0.4674	0.1181	–3.9562

6 Assume that changes in the civilian unemployment rate are covariance stationary and that an AR(1) model is a good description for the time series of changes in the unemployment rate. Specifically, we have $\Delta UER_t = -0.0405 - 0.4674\Delta UER_{t-1}$ (using the coefficient estimates given in the previous problem). Given this equation, what is the mean-reverting level to which changes in the unemployment rate converge?

7 Suppose the following model describes changes in the civilian unemployment rate: $\Delta UER_t = -0.0405 - 0.4674\Delta UER_{t-1}$. The current change (first difference) in the unemployment rate is 0.0300. Assume that the mean-reverting level for changes in the unemployment rate is –0.0276.

A What is the best prediction of the next change?

B What is the prediction of the change following the next change?

C Explain your answer to Part B in terms of equilibrium.

8 Table 3 gives the actual sales, log of sales, and changes in the log of sales of Cisco Systems for the period 1Q:2001 to 4Q:2001.

Table 3

Date Quarter: Year	Actual Sales ($ Millions)	Log of Sales	Changes in Log of Sales $\Delta\ln(\text{Sales}_t)$
1Q:2001	6,519	8.7825	0.1308
2Q:2001	6,748	8.8170	0.0345
3Q:2001	4,728	8.4613	–0.3557
4Q:2001	4,298	8.3659	–0.0954
1Q:2002			
2Q:2002			

Forecast the first- and second-quarter sales of Cisco Systems for 2002 using the regression $\Delta\ln(\text{Sales}_t) = 0.0661 + 0.4698\Delta\ln(\text{Sales}_{t-1})$.

9 Table 4 gives the actual change in the log of sales of Cisco Systems from 1Q:2001 to 4Q:2001, along with the forecasts from the regression model $\Delta\ln(\text{Sales}_t) = 0.0661 + 0.4698\Delta\ln(\text{Sales}_{t-1})$ estimated using data from 3Q:1991 to 4Q:2000. (Note that the observations after the fourth quarter of 2000 are out of sample.)

Table 4

Date	Actual Values of Changes in the Log of Sales $\Delta\ln(\text{Sales}_t)$	Forecast Values of Changes in the Log of Sales $\Delta\ln(\text{Sales}_t)$
1Q:2001	0.1308	0.1357
2Q:2001	0.0345	0.1299
3Q:2001	–0.3557	0.1271
4Q:2001	–0.0954	0.1259

A Calculate the RMSE for the out-of-sample forecast errors.

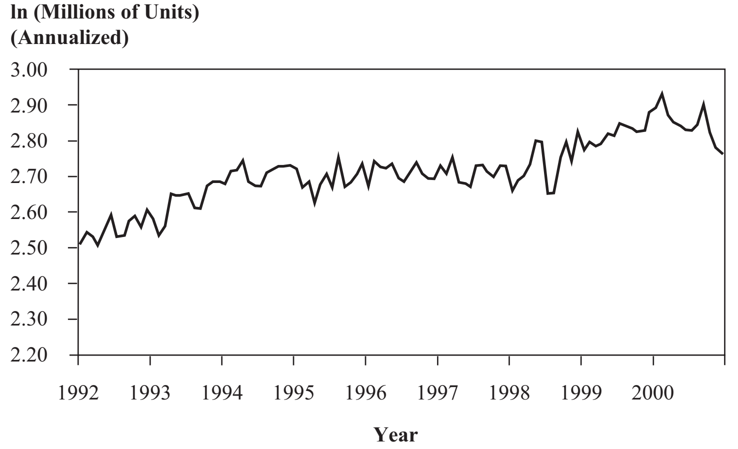
B Compare the forecasting performance of the model given with that of another model having an out-of-sample RMSE of 20 percent.

10 **A** The AR(1) model for the civilian unemployment rate, $\Delta UER_t = -0.0405 - 0.4674\Delta UER_{t-1}$, was developed with five years of data. What would be the drawback to using the AR(1) model to predict changes in the civilian unemployment rate 12 months or more ahead, as compared with one month ahead?

B For purposes of estimating a predictive equation, what would be the drawback to using 30 years of civilian unemployment data rather than only five years?

11 Figure 4 shows monthly observations on the natural log of lightweight vehicle sales, ln ($Sales_t$), for the period January 1992 to December 2000.

Figure 4 Lightweight Vehicle Sales

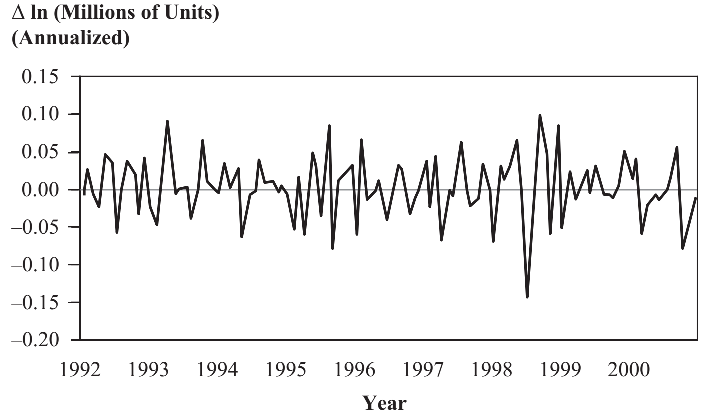


A Using the figure, comment on whether the specification ln ($Sales_t$) = b_0 + b_1[ln ($Sales_{t-1}$)] + ε_t is appropriate.

B State an appropriate transformation of the time series.

12 Figure 5 shows a plot of first differences in the log of monthly lightweight vehicle sales over the same period as in Problem 11. Has differencing the data made the resulting series, Δln ($Sales_t$) = ln ($Sales_t$) – ln ($Sales_{t-1}$), covariance stationary?

Figure 5 Change in Natural Log of Lightweight Vehicle Sales



13 Using monthly data from January 1992 to December 2000, we estimate the following equation for lightweight vehicle sales: Δln ($Sales_t$) = 2.7108 + 0.3987Δln ($Sales_{t-1}$) + ε_t. Table 5 gives sample autocorrelations of the errors from this model.

Table 5 Different Order Autocorrelations of Differences in the Logs of Vehicle Sales

Lag	Autocorrelation	Standard Error	t-Statistic
1	0.9358	0.0962	9.7247
2	0.8565	0.0962	8.9005
3	0.8083	0.0962	8.4001
4	0.7723	0.0962	8.0257
5	0.7476	0.0962	7.7696
6	0.7326	0.0962	7.6137
7	0.6941	0.0962	7.2138
8	0.6353	0.0962	6.6025
9	0.5867	0.0962	6.0968
10	0.5378	0.0962	5.5892
11	0.4745	0.0962	4.9315
12	0.4217	0.0962	4.3827

A Use the information in the table to assess the appropriateness of the specification given by the equation.

B If the residuals from the AR(1) model above violate a regression assumption, how would you modify the AR(1) specification?

14 Figure 6 shows the quarterly sales of Cisco Systems from 1Q:1991 to 4Q:2000.

Figure 6 Quarterly Sales at Cisco

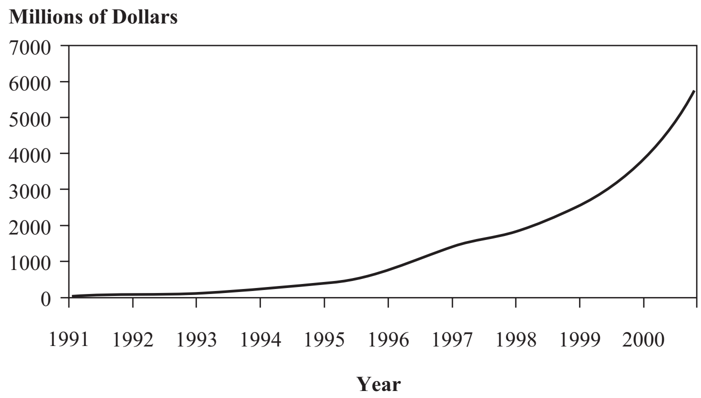


Table 6 gives the regression statistics from estimating the model $\Delta\ln(\text{Sales}_t) = b_0 + b_1\Delta\ln(\text{Sales}_{t-1}) + \varepsilon_t$.

Table 6 Change in the Natural Log of Sales for Cisco Systems Quarterly Observations, 3Q:1991–4Q:2000

Regression Statistics	
R-squared	0.2899
Standard error	0.0408
Observations	38
Durbin–Watson	1.5707

	Coefficient	Standard Error	t-Statistic
Intercept	0.0661	0.0175	3.7840
Δln (Sales$_{t-1}$)	0.4698	0.1225	3.8339

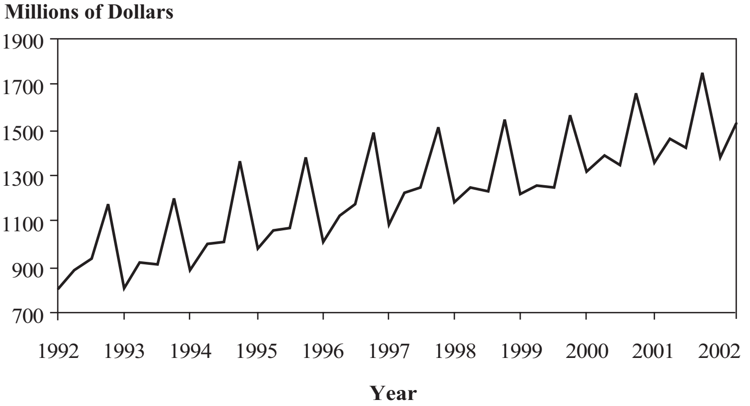
A Describe the salient features of the quarterly sales series.

B Describe the procedures we should use to determine whether the AR(1) specification is correct.

C Assuming the model is correctly specified, what is the long-run change in the log of sales toward which the series will tend to converge?

15 Figure 7 shows the quarterly sales of Avon Products from 1Q:1992 to 2Q:2002. Describe the salient features of the data shown.

Figure 7 Quarterly Sales at Avon



16 Table 7 below shows the autocorrelations of the residuals from an AR(1) model fit to the changes in the gross profit margin (GPM) of The Home Depot, Inc.

Table 7 Autocorrelations of the Residuals from Estimating the Regression $\Delta GPM_t = 0.0006 - 0.3330_1 \Delta GPM_{t-1} + \varepsilon_t$ 1Q:1992–4Q:2001 (40 Observations)

Lag	Autocorrelation
1	–0.1106
2	–0.5981

Table 7 (Continued)

Lag	Autocorrelation
3	–0.1525
4	0.8496
5	–0.1099

Table 8 shows the output from a regression on changes in the GPM for Home Depot, where we have changed the specification of the AR regression.

Table 8 Change in Gross Profit Margin for Home Depot 1Q:1992–4Q:2001

Regression Statistics	
R-squared	0.9155
Standard error	0.0057
Observations	40
Durbin–Watson	2.6464

	Coefficient	Standard Error	t-Statistic
Intercept	–0.0001	0.0009	–0.0610
ΔGPM_{t-1}	–0.0608	0.0687	–0.8850
ΔGPM_{t-4}	0.8720	0.0678	12.8683

A Identify the change that was made to the regression model.

B Discuss the rationale for changing the regression specification.

17 Suppose we decide to use an autoregressive model with a seasonal lag because of the seasonal autocorrelation in the previous problem. We are modeling quarterly data, so we estimate Equation 15: $(\ln \text{Sales}_t - \ln \text{Sales}_{t-1}) = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + b_2(\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}) + \varepsilon_t$. Table 9 shows the regression statistics from this equation.

Table 9 Log Differenced Sales: AR(1) Model with Seasonal Lag Johnson & Johnson Quarterly Observations, January 1985–December 2001

Regression Statistics	
R-squared	0.4220
Standard error	0.0318

(continued)

Table 9 (Continued)

Regression Statistics	
Observations	68
Durbin–Watson	1.8784

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0121	0.0053	2.3055
Lag 1	–0.0839	0.0958	–0.8757
Lag 4	0.6292	0.0958	6.5693

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	0.0572	0.1213	0.4720
2	–0.0700	0.1213	–0.5771
3	0.0065	0.1213	–0.0532
4	–0.0368	0.1213	–0.3033

A Using the information in Table 9, determine if the model is correctly specified.

B If sales grew by 1 percent last quarter and by 2 percent four quarters ago, use the model to predict the sales growth for this quarter.

18 Describe how to test for autoregressive conditional heteroskedasticity (ARCH) in the residuals from the AR(1) regression on first differences in the civilian unemployment rate, $\Delta UER_t = b_0 + b_1 \Delta UER_{t-1} + \varepsilon_t$.

19 Suppose we want to predict the annualized return of the five-year T-bill using the annualized return of the three-month T-bill with monthly observations from January 1993 to December 2002. Our analysis produces the data shown in Table 10.

Table 10 Regression with 3-Month T-Bill as the Independent Variable and 5-Year Treasury Bill as the Dependent Variable Monthly Observations, January 1993 to December 2002

Regression Statistics	
R-squared	0.5829
Standard error	0.6598
Observations	120
Durbin–Watson	0.1130

	Coefficient	Standard Error	*t*-Statistic
Intercept	3.0530	0.2060	14.8181
Three-month	0.5722	0.0446	12.8408

Can we rely on the regression model in Table 10 to produce meaningful predictions? Specify what problem might be a concern with this regression.

The following information relates to Questions 20–26

Angela Martinez, an energy sector analyst at an investment bank, is concerned about the future level of oil prices and how it might affect portfolio values. She is considering whether to recommend a hedge for the bank portfolio's exposure to changes in oil prices. Martinez examines West Texas Intermediate (WTI) monthly crude oil price data, expressed in US dollars per barrel, for the 181-month period from August 2000 through August 2015. The end-of-month WTI oil price was \$51.16 in July 2015 and \$42.86 in August 2015 (Month 181).

After reviewing the time-series data, Martinez determines that the mean and variance of the time series of oil prices are not constant over time. She then runs the following four regressions using the WTI time-series data.

- Linear trend model: Oil price$_t$ = b_0 + b$_1 t$ + e_t
- Log-linear trend model: ln Oil price$_t$ = b_0 + $b_1 t$ + e_t
- AR(1) model: Oil price$_t$ = b_0 + b_1Oil price$_{t-1}$ + e_t
- AR(2) model: Oil price$_t$ = b_0 + b_1Oil price$_{t-1}$ + b_2Oil price$_{t-2}$ + e_t

Exhibit 1 presents selected data from all four regressions, and Exhibit 2 presents selected autocorrelation data from the AR(1) models.

Exhibit 1 Crude Oil Price per Barrel, August 2000–August 2015

	Regression Statistics (t-statistics for coefficients are reported in parentheses)			
	Linear	**Log-Linear**	**AR(1)**	**AR(2)**
R^2	0.5703	0.6255	0.9583	0.9656
Standard error	18.6327	0.3034	5.7977	5.2799
Observations	181	181	180	179
Durbin–Watson	0.10	0.08	1.16	2.08
RMSE			2.0787	2.0530
Coefficients:				
Intercept	28.3278	3.3929	1.5948	2.0017
	(10.1846)	(74.9091)	(1.4610)	(1.9957)
t (Trend)	0.4086	0.0075		
	(15.4148)	(17.2898)		
Oil Price$_{t-1}$			0.9767	1.3946
			(63.9535)	(20.2999)
Oil Price$_{t-2}$				–0.4249
				(–6.2064)

In Exhibit 1, at the 5% significance level, the lower critical value for the Durbin–Watson test statistic is 1.75 for both the linear and log-linear regressions.

Exhibit 2 Autocorrelations of the Residual from AR(1) Model

Lag	Autocorrelation	*t*-Statistic
1	0.4157	5.5768
2	0.2388	3.2045
3	0.0336	0.4512
4	–0.0426	–0.5712

Note: At the 5% significance level, the critical value for a *t*-statistic is 1.97.

After reviewing the data and regression results, Martinez draws the following conclusions.

Conclusion 1 The time series for WTI oil prices is covariance stationary.

Conclusion 2 Out-of-sample forecasting using the AR(1) model appears to be more accurate than that of the AR(2) model.

20 Based on Exhibit 1, the predicted WTI oil price for October 2015 using the linear trend model is *closest* to:

A $29.15.

B $74.77.

C $103.10.

21 Based on Exhibit 1, the predicted WTI oil price for September 2015 using the log-linear trend model is *closest* to:

A $29.75.

B $29.98.

C $116.50.

22 Based on the regression output in Exhibit 1, there is evidence of positive serial correlation in the errors in:

A the linear trend model but not the log-linear trend model.

B both the linear trend model and the log-linear trend model.

C neither the linear trend model nor the log-linear trend model.

23 Martinez's Conclusion 1 is:

A correct.

B incorrect because the mean and variance of WTI oil prices are not constant over time.

C incorrect because the Durbin–Watson statistic of the AR(2) model is greater than 1.75.

24 Based on Exhibit 1, the forecasted oil price in September 2015 based on the AR(2) model is *closest* to:

A $38.03.

B $40.04.

C $61.77.

25 Based on the data for the AR(1) model in Exhibits 1 and 2, Martinez can conclude that the:

A residuals are not serially correlated.

B autocorrelations do not differ significantly from zero.

C standard error for each of the autocorrelations is 0.0745.

26 Based on the mean-reverting level implied by the AR(1) model regression output in Exhibit 1, the forecasted oil price for September 2015 is *most likely* to be:

A less than \$42.86.

B equal to \$42.86.

C greater than \$42.86.

The following information relates to Question 27–35

Max Busse is an analyst in the research department of a large hedge fund. He was recently asked to develop a model to predict the future exchange rate between two currencies. Busse gathers monthly exchange rate data from the most recent 10-year period and runs a regression based on the following AR(1) model specification:

Regression 1: $x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$, where x_t is the exchange rate at time t.

Based on his analysis of the time series and the regression results, Busse reaches the following conclusions:

Conclusion 1 The variance of x_t increases over time.

Conclusion 2 The mean-reverting level is undefined.

Conclusion 3 b_0 does not appear to be significantly different from 0.

Busse decides to do additional analysis by first-differencing the data and running a new regression.

Regression 2: $y_t = b_0 + b_1 y_{t-1} + \varepsilon_t$, where $y_t = x_t - x_{t-1}$.

Exhibit 1 shows the regression results.

Exhibit 1 First-Differenced Exchange Rate AR(1) Model: Month-End Observations, Last 10 Years

Regression Statistics	
R^2	0.0017
Standard error	7.3336
Observations	118
Durbin–Watson	1.9937

	Coefficient	Standard Error	*t*-Statistic
Intercept	–0.8803	0.6792	–1.2960
$x_{t-1} - x_{t-2}$	0.0412	0.0915	0.4504

(continued)

Exhibit 1 (Continued)

Autocorrelations of the Residual			
Lag	**Autocorrelation**	**Standard Error**	***t*-Statistic**
1	0.0028	0.0921	0.0300
2	0.0205	0.0921	0.2223
3	0.0707	0.0921	0.7684
4	0.0485	0.0921	0.5271

Note: The critical t-statistic at the 5% significance level is 1.98.

Busse decides that he will need to test the data for nonstationarity using a Dickey–Fuller test. To do so, he knows he must model a transformed version of Regression 1.

Busse's next assignment is to develop a model to predict future quarterly sales for PoweredUP, Inc., a major electronics retailer. He begins by running the following regression:

Regression 3: $\ln \text{Sales}_t - \ln \text{Sales}_{t-1} = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + \varepsilon_t$.

Exhibit 2 presents the results of this regression.

Exhibit 2 Log Differenced Sales: AR(1) Model PoweredUP, Inc., Last 10 Years of Quarterly Sales

Regression Statistics	
R^2	0.2011
Standard error	0.0651
Observations	38
Durbin–Watson	1.9677

	Coefficient	Standard Error	t-Statistic
Intercept	0.0408	0.0112	3.6406
$\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}$	−0.4311	0.1432	−3.0099

Autocorrelations of the Residual			
Lag	**Autocorrelation**	**Standard Error**	***t*-Statistic**
1	0.0146	0.1622	0.0903
2	−0.1317	0.1622	−0.8119
3	−0.1123	0.1622	−0.6922
4	0.6994	0.1622	4.3111

Note: The critical t-statistic at the 5% significance level is 2.02.

Because the regression output from Exhibit 2 raises some concerns, Busse runs a different regression. These regression results, along with quarterly sales data for the past five quarters, are presented in Exhibits 3 and 4, respectively.

Exhibit 3 Log Differenced Sales: AR(1) Model with Seasonal Lag PoweredUP, Inc., Last 10 Years of Quarterly Sales

Regression Statistics

R^2	0.6788
Standard error	0.0424
Observations	35
Durbin–Watson	1.8799

	Coefficient	Standard Error	*t*-Statistic
Intercept	0.0092	0.0087	1.0582
$\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}$	–0.1279	0.1137	–1.1252
$\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}$	0.7239	0.1093	6.6209

Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	*t*-Statistic
1	0.0574	0.1690	0.3396
2	0.0440	0.1690	0.2604
3	0.1923	0.1690	1.1379
4	–0.1054	0.1690	–0.6237

Note: The critical *t*-statistic at the 5% significance level is 2.03.

Exhibit 4 Most Recent Quarterly Sales Data (in billions)

Dec 2015 (Sales_{t-1})	\$3.868
Sept 2015 (Sales_{t-2})	\$3.780
June 2015 (Sales_{t-3})	\$3.692
Mar 2014 (Sales_{t-4})	\$3.836
Dec 2014 (Sales_{t-5})	\$3.418

After completing his work on PoweredUP, Busse is asked to analyze the relationship of oil prices and the stock prices of three transportation companies. His firm wants to know whether the stock prices can be predicted by the price of oil. Exhibit 5 shows selected information from the results of his analysis.

Exhibit 5 Analysis Summary of Stock Prices for Three Transportation Stocks and the Price of Oil

	Unit Root?	Linear or Exponential Trend?	Serial Correlation of Residuals in Trend Model?	ARCH(1)?	Comments
Company #1	Yes	Exponential	Yes	Yes	Not co-integrated with oil price
Company #2	Yes	Linear	Yes	No	Co-integrated with oil price

(continued)

Exhibit 5 (Continued)

	Unit Root?	Linear or Exponential Trend?	Serial Correlation of Residuals in Trend Model?	ARCH(1)?	Comments
Company #3	No	Exponential	Yes	No	Not co-integrated with oil price
Oil Price	Yes				

To assess the relationship between oil prices and stock prices, Busse runs three regressions using the time series of each company's stock prices as the dependent variable and the time series of oil prices as the independent variable.

27 Which of Busse's conclusions regarding the exchange rate time series is consistent with both the properties of a covariance-stationary time series and the properties of a random walk?

A Conclusion 1

B Conclusion 2

C Conclusion 3

28 Based on the regression output in Exhibit 1, the first-differenced series used to run Regression 2 is consistent with:

A a random walk.

B covariance stationarity.

C a random walk with drift.

29 Based on the regression results in Exhibit 1, the *original* time series of exchange rates:

A has a unit root.

B exhibits stationarity.

C can be modeled using linear regression.

30 In order to perform the nonstationarity test, Busse should transform the Regression 1 equation by:

A adding the second lag to the equation.

B changing the regression's independent variable.

C subtracting the independent variable from both sides of the equation.

31 Based on the regression output in Exhibit 2, what should lead Busse to conclude that the Regression 3 equation is not correctly specified?

A The Durbin–Watson statistic

B The t-statistic for the slope coefficient

C The t-statistics for the autocorrelations of the residual

32 Based on the regression output in Exhibit 3 and sales data in Exhibit 4, the forecasted value of quarterly sales for March 2016 for PoweredUP is *closest* to:

A \$4.193 billion.

B \$4.205 billion.

C \$4.231 billion.

33 Based on Exhibit 5, Busse should conclude that the variance of the error terms for Company #1:

A is constant.

B can be predicted.

C is homoskedastic.

34 Based on Exhibit 5, for which company would the regression of stock prices on oil prices be expected to yield valid coefficients that could be used to estimate the long-term relationship between stock price and oil price?

A Company #1

B Company #2

C Company #3

35 Based on Exhibit 5, which single time-series model would *most likely* be appropriate for Busse to use in predicting the future stock price of Company #3?

A Log-linear trend model

B First-differenced AR(2) model

C First-differenced log AR(1) model

SOLUTIONS

1 A The estimated forecasting equation is $\text{UER}_t = 5.5098 - 0.0294(t)$. The data begin in January 1996, and July 1996 is period 7. Thus the linear trend model predicts the unemployment rate to be $\text{UER}_7 = 5.5098 - 0.0294\ (7) = 5.3040$ or approximately 5.3 percent.

B The DW statistic is designed to detect positive serial correlation of the errors of a regression equation. Under the null hypothesis of no positive serial correlation, the DW statistics is 2.0. Positive serial correlation will lead to a DW statistic that is less than 2.0. From the table in Problem 1, we see that the DW statistic is 0.9099. To see whether this result is significantly less than 2.0, refer to the Durbin–Watson table in Appendix E at the end of this volume, in the column marked $k = 1$ (one independent variable) and the row corresponding to 60 observations. We see that $d_l = 1.55$. Because our DW statistic is clearly less than d_l, we reject the null hypothesis of no serial correlation at the 0.05 significance level.

The presence of serial correlation in the error term violates one of the regression assumptions. The standard errors of the estimated coefficients will be biased downward, so we cannot conduct hypothesis testing on the coefficients.

2 The difference between UER and its forecast value, PRED, is the forecast error. In an appropriately specified regression model, the forecast errors are randomly distributed around the regression line and have a constant variance. We can see that the errors from this model specification are persistent. The errors tend first to be above the regression line and then, starting in 1997, they tend to be below the regression line until 2000 when they again are persistently above the regression line. This persistence suggests that the errors are positively serially correlated. Therefore, we conclude that the model is not appropriate for making estimates.

3 A log-linear model captures growth at a constant rate. The log-linear model $\ln(\text{Sales}_t) = b_0 + b_1 t + \varepsilon_t$ would be the simplest model consistent with a constant growth rate for monthly sales. Note that we would need to confirm that the regression assumptions are satisfied before accepting the model as valid.

4 A The plot of the series ΔUER_t seems to fluctuate around a constant mean; its volatility appears to be constant throughout the period. Our initial judgment is that the differenced series is covariance stationary.

B The change in the unemployment rate seems covariance stationary, so we should first estimate an AR(1) model and test to see whether the residuals from this model have significant serial correlation. If the residuals do not display significant serial correlation, we should use the AR(1) model. If the residuals do display significant serial correlation, we should try an AR(2) model and test for serial correlation of the residuals of the AR(2) model. We should continue this procedure until the errors from the final AR(p) model are serially uncorrelated.

5 The DW statistic cannot be appropriately used for a regression that has a lagged value of the dependent variable as one of the explanatory variables. To test for serial correlation, we need to examine the autocorrelations.

6 When a covariance-stationary series is at its mean-reverting level, the series will tend not to change until it receives a shock (ε_t). So, if the series ΔUER_t is at the mean-reverting level, $\Delta\text{UER}_t = \Delta\text{UER}_{t-1}$. This implies that $\Delta\text{UER}_t = -0.0405$

− 0.4674ΔUER$_t$, so that (1 + 0.4674) ΔUER$_t$ = −0.0405 and ΔUER$_t$ = −0.0405/(1 + 0.4674) = −0.0276. The mean-reverting level is −0.0276. In an AR(1) model, the general expression for the mean-reverting level is $b_0/(1 - b_1)$.

7 **A** The predicted change in the unemployment rate for next period is −5.45 percent, found by substituting 0.0300 into the forecasting model: −0.0405 − 0.4674(0.03) = −0.0545.

B If we substitute our one-period-ahead forecast of −0.0545 into the model (using the chain rule of forecasting), we get a two-period ahead forecast of −0.0150 or −1.5 percent.

C The answer to Part B is quite close to the mean-reverting level of − 0.0276. A stationary time series may need many periods to return to its equilibrium, mean-reverting level.

8 The forecast of sales is \$4,391 million for the first quarter of 2002 and \$4,738 million for the second quarter of 2002, as the following table shows.

Date	Sales (\$ Millions)	Log of Sales	Actual Values of Changes in the Log of Sales Δln (Sales$_t$)	Forecast Values of Changes in the Log of Sales Δln (Sales$_t$)
1Q:2001	6,519	8.7825	0.1308	
2Q:2001	6,748	8.8170	0.0345	
3Q:2001	4,728	8.4613	−0.3557	
4Q:2001	4,298	8.3659	−0.0954	
1Q:2002	4,391	8.3872		0.0213
2Q:2002	4,738	8.4633		0.0761

We find the forecasted change in the log of sales for the first quarter of 2002 by inputting the value for the change in the log of sales from the previous quarter into the equation Δln (Sales$_t$) = 0.0661 + 0.4698Δln (Sales$_{t-1}$). Specifically, Δln (Sales$_t$) = 0.0661 + 0.4698(−0.0954) = 0.0213, which means that we forecast the log of sales in the first quarter of 2002 to be 8.3659 + 0.0213 = 8.3872.

Next, we forecast the change in the log of sales for the second quarter of 2002 as Δln (Sales$_t$) = 0.0661 + 0.4698(0.0213) = 0.0761. Note that we have to use our first-quarter 2002 estimated value of the change in the log of sales as our input for Δln (Sales$_{t-1}$) because we are forecasting past the period for which we have actual data.

With a forecasted change of 0.0761, we forecast the log of sales in the second quarter of 2002 to be 8.3872 + 0.0761 = 8.4633.

We have forecasted the log of sales in the first and second quarters of 2002 to be 8.3872 and 8.4633, respectively. Finally, we take the antilog of our estimates of the log of sales in the first and second quarters of 2002 to get our estimates of the level of sales: $e^{8.3872}$ = 4,391 and $e^{8.4633}$ = 4,738, respectively, for sales of \$4,391 million and \$4,738 million.

9 **A** The RMSE of the out-of-sample forecast errors is approximately 27 percent. Out-of-sample error refers to the difference between the realized value and the forecasted value of Δln (Sales$_t$) for dates beyond the estimation period. In this case, the out-of-sample period is 1Q:2001 to 4Q:2001. These are the four quarters for which we have data that we did not use to obtain the estimated model Δln (Sales$_t$) = 0.0661 + 0.4698Δln (Sales$_{t-1}$).

The steps to calculate RMSE are as follows:

i. Take the difference between the actual and the forecast value. This is the error.

ii. Square the error.

iii. Sum the squared errors.

iv. Divide by the number of forecasts.

v. Take the square root of the average.

We show the calculations for RMSE in the table below.

Actual Values of Changes in the Log of Sales $\Delta\ln(\text{Sales}_t)$	Forecast Values of Changes in the Log of Sales $\Delta\ln(\text{Sales}_t)$	Error (Column 1 – Column 2)	Squared Error (Column 3 Squared)
0.1308	0.1357	–0.0049	0.0000
0.0345	0.1299	–0.0954	0.0091
–0.3557	0.1271	–0.4828	0.2331
–0.0954	0.1259	–0.2213	0.0490
		Sum	0.2912
		Mean	0.0728
		RMSE	0.2698

B The lower the RMSE, the more accurate the forecasts of a model in forecasting. Therefore, the model with the RMSE of 20 percent has greater accuracy in forecasting than the model in Part A, which has an RMSE of 27 percent.

10 A Predictions too far ahead can be nonsensical. For example, the AR(1) model we have been examining, $\Delta\text{UER}_t = -0.0405 - 0.4674\Delta\text{UER}_{t-1}$, taken at face value, predicts declining civilian unemployment into the indefinite future. Because the civilian unemployment rate will probably not go below 3 percent frictional unemployment and cannot go below 0 percent unemployment, this model's long-range forecasts are implausible. The model is designed for short-term forecasting, as are many time-series models.

B Using more years of data for estimation may lead to nonstationarity even in the series of first differences in the civilian unemployment rate. As we go further back in time, we increase the risk that the underlying civilian unemployment rate series has more than one regime (or true model). If the series has more than one regime, fitting one model to the entire period would not be correct. Note that when we have good reason to believe that a time series is stationary, a longer series of data is generally desirable.

11 A The graph of $\ln(\text{Sales}_t)$ appears to trend upward over time. A series that trends upward or downward over time often has a unit root and is thus not covariance stationary. Therefore, using an AR(1) regression on the undifferenced series is probably not correct. In practice, we need to examine regression statistics to confirm visual impressions such as this.

B The most common way to transform a time series with a unit root into a covariance-stationary time series is to difference the data—that is, to create a new series $\Delta\ln(\text{Sales}_t) = \ln(\text{Sales}_t) - \ln(\text{Sales}_{t-1})$.

12 The plot of the series $\Delta\ln(\text{Sales}_t)$ appears to fluctuate around a constant mean; its volatility seems constant throughout the period. Differencing the data appears to have made the time series covariance stationary.

13 A In a correctly specified regression, the residuals must be serially uncorrelated. We have 108 observations, so the standard error of the autocorrelation is $1/\sqrt{T}$, or in this case $1/\sqrt{108} = 0.0962$. The t-statistic for each lag is significant at the 0.01 level. We would have to modify the model specification before continuing with the analysis.

B Because the residuals from the AR(1) specification display significant serial correlation, we should estimate an AR(2) model and test for serial correlation of the residuals of the AR(2) model. If the residuals from the AR(2) model are serially uncorrelated, we should then test for seasonality and ARCH behavior. If any serial correlation remains in the residuals, we should estimate an AR(3) process and test the residuals from that specification for serial correlation. We should continue this procedure until the errors from the final AR(p) model are serially uncorrelated. When serial correlation is eliminated, we should test for seasonality and ARCH behavior.

14 A The series has a steady upward trend of growth, suggesting an exponential growth rate. This finding suggests transforming the series by taking the natural log and differencing the data.

B First, we should determine whether the residuals from the AR(1) specification are serially uncorrelated. If the residuals are serially correlated, then we should try an AR(2) specification and then test the residuals from the AR(2) model for serial correlation. We should continue in this fashion until the residuals are serially uncorrelated, then look for seasonality in the residuals. If seasonality is present, we should add a seasonal lag. If no seasonality is present, we should test for ARCH. If ARCH is not present, we can conclude that the model is correctly specified.

C If the model $\Delta\ln(\text{Sales}_t) = b_0 + b_1[\Delta\ln(\text{Sales}_{t-1})] + \varepsilon_t$ is correctly specified, then the series $\Delta\ln(\text{Sales}_t)$ is covariance stationary. So, this series tends to its mean-reverting level, which is $b_0/(1 - b_1)$ or $0.0661/(1 - 0.4698) = 0.1247$.

15 The quarterly sales of Avon show an upward trend and a clear seasonal pattern, as indicated by the repeated regular cycle.

16 A A second explanatory variable, the change in the gross profit margin lagged four quarters, ΔGPM_{t-4}, was added.

B The model was augmented to account for seasonality in the time series (with quarterly data, significant autocorrelation at the fourth lag indicates seasonality). The standard error of the autocorrelation coefficient equals 1 divided by the square root of the number of observations: $1/\sqrt{40}$ or 0.1581. The autocorrelation at the fourth lag (0.8496) is significant: $t = 0.8496/0.1581 = 5.37$. This indicates seasonality, and accordingly we added ΔGPM_{t-4}. Note that in the augmented regression, the coefficient on ΔGPM_{t-4} is highly significant. (Although the autocorrelation at second lag is also significant, the fourth lag is more important because of the rationale of seasonality. Once the fourth lag is introduced as an independent variable, we might expect that the second lag in the residuals would not be significant.)

17 A In order to determine whether this model is correctly specified, we need to test for serial correlation among the residuals. We want to test whether we can reject the null hypothesis that the value of each autocorrelation is 0 against the alternative hypothesis that each is not equal to 0. At the 0.05 significance level, with 68 observations and three parameters, this model has 65 degrees of freedom. The critical value of the t-statistic needed to reject the null hypothesis is thus about 2.0. The absolute value of the t-statistic for each autocorrelation is below 0.60 (less than 2.0), so we cannot reject the null hypothesis that each autocorrelation is not significantly different from 0. We have determined that the model is correctly specified.

B If sales grew by 1 percent last quarter and by 2 percent four quarters ago, then the model predicts that sales growth this quarter will be 0.0121 – 0.0839 ln(1.01) + 0.6292 ln(1.02) = $e^{0.02372}$ – 1 = 2.40%.

18 We should estimate the regression $\Delta\text{UER}_t = b_0 + b_1\Delta\text{UER}_{t-1} + \varepsilon_t$ and save the residuals from the regression. Then we should create a new variable, $\hat{\varepsilon}_t^2$, by squaring the residuals. Finally, we should estimate $\hat{\varepsilon}_t^2 = a_0 + a_1\hat{\varepsilon}_{t-1}^2 + u_t$ and test to see whether a_1 is statistically different from 0.

19 To determine whether we can use linear regression to model more than one time series, we should first determine whether any of the time series has a unit root. If none of the time series has a unit root, then we can safely use linear regression to test the relations between the two time series. Note that if one of the two variables has a unit root, then our analysis would not provide valid results; if both of the variables have unit roots, then we would need to evaluate whether the variables are cointegrated.

20 C is correct. The predicted value for period t from a linear trend is calculated as $\hat{y}_t = \hat{b}_0 + \hat{b}_1(t)$.

October 2015 is the second month out of sample, or t = 183. So, the predicted value for October 2015 is calculated as

$$\hat{y}_t = 28.3278 + 0.4086(183) = \$103.10.$$

Therefore, the predicted WTI oil price for October 2015 based on the linear trend model is \$103.10.

21 C is correct. The predicted value for period t from a log-linear trend is calculated as $\ln\hat{y}_t = \hat{b}_0 + \hat{b}_1(t)$.

September 2015 is the first month out of sample, or t = 182. So, the predicted value for September 2015 is calculated as follows:

$$\ln\hat{y}_t = 3.3929 + 0.0075(182)$$

$$\ln\hat{y}_t = 4.7579$$

$$\hat{y}_t = e^{4.7579} = \$116.50$$

Therefore, the predicted WTI oil price for September 2015, based on the log-linear trend model, is \$116.50.

22 B is correct. The Durbin–Watson statistic for the linear trend model is 0.10 and, for the log-linear trend model, 0.08. Both of these values are below the critical value of 1.75. Therefore, we can reject the hypothesis of no positive serial correlation in the regression errors in both the linear trend model and the log-linear trend model.

23 B is correct. There are three requirements for a time series to be covariance stationary. First, the expected value of the time series must be constant and finite in all periods. Second, the variance of the time series must be constant and finite in all periods. Third, the covariance of the time series with itself for a fixed number of periods in the past or future must be constant and finite in all periods. Martinez concludes that the mean and variance of the time series of WTI oil prices are not constant over time. Therefore, the time series is not covariance stationary.

24 B is correct. The last two observations in the WTI time series are July and August 2015, when the WTI oil price was \$51.16 and \$42.86, respectively. Therefore, September 2015 represents a one-period-ahead forecast. The one-period-ahead forecast from an AR(2) model is calculated as

$$\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1 x_t + \hat{b}_2 x_{t+1}$$

So, the one-period-ahead (September 2015) forecast is calculated as

$$\hat{x}_{t+1} = 2.0017 + 1.3946(\$42.86) - 0.4249(\$51.16) = \$40.04.$$

Therefore, the September 2015 forecast based on the AR(2) model is \$40.04.

25 C is correct. The standard error of the autocorrelations is calculated as $\frac{1}{\sqrt{T}}$, where T represents the number of observations used in the regression.

Therefore, the standard error for each of the autocorrelations is $\frac{1}{\sqrt{180}} = 0.0745$.

Martinez can conclude that the residuals are serially correlated and are significantly different from zero because two of the four autocorrelations in Exhibit 2 have a t-statistic in absolute value that is greater than the critical value of 1.97.

Choices A and B are incorrect because two of the four autocorrelations have a t-statistic in absolute value that is greater than the critical value of the t-statistic of 1.97.

26 C is correct. The mean-reverting level from the AR(1) model is calculated as

$$\hat{x}_t = \frac{b_0}{1 - b_1} = \frac{1.5948}{1 - 0.9767} = \$68.45$$

Therefore, the mean-reverting WTI oil price from the AR(1) model is \$68.45. The forecasted oil price in September 2015 will likely be greater than \$42.86 because the model predicts that the price will rise in the next period from the August 2015 price of \$42.86.

27 C is correct. A random walk can be described by the equation $x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$, where $b_0 = 0$ and $b_1 = 1$. So $b_0 = 0$ is a characteristic of a random walk time series. A covariance-stationary series must satisfy the following three requirements:

1. The expected value of the time series must be constant and finite in all periods.
2. The variance of the time series must be constant and finite in all periods.
3. The covariance of the time series with itself for a fixed number of periods in the past or future must be constant and finite in all periods.

$b_0 = 0$ does not violate any of these three requirements and is thus consistent with the properties of a covariance-stationary time series.

28 B is correct. The critical t-statistic at a 5% confidence level is 1.98. As a result, neither the intercept nor the coefficient on the first lag of the first-differenced exchange rate in Regression 2 differs significantly from zero. Also, the residual autocorrelations do not differ significantly from zero. As a result, Regression 2 can be reduced to $y_t = \varepsilon_t$ with a mean-reverting level of $b_0/(1 - b_1) = 0/1 = 0$. Therefore, the variance of y_t in each period is $\text{Var}(\varepsilon_t) = \sigma^2$. The fact that the residuals are not autocorrelated is consistent with the covariance of the times series, with itself being constant and finite at different lags. Because the variance and the mean of y_t are constant and finite in each period, we can also conclude that y_t is covariance stationary.

29 A is correct. If the exchange rate series is a random walk, then the first-differenced series will yield $b_0 = 0$ and $b_1 = 0$, and the error terms will not be serially correlated. The data in Exhibit 1 show that this is the case: Neither the intercept nor the coefficient on the first lag of the first-differenced exchange rate in Regression 2 differs significantly from zero because the t-statistics of both coefficients are less than the critical t-statistic of 1.98. Also, the residual autocorrelations do not differ significantly from zero because the t-statistics of all autocorrelations are less than the critical t-statistic of 1.98. Therefore, because all random walks have unit roots, the exchange rate time series used to run Regression 1 has a unit root.

30 C is correct. To conduct the Dickey–Fuller test, one must subtract the independent variable, x_{t-1}, from both sides of the original AR(1) model. This results in a change of the dependent variable (from x_t to $x_t - x_{t-1}$) and a change in the regression's slope coefficient (from b_1 to $b_1 - 1$) but not a change in the independent variable.

31 C is correct. The regression output in Exhibit 2 suggests there is serial correlation in the residual errors. The fourth autocorrelation of the residual has a value of 0.6994 and a t-statistic of 4.3111, which is greater than the t-statistic critical value of 2.02. Therefore, the null hypothesis that the fourth autocorrelation is equal to zero can be rejected. This indicates strong and significant seasonal autocorrelation, which means the Regression 3 equation is misspecified.

32 C is correct. The quarterly sales for March 2016 is calculated as follows:

$$\begin{aligned}
\ln \text{Sales}_t - \ln \text{Sales}_{t-1} &= b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + b_2(\ln \text{Sales}_{t-4} \\
&\quad - \ln \text{Sales}_{t-5}).\\
\ln \text{Sales}_t - \ln 3.868 &= 0.0092 - 0.1279(\ln 3.868 - \ln 3.780) + 0.7239(\ln \\
&\quad 3.836 - \ln 3.418).\\
\ln \text{Sales}_t - 1.35274 &= 0.0092 - 0.1279(1.35274 - 1.32972) + \\
&\quad 0.7239(1.34443 - 1.22906).\\
\ln \text{Sales}_t &= 1.35274 + 0.0092 - 0.1279(0.02301) + \\
&\quad 0.7239(0.11538).\\
\ln \text{Sales}_t &= 1.44251.\\
\text{Sales}_t &= e^{1.44251} = 4.231.
\end{aligned}$$

33 B is correct. Exhibit 5 shows that the time series of the stock prices of Company #1 exhibits heteroskedasticity, as evidenced by the fact that the time series is ARCH(1). If a time series is ARCH(1), then the variance of the error in one period depends on the variance of the error in previous periods. Therefore, the variance of the errors in period $t + 1$ can be predicted in period t using the formula

$$\hat{\sigma}^2_{t+1} = \hat{a}_0 + \hat{a}_1\hat{\varepsilon}^2_t$$

34 B is correct. When two time series have a unit root but are co-integrated, the error term in the linear regression of one time series on the other will be covariance stationary. Exhibit 5 shows that the series of stock prices of Company #2 and the oil prices both contain a unit root, and the two time series are co-integrated. As a result, the regression coefficients and standard errors are consistent and can be used for hypothesis tests. Although the co-integrated regression estimates the long-term relation between the two series, it may not be the best model of the short-term relationship.

35 C is correct. As a result of the exponential trend in the time series of stock prices for Company #3, Busse would want to take the natural log of the series and then first-difference it. Because the time series also has serial correlation in the residuals from the trend model, Busse should use a more complex model, such as an autoregressive (AR) model.

QUANTITATIVE METHODS
STUDY SESSION

3

Quantitative Methods (2)

This study session provides coverage on techniques that underlie how financial technology (fintech) is affecting areas within the investment industry, such as investment analysis, automated advice, and risk management. The first reading introduces techniques in machine learning (ML) that involve clustering, simplifying, classifying, and predicting relationships in the large datasets that are often found in finance. The next reading examines how data projects involving large datasets are structured with an application to sentiment analysis in investment analysis using machine learning techniques for natural language processing (NLP). The session concludes with coverage of probability-based techniques for assessing risk, with a focus on simulation models.

READING ASSIGNMENTS

Reading 7	Machine Learning by Kathleen DeRose, CFA, and Christophe Le Lannou
Reading 8	Big Data Projects by Sree Mallikarjun, PhD, and Ahmed Abbasi, PhD
Reading 9	Excerpt from "Probabilistic Approaches: Scenario Analysis, Decision Trees, and Simulations" by Aswath Damodaran

READING

7

Machine Learning

by Kathleen DeRose, CFA, and Christophe Le Lannou

Kathleen DeRose, CFA, is at New York University, Stern School of Business (USA). Christophe Le Lannou is at dataLearning (United Kingdom).

LEARNING OUTCOMES

Mastery	The candidate should be able to:
☐	**a.** distinguish between supervised machine learning, unsupervised machine learning, and deep learning;
☐	**b.** describe overfitting and identify methods of addressing it;
☐	**c.** describe supervised machine learning algorithms—including penalized regression, support vector machine, k-nearest neighbor, classification and regression tree, ensemble learning, and random forest—and determine the problems for which they are best suited;
☐	**d.** describe unsupervised machine learning algorithms—including principal components analysis, k-means clustering, and hierarchical clustering—and determine the problems for which they are best suited;
☐	**e.** describe neural networks, deep learning nets, and reinforcement learning.

INTRODUCTION

1

Investment firms are increasingly using technology at every step of the investment management value chain—from improving their understanding of clients, to uncovering new sources of alpha, to executing trades more efficiently. Machine learning techniques, a central part of that technology, are the subject of this reading. These techniques first appeared in finance in the 1990s and have since flourished with the explosion of data and cheap computing power.

This reading provides a high-level view of machine learning (ML). It covers a selection of key ML algorithms and their investment applications. Investment practitioners should be equipped with a basic understanding of the types of investment problems that machine learning can address, an idea of how the algorithms work, and the vocabulary to interact with machine learning and data science experts. While investment practitioners need not master the details and mathematics of machine learning, as domain experts in investments they can play an important role by being able to source appropriate model inputs, interpret model outputs, and translate outputs into appropriate investment actions.

Section 2 gives an overview of machine learning in investment management. Section 3 defines machine learning and the types of problems that can be addressed by supervised and unsupervised learning. Section 4 describes evaluating machine learning algorithm performance. Key supervised machine learning algorithms are covered in Section 5, while Section 6 describes key unsupervised machine learning algorithms. Neural networks, deep learning nets, and reinforcement learning are covered in Section 7. The reading concludes with a summary.

2 MACHINE LEARNING AND INVESTMENT MANAGEMENT

The growing volume and exploding diversity of data, as well as the increasing economic value of insights extracted from these data, have inspired a resurgence in data science. This newly evolving field combines mathematics, computer science, and business analytics. It also strikes out in a new direction that relies on learning—from basic learning functions that map relationships between variables, to advanced neural networks that mimic physical processes that absorb, order, and adapt to information.

Machine learning has theoretical and practical implications for investment management. For example, machine learning could potentially reshape received wisdom about asset risk premiums and reconfigure investment management business processes. Large data sets and learning models are already affecting investment management practices—from client profiling, to asset allocation, to stock selection, to portfolio construction and risk management, to trading.

Machine learning applications are at each step of the asset and wealth management value chain. Chatbots answer basic retirement savings questions, learning from their interactions with investors. Machine learning methods can be used to generate alpha signals used in security selection by creating a non-linear forecast for a single time series, by deriving a forecast from a suite of predefined factors, or even by choosing input signals from existing or newly found data. For example, researchers using textual analysis have found that year-over-year changes in annual (10K) and quarterly (10Q) filings, particularly negative changes in the management discussion and risk sections, can strongly predict equity returns.

Machine learning methods can help calculate target portfolio weights that incorporate client restrictions and then dynamically weight them to maximize a Sharpe ratio or better estimate the variance–covariance matrix by using such techniques as principal components analysis, which reduces the number of variables needed to explain the variation in the data. Research suggests that machine learning solutions outperform mean–variance optimization in portfolio construction. Machine learning techniques are already creating better order flow management tools with non-linear trading algorithms that reduce the costs of implementing portfolio decisions. These developments have caused an evolution in automated tools, automated processes, and algorithmically automated businesses, like robo-advising.

3 WHAT IS MACHINE LEARNING?

We now discuss some fundamental concepts of machine learning, including a definition of machine learning and an overview of key types of machine learning, such as supervised and unsupervised ML.

3.1 Defining Machine Learning

Statistical approaches and machine learning techniques both analyze observations to reveal some underlying process; however, they diverge in their assumptions, terminology, and techniques. Statistical approaches rely on foundational assumptions and explicit models of structure, such as observed samples that are assumed to be drawn from a specified underlying probability distribution. These *a priori* restrictive assumptions can fail in reality.

In contrast, machine learning seeks to extract knowledge from large amounts of data with no such restrictions. The goal of machine learning algorithms is to automate decision-making processes by generalizing (i.e., "learning") from known examples to determine an underlying structure in the data. The emphasis is on the ability of the algorithm to generate structure or predictions from data without any human help. An elementary way to think of ML algorithms is to "find the pattern, apply the pattern."

Machine learning techniques are better able than statistical approaches to handle problems with many variables (high dimensionality) or with a high degree of non-linearity. ML algorithms are particularly good at detecting change, even in highly non-linear systems, because they can detect the preconditions of a model's break or anticipate the probability of a regime switch.

Machine learning is broadly divided into three distinct classes of techniques: supervised learning, unsupervised learning, and deep learning.

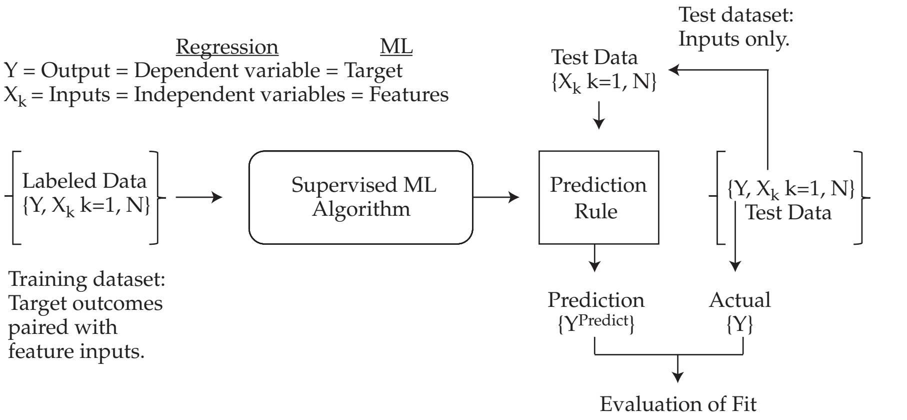
3.2 Supervised Learning

Supervised learning involves ML algorithms that infer patterns between a set of inputs (the *X*'s) and the desired output (*Y*). The inferred pattern is then used to map a given input set into a predicted output. Supervised learning requires a **labeled data set**, one that contains matched sets of observed inputs and the associated output. Applying the ML algorithm to this data set to infer the pattern between the inputs and output is called "training" the algorithm. Once the algorithm has been trained, the inferred pattern can be used to predict output values based on new inputs (i.e., ones not in the training data set).

Multiple regression is an example of supervised learning. A regression model takes matched data (*X*'s, *Y*) and uses it to estimate parameters that characterize the relationship between *Y* and the *X*'s. The estimated parameters can then be used to predict *Y* on a new, different set of *X*'s. The difference between the predicted and actual *Y* is used to evaluate how well the regression model predicts out-of-sample (i.e., using new data).

The terminology used with ML algorithms differs from that used in regression. Exhibit 1 provides a visual of the supervised learning model training process and a translation between regression and ML terminologies.

Exhibit 1 Overview of Supervised Learning



In supervised machine learning, the dependent variable (Y) is the **target** and the independent variables (X's) are known as **features**. The labeled data (training data set) is used to train the supervised ML algorithm to infer a pattern-based prediction rule. The fit of the ML model is evaluated using labeled test data in which the predicted targets ($Y^{Predict}$) are compared to the actual targets (Y^{Actual}).

An example of supervised learning is the case in which ML algorithms are used to predict whether credit card transactions are fraudulent or legitimate. In the credit card example, the target is a binary variable with a value of 1 for "fraudulent" or 0 for "non-fraudulent." The features are the transaction characteristics. The chosen ML algorithm uses these data to train a model to predict fraud more accurately in new transactions. The ML program "learns from experience" if the percentage of correctly predicted credit card transactions increases as the input from a growing credit card database increases. One possible ML algorithm to use here would be to fit a logistic regression model to the data to provide an estimate of the probability a transaction is fraudulent.

Supervised learning can be divided into two categories of problems, regression problems and classification problems, with the distinction between them being determined by the nature of the target (Y) variable. If the target variable is continuous, then the task is one of regression (even if the ML technique used is not "regression," note this nuance of ML terminology). If the target variable is categorical or ordinal (i.e., a ranked category), then it is a classification problem. Regression and classification use different ML techniques.

Regression focuses on making predictions of continuous target variables. Most readers are already familiar with multiple linear regression (e.g., ordinary least squares) models, but other supervised learning techniques exist that include non-linear models. These non-linear models are useful for problems involving large data sets with large numbers of features, many of which may be correlated. Some examples of problems belonging to the regression category are using historical stock market returns to forecast stock price performance or using historical corporate financial ratios to forecast the probability of bond default.

Classification focuses on sorting observations into distinct categories. In a regression problem, when the dependent variable (target) is categorical, the model relating the outcome to the independent variables (features) is called a "classifier." Many classification models are binary classifiers, as in the case of fraud detection for credit card transactions. Multi-category classification is not uncommon, as in the case of classifying firms into multiple credit rating categories. In assigning ratings, the outcome variable is ordinal, meaning the categories have a distinct order or ranking (e.g., from low to high creditworthiness). Ordinal variables are intermediate between categorical variables and continuous variables on a scale of measurement.

3.3 Unsupervised Learning

Unsupervised learning is machine learning that does not make use of labeled data. More formally, in unsupervised learning, we have inputs (*X*'s) that are used for analysis without any target (*Y*) being supplied. In unsupervised learning, because the ML algorithm is not given labeled training data, the algorithm seeks to discover structure within the data themselves. As such, unsupervised learning is useful for exploring new data sets as it can provide human experts with insights into a data set too big or too complex to visualize.

Two important types of problems that are well suited to unsupervised machine learning are reducing the dimension of data and sorting data into clusters, known as dimension reduction and clustering, respectively.

Dimension reduction focuses on reducing the number of features while retaining variation across observations to preserve the information contained in that variation. Dimension reduction may have several purposes. It may be applied to data with a large number of features to produce a lower dimensional representation (i.e., with fewer features) that can fit, for example, on a computer screen. Dimension reduction is also used in many quantitative investment and risk management applications where it is critical to identify the major factors underlying asset price movements.

Clustering focuses on sorting observations into groups (clusters) such that observations in the same cluster are more similar to each other than they are to observations in other clusters. Groups are formed based on a set of criteria that may or may not be prespecified (such as the number of groups). Clustering has been used by asset managers to sort companies into empirically determined groupings (e.g., based on their financial statement data) rather than conventional groupings (e.g., based on sectors or countries).

3.4 Deep Learning and Reinforcement Learning

More broadly in the field of artificial intelligence, additional categories of machine learning algorithms are distinguished. In **deep learning**, sophisticated algorithms address highly complex tasks, such as image classification, face recognition, speech recognition, and natural language processing. In **reinforcement learning**, a computer learns from interacting with itself (or data generated by the same algorithm). Deep learning and reinforcement learning are themselves based on **neural networks** (NNs, also called *artificial neural networks*, or ANNs). NNs include highly flexible ML algorithms that have been successfully applied to a variety of tasks characterized by non-linearities and interactions among features. Besides being commonly used for classification and regression, neural networks are also the foundation for deep learning and reinforcement learning, which can be either supervised or unsupervised.

3.5 Summary of ML Algorithms and How to Choose Among Them

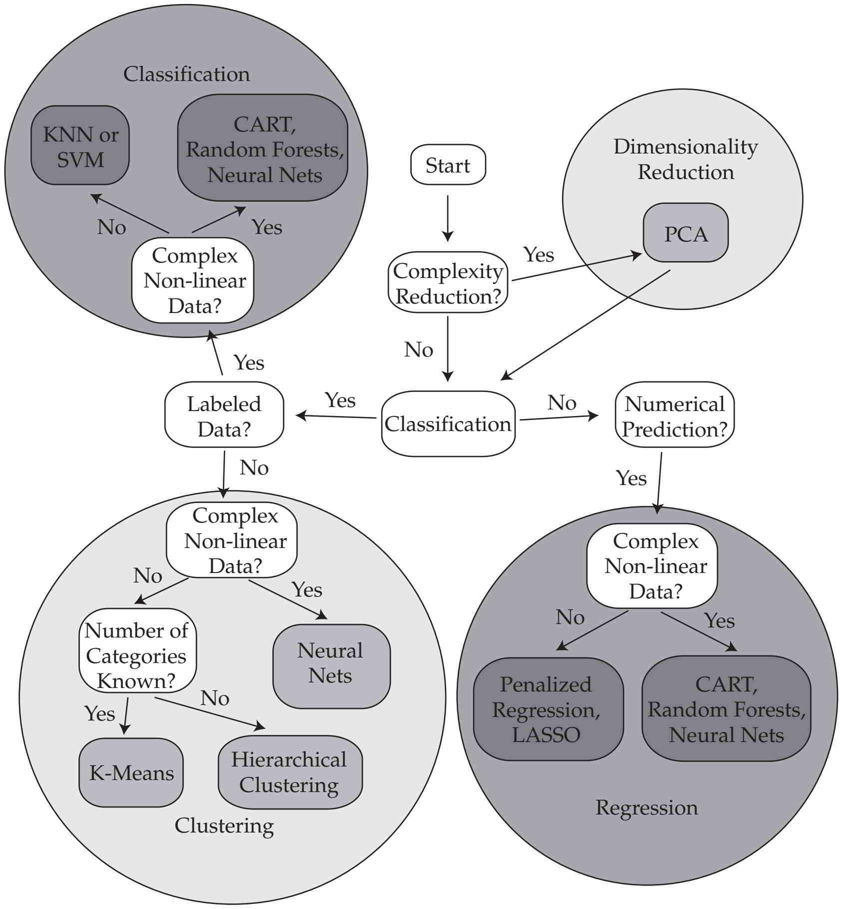
Exhibit 2 is a guide to the various machine learning algorithms organized by algorithm type (supervised or unsupervised) and by type of variables (continuous, categorical, or both). We will not cover linear or logistic regression since they are covered elsewhere in quantitative methods. The extensions of linear regression, such as penalized regression and least absolute shrinkage and selection operator (LASSO), as well as the other ML algorithms shown in Exhibit 2, will be covered in this reading.

Exhibit 2 Guide to ML Algorithms

	ML Algorithm Type	
Variables	**Supervised (Target Variable)**	**Unsupervised (No Target Variable)**
Continuous	**Regression** • Linear; Penalized Regression/LASSO • Logistic • Classification and Regression Tree (CART) • Random Forest	**Dimensionality Reduction** • Principal Components Analysis (PCA) **Clustering** • K-Means • Hierarchical
Categorical	**Classification** • Logit • Support Vector Machine (SVM) • K-Nearest Neighbor (KNN) • Classification and Regression Tree (CART)	**Dimensionality Reduction** • Principal Components Analysis (PCA) **Clustering** • K-Means • Hierarchical
Continuous or Categorical	Neural Networks Deep Learning Reinforcement Learning	Neural Networks Deep Learning Reinforcement Learning

Exhibit 3 presents a stylized decision flowchart for choosing among the machine learning algorithms shown in Exhibit 2. The dark-shaded ovals contain the supervised ML algorithms; the light-shaded ovals contain the unsupervised ML algorithms; and the key questions to consider are shown in the unshaded boxes.

Exhibit 3 Stylized Decision Flowchart for Choosing ML Algorithms



First, start by asking, are the data complex, having many features that are highly correlated? If yes, then dimensionality reduction using principal components analysis (PCA) is appropriate.

Next, is the problem one of classification or numerical prediction? If numerical prediction, then depending on whether or not the data have non-linear characteristics, the choice of ML algorithms is from a set of regression algorithms—either penalized regression/LASSO for linear data or CART, random forest, and neural networks for non-linear data.

If the problem is one of classification, then depending on whether or not the data are labeled, the choice is either from a set of classification algorithms using labeled data or from a set of clustering algorithms using unlabeled data.

If the data are labeled, then depending on whether or not the data have non-linear characteristics, the choice of classification algorithm would be k-nearest neighbor (KNN) and support vector machine (SVM) for linear data or CART, random forest, and neural networks for non-linear data.

Finally, if the data are unlabeled, the choice of clustering algorithm depends on whether or not the data have non-linear characteristics. The choice of clustering algorithm would be neural networks for non-linear data or for linear data, k-means with a known number of categories, and hierarchical clustering with an unknown number of categories.

Following a description of how to evaluate ML algorithm performance, we will define all of the ML algorithms shown in Exhibit 3 and then explain their applications in investment management.

EXAMPLE 1

Machine Learning Overview

1 Which of the following *best* describes machine learning? Machine learning:
   A is a type of computer algorithm.
   B is a set of computer-driven approaches aimed at generating structure or predictions from data without human intervention by finding a pattern and then applying the pattern.
   C is a set of computer-driven approaches adapted to extracting information from linear, labeled data sets.

2 Which of the following statements is *most* accurate? When attempting to discover groupings of data without any target (Y) variable:
   A an unsupervised ML algorithm is used.
   B an ML algorithm that is given labeled training data is used.
   C a supervised ML algorithm is used.

3 Which of the following statements concerning supervised learning *best* distinguishes it from unsupervised learning? Supervised learning involves:
   A training on labeled data to infer a pattern-based prediction rule.
   B training on unlabeled data to infer a pattern-based prediction rule.
   C learning from unlabeled data by discovering underlying structure in the data themselves.

4 Which of the following *best* describes dimension reduction? Dimension reduction:
   A focuses on classifying observations in a data set into known groups using labeled training data.
   B focuses on clustering observations in a data set into unknown groups using unlabeled data.
   C focuses on reducing the number of features in a data set while retaining variation across observations to preserve the information in that variation.

Solution to 1:

B is correct. A is incorrect because machine learning is not best described as a type of computer algorithm. C is incorrect because machine learning is not limited to extracting information from linear, labeled data sets.

Solution to 2:

A is correct. B is incorrect because the term "labeled training data" means the target (Y) is provided. C is incorrect because a supervised ML algorithm is meant to predict a target (Y) variable.

Solution to 3:

A is correct. B is incorrect because supervised learning uses labeled training data. C is incorrect because it describes unsupervised learning.

Solution to 4:

C is correct. A is incorrect because it describes classification, not dimension reduction. B is incorrect because it describes clustering, not dimension reduction.

OVERVIEW OF EVALUATING ML ALGORITHM PERFORMANCE

4

Machine learning algorithms promise several advantages relative to a structured statistical approach in exploring and analyzing the structure of very large data sets. ML algorithms have the ability to uncover complex interactions between feature variables and the target variable, and they can process massive amounts of data quickly. Moreover, many ML algorithms can easily capture non-linear relationships and may be able to recognize and predict structural changes between features and the target. These advantages mainly derive from the non-parametric and non-linear models that allow more flexibility when inferring relationships.

The flexibility of ML algorithms comes with a price, however. ML algorithms can produce overly complex models with results that are difficult to interpret, may be sensitive to particular changes in the data, and may fit the training data too well. An ML algorithm that fits the training data too well will typically not predict well using new data. This problem is known as **overfitting**, and it means that the fitted algorithm does not **generalize** well to new data. A model that generalizes well is a model that retains its explanatory power when predicting out-of-sample (i.e., using new data). An overfit model has incorporated the noise or random fluctuations in the training data into its learned relationship. The problem is that these aspects do not apply to any new data the algorithm receives and so will negatively impact the model's ability to generalize. The evaluation of any ML algorithm thus focuses on its prediction error on new data rather than on its goodness-of-fit on the data in which the algorithm was fitted (i.e., trained).

Generalization is an objective in model building, so the problem of overfitting is a challenge to attaining that objective. These two concepts are the focus of the discussion below.

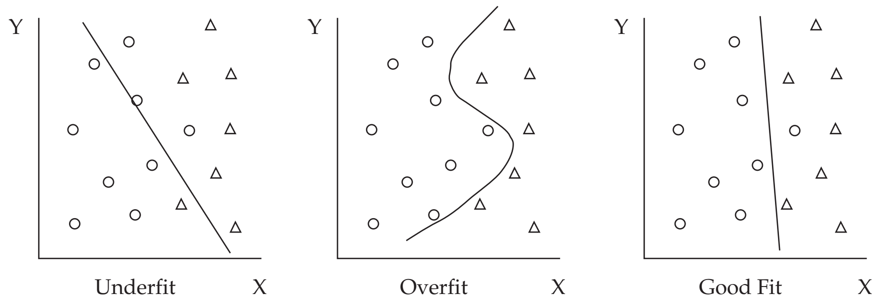
4.1 Generalization and Overfitting

To properly describe generalization and overfitting of an ML model, it is important to note the partitioning of the data set to which the model will be applied. The data set is typically divided into three non-overlapping samples: 1) **training sample** used to train the model; 2) **validation sample** for validating and tuning the model; and 3) **test sample** for testing the model's ability to predict well on new data. The training sample is often referred to as being "in-sample," while the validation and test samples are commonly referred to as being "out-of-sample." We will return shortly to the topic of partitioning the data set.

To be valid and useful, any supervised machine learning model must generalize well beyond the training data. The model should retain its explanatory power when tested out-of-sample. As mentioned, one common reason for failure to generalize is overfitting. Think of overfitting as tailoring a custom suit that fits only one person. Continuing the analogy, underfitting is similar to making a baggy suit that fits no one, whereas robust fitting, the desired result, is similar to fashioning a universal suit that fits all similar people.

The concepts of underfitting, overfitting, and good (or robust) fitting are illustrated in Exhibit 4. Underfitting means the model does not capture the relationships in the data. The left graph shows four errors in this underfit model (three misclassified circles and one misclassified triangle). Overfitting means training a model to such a degree of specificity to the training data that the model begins to incorporate noise coming from quirks or spurious correlations; it mistakes randomness for patterns and relationships. The algorithm may have memorized the data, rather than learned from it, so it has perfect hindsight but no foresight. The main contributors to overfitting are thus high noise levels in the data and too much complexity in the model. The middle graph shows no errors in this overfit model. **Complexity** refers to the number of features, terms, or branches in the model and to whether the model is linear or non-linear (non-linear is more complex). As models become more complex, overfitting risk increases. A good fit/robust model fits the training (in-sample) data well and generalizes well to out-of-sample data, both within acceptable degrees of error. The right graph shows that the good fitting model has only one error, the misclassified circle.

Exhibit 4 Under-, Over-, and Good-Fitting



4.2 Errors and Overfitting

To capture these effects and calibrate degree of fit, data scientists compare error rates in- and out-of-sample as a function of both the data and the algorithm. Total in-sample errors (E_{in}) are generated by the predictions of the fitted relationship relative to actual target outcomes on the training sample. Total out-of-sample errors (E_{out}) are from either the validation or test samples. Low or no in-sample error but large out-of-sample error are indicative of poor generalization. Data scientists decompose the total out-of-sample error into three sources:

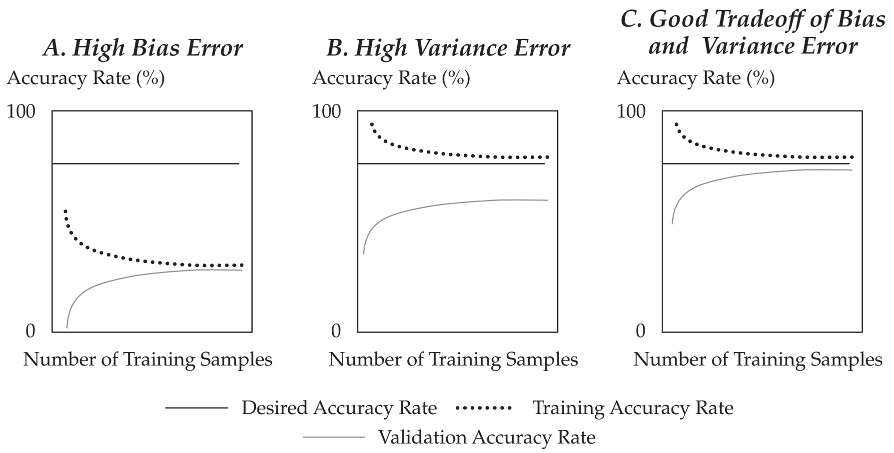
1 **Bias error**, or the degree to which a model fits the training data. Algorithms with erroneous assumptions produce high bias with poor approximation, causing underfitting and high in-sample error.

2 **Variance error**, or how much the model's results change in response to new data from validation and test samples. Unstable models pick up noise and produce high variance, causing overfitting and high out-of-sample error.

3 **Base error** due to randomness in the data.

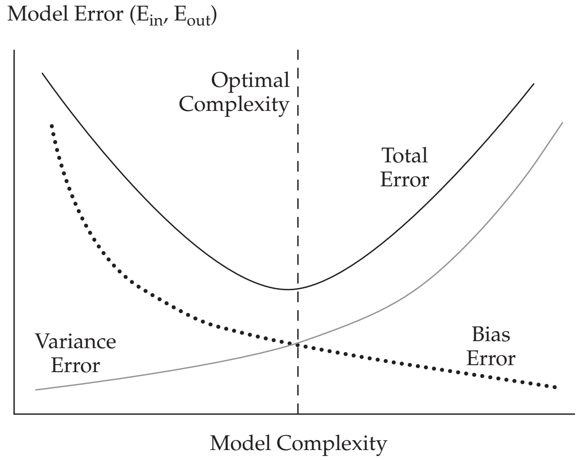
A **learning curve** plots the accuracy rate (= 1 – error rate) in the validation or test samples (i.e., out-of-sample) against the amount of data in the training sample, so it is useful for describing under- and overfitting as a function of bias and variance errors. If the model is robust, out-of-sample accuracy increases as the training sample size increases. This implies that error rates experienced in the validation or test samples (E_{out}) and in the training (E_{in}) sample converge toward each other and toward a desired error rate (or, alternatively, the base error). In an underfitted model with high bias error, shown in the left panel of Exhibit 5, high error rates cause convergence below the desired accuracy rate. Adding more training samples will not improve the model. In an overfitted model with high variance error, shown in the middle panel of Exhibit 5, the validation sample and training sample error rates fail to converge. In building models, data scientists try to simultaneously minimize both bias and variance errors while selecting an algorithm with good predictive or classifying power, as seen in the right panel of Exhibit 5.

Exhibit 5 Learning Curves: Accuracy in Validation and Training Samples vs. Training Sample Size



Out-of-sample error rates are also a function of model complexity. As complexity increases in the training set, error rates (E_{in}) fall and bias error shrinks. As complexity increases in the test set, however, error rates (E_{out}) rise and variance error rises. Typically, linear functions are more susceptible to bias error and underfitting, while non-linear functions are more prone to variance error and overfitting. Therefore, an optimal point of model complexity exists where the bias and variance error curves intersect and in- and out-of-sample error rates are minimized. A **fitting curve**, which shows in- and out-of-sample error rates (E_{in} and E_{out}) on the y-axis plotted against model complexity on the x-axis, is presented in Exhibit 6 and illustrates this trade-off.

Exhibit 6 Fitting Curve Shows Trade-Off Between Bias and Variance Errors and Model Complexity



Finding the optimal point (managing overfitting risk)—the sweet spot just before the total error rate starts to rise (due to increasing variance error)—is a core part of the machine learning process and the key to successful generalization. Data scientists express the trade-off between overfitting and generalization as a trade-off between *cost* (the difference between in- and out-of-sample error rates) and *complexity*. They use the trade-off between cost and complexity to calibrate and visualize under- and overfitting and to optimize their models.

4.3 Preventing Overfitting in Supervised Machine Learning

We have seen that overfitting impairs generalization, but overfitting potential is endemic to the supervised machine learning process due to the presence of noise. So, how do data scientists combat this risk? Two common guiding principles and two methods are used to reduce overfitting: 1) preventing the algorithm from getting too complex during selection and training, which requires estimating an overfitting penalty; and 2) proper data sampling achieved by using **cross-validation**, a technique for estimating out-of-sample error directly by determining the error in validation samples.

The first strategy comes from "Occam's razor," the problem-solving principle that the simplest solution tends to be the correct one. In supervised machine learning, it means limiting the number of features and penalizing algorithms that are too complex or too flexible by constraining them to include only parameters that reduce out-of-sample error.

The second strategy comes from the principle of avoiding sampling bias. Good experimental designs have representative samples and control groups, and both should be drawn from the same theoretical distribution. But sampling bias can creep into machine learning in many ways. The challenge is having a large enough data set to make both training and testing possible on representative samples. An unrepresentative sample or reducing the training sample size too much could obscure its true patterns, thereby increasing bias. In supervised machine learning, the technique for reducing sampling bias is through careful partitioning of the data set into three groups: 1) training sample, the set of labeled training data where the target variable (Y) is known; 2) validation sample, the set of data for comparing various solutions and tuning the selected model, thereby validating the model; and 3) test sample, the set of data held aside for testing to confirm the model's predictive or classifying power. The goal, of course, is to deploy the tested model on fresh data from the same domain.

To mitigate the problem of such **holdout samples** (i.e., data samples not used to train the model) reducing the training set size too much, modelers use special cross-validation techniques. One such technique is **k-fold cross-validation**, in which the data (excluding test sample and fresh data) are shuffled randomly and then are divided into k equal sub-samples, with $k - 1$ samples used as training samples and one sample, the kth, used as a validation sample. Note that k is typically set at 5 or 10. This process is then repeated k times, which helps minimize both bias and variance by insuring that each data point is used in the training set $k - 1$ times and in the validation set once. The average of the k validation errors (mean E_{val}) is then taken as a reasonable estimate of the model's out-of-sample error (E_{out}).

In sum, mitigating overfitting risk by avoiding excessive out-of-sample error is critical to creating a supervised machine learning model that generalizes well to fresh data sets drawn from the same domain. The main techniques used to mitigate overfitting risk in model construction are complexity reduction and cross-validation.

EXAMPLE 2

Evaluating ML Algorithm Performance

Shreya Anand is a portfolio manager based in the Mumbai headquarters office of an investment firm, where she runs a high-dividend yield fund for wealthy clients. Anand has some knowledge of data science from her course in university. She is interested in classifying companies in the NIFTY 200 Index—an index of large- and mid-cap companies listed on the National Stock Exchange of India—into two categories: dividend increase and no dividend increase. She assembles data for training, validating, and testing an ML-based model that consists of 1,000 observations of NIFTY 200 companies, each consisting of 25 features (fundamental and technical) and the labeled target (dividend increase or no dividend increase).

After training her model, Anand discovers that while it is good at correctly classifying using the training sample, it does not perform well using new data. In consulting her colleagues about this issue, Anand hears conflicting explanations about what constitutes good generalization in an ML model:

Statement 1 The model retains its explanatory power when predicting using new data (i.e., out-of-sample).

Statement 2 The model shows low explanatory power after training using in-sample data (i.e., training data).

Statement 3 The model loses its explanatory power when predicting using new data (i.e., out-of-sample).

1 Which statement made to Anand is most accurate?

A Statement 1.

B Statement 2.

C Statement 3.

2 Anand's model is *most likely* being impaired by which of the following:

A Underfitting and bias error

B Overfitting and variance error.

C Overfitting and bias error.

3 By implementing which one of the following actions can Anand address the problem?

A Estimate and incorporate into the model a penalty that decreases in size with the number of included features.

B Use the k-fold cross-validation technique to estimate the model's out-of-sample error, and then adjust the model accordingly.

C Use an unsupervised learning model.

Solution to 1:

A, Statement 1, is correct. B, Statement 2, is incorrect because it describes a poorly fitting model with high bias. C, Statement 3, is incorrect because it describes an overfitted model with poor generalization.

Solution to 2:

B is correct. Anand's model is good at correctly classifying using the training sample, but it does not perform well using new data. The model is overfitted, so it has high variance error.

Solution to 3:

B is correct. A is incorrect because the penalty should increase in size with the number of included features. C is incorrect because Anand is using labeled data for classification, and unsupervised learning models do not use labeled data.

5 SUPERVISED MACHINE LEARNING ALGORITHMS

Supervised machine learning models are trained using labeled data, and depending on the nature of the target (Y) variable, they can be divided into two types: regression for a continuous target variable and classification for a categorical or ordinal target variable. As shown in Exhibits 2 and 3 under regression, we will now cover penalized regression and LASSO. Then as shown under classification, we will introduce support vector machine (SVM), k-nearest neighbor (KNN), and classification and regression tree (CART) algorithms. Note that CART, as its name implies, can be used for both classification and regression problems.

In the following discussion, assume we have a number of observations of a target variable, Y, and n real-valued features, $X_1, \ldots, X_n$, that we may use to establish a relationship (regression or classification) between X (a vector of the X_i) and Y for each observation in our data set.

5.1 Penalized Regression

Penalized regression is a computationally efficient technique used in prediction problems. In practice, penalized regression has been useful for reducing a large number of features to a manageable set and for making good predictions in a variety of large data sets, especially where features are correlated (i.e., when classical linear regression breaks down).

In a large data set context, we may have many features that potentially could be used to explain Y. When a model is fit to training data, the model may so closely reflect the characteristics of the specific training data that the model does not perform well on new data. Features may be included that reflect what is noise or randomness in the training data set that will not be present in new or future data used for making predictions. That is the problem of overfitting, and penalized regression can be described as a technique to avoid overfitting. In prediction, out-of-sample performance is key, so relatively parsimonious models (that is, models in which each variable plays an essential role) tend to work well because they are less subject to overfitting.

Let us suppose that we standardize our data, so the features have a mean of 0 and a variance of 1. Standardization of features will allow us to compare the magnitudes of regression coefficients for the feature variables. In ordinary linear regression (i.e., ordinary least squares, or OLS), the regression coefficients $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_K$ are chosen to *minimize* the sum of the squared residuals (i.e., the sum of the squared difference between the actual values, Y_i, and the predicted values, $\hat{Y}_i$), or:

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Penalized regression includes a constraint such that the regression coefficients are chosen to minimize the sum of squared residuals *plus* a penalty term that increases in size with the number of included features. So, in a penalized regression, a feature must make a sufficient contribution to model fit to offset the penalty from including it. Therefore, only the more important features for explaining Y will remain in the penalized regression model.

In one popular type of penalized regression, **LASSO** (least absolute shrinkage and selection operator), the penalty term has the following form, with $\lambda > 0$:

$$\text{Penalty term} = \lambda \sum_{k=1}^{K} \left| \hat{b}_k \right|.$$

In addition to minimizing the sum of the squared residuals, LASSO also involves minimizing the sum of the absolute values of the regression coefficients (see the following expression). The greater the number of included features (i.e., variables with non-zero coefficients), the larger the penalty term. Therefore, penalized regression ensures that a feature is included only if the sum of squared residuals declines by more than the penalty term increases. All types of penalized regression involve a trade-off of this type. Also, since LASSO eliminates the least important features from the model, it automatically performs feature selection.

$$\sum_{i=1}^{n} \left(Y_i - Y_i\right)^2 + \lambda \sum_{k=1}^{K} \left| \hat{b}_K \right|$$

Lambda (λ) is a **hyperparameter**—a parameter whose value must be set by the researcher before learning begins—of the regression model and will determine the balance between fitting the model versus keeping the model parsimonious. Note that in the case where $\lambda = 0$, then the LASSO penalized regression is equivalent to an OLS regression. When using LASSO or other penalized regression techniques, the penalty term is added only during the model building process (i.e., when fitting the model to the training data). Once the model has been built, the penalty term is no longer needed, and the model is then evaluated by the sum of the squared residuals generated using the test data set.

With today's availability of fast computation algorithms, investment analysts are increasingly using LASSO and other regularization techniques to remove less pertinent features and build parsimonious models. **Regularization** describes methods that reduce statistical variability in high dimensional data estimation problems—in this case, reducing regression coefficient estimates toward zero and thereby avoiding complex models and the risk of overfitting. LASSO has been used, for example, for forecasting default probabilities in industrial sectors where scores of potential features, many collinear, have been reduced to fewer than 10 variables, which is important given the relatively small number (about 100) of observations of default.

Regularization methods can also be applied to non-linear models. A long-term challenge of the asset management industry in applying mean–variance optimization has been the estimation of stable covariance matrixes and asset weights for large portfolios. Asset returns typically exhibit strong multi-collinearity, making the estimation of the covariance matrix highly sensitive to noise and outliers, so the resulting optimized asset weights are highly unstable. Regularization methods have been used to address this problem.

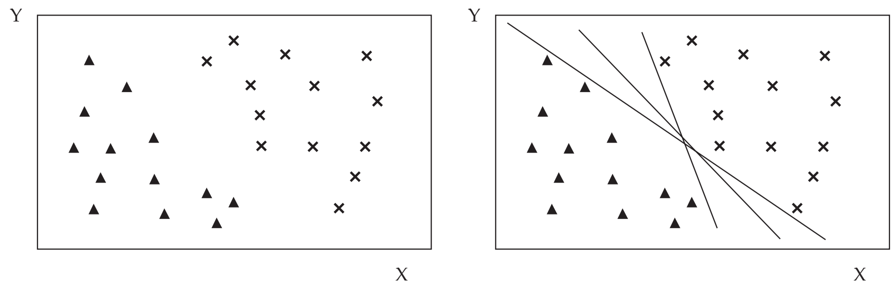
In prediction, only out-of-sample performance (i.e., prediction accuracy) really matters. The relatively parsimonious models produced by applying penalized regression methods, like LASSO, tend to work well because they are less subject to overfitting.

5.2 Support Vector Machine

Support vector machine (SVM) is one of the most popular algorithms in machine learning. It is a powerful supervised algorithm used for classification, regression, and outlier detection. Despite its complicated sounding name, the notion is relatively straightforward and best explained with a few pictures. The left panel in Exhibit 7 presents a simple data set with two features (x and y coordinates) labeled in two groups (triangles and crosses). These binary labeled data are noticeably separated into two distinct regions, which could represent stocks with positive and negative returns in a given year. These two regions can be easily separated by numerous straight lines; three of them are shown in the right panel of Exhibit 7. The data are thus linearly separable, and any of the straight lines shown would be called a **linear classifier**—a binary classifier that makes its classification decision based on a linear combination of the features of each data point.

Exhibit 7: Scatterplots and Linear Separation of Labeled Data

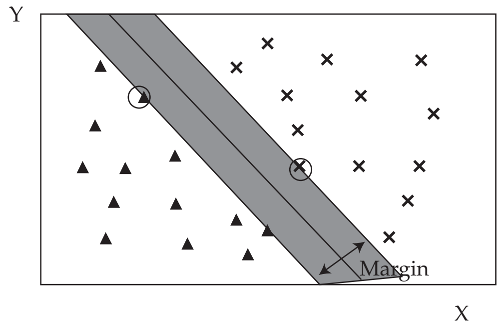
A. Data Labeled in Two Groups | *B. Data is Linearly Separable*



With two dimensions or features (x and y) linear classifiers can be represented as straight lines. Observations with n features can be represented in an n-dimension space, and the data set would be linearly separable if the observations can be separated into two distinct regions by a linear space boundary. The general term for such a space boundary is an n-dimensional hyperplane, which with n = 1 is called a line and with n = 2 is called a plane.

Support vector machine (SVM) is a linear classifier that determines the hyperplane that optimally separates the observations into two sets of data points. The intuitive idea behind the SVM algorithm is maximizing the probability of making a correct prediction (here, that an observation is a triangle or a cross) by determining the boundary that is the furthest away from all the observations. In Exhibit 8, SVM separates the data by the maximum margin, where the margin is the shaded strip that divides the observations into two groups. The straight line in the middle of the shaded strip is the discriminant boundary or boundary, for short. We can see that the SVM algorithm produces the widest shaded strip (i.e., the one with the maximum margin on either side of the boundary). The margin is determined by the observations closest to the boundary (the circled points) in each set, and these observations are called support vectors. Adding more training data away from the support vectors will not affect the boundary. In our training data sets, however, the result will be sensitive to the support vectors.

Exhibit 8 Linear Support Vector Machine Classifier



In Exhibit 8, SVM is classifying all observations perfectly. Many real-world data sets, however, are not perfectly linearly separable. Some observations may fall on the wrong side of the boundary and be misclassified by the SVM algorithm. The SVM algorithm handles this problem by an adaptation called **soft margin classification**, which adds a penalty to the objective function for observations in the training set that are misclassified. In essence, the SVM algorithm will choose a discriminant boundary that optimizes the trade-off between a wider margin and a lower total error penalty.

As an alternative to soft margin classification, a non-linear SVM algorithm can be run by introducing more advanced, non-linear separation boundaries. These algorithms will reduce the number of misclassified instances in the training data sets but will have more features, thus adding to the model's complexity.

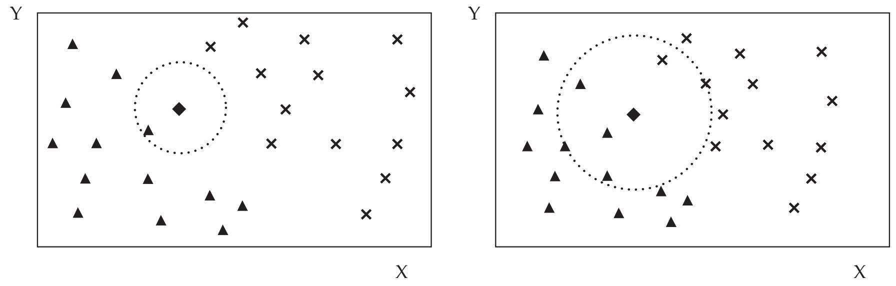
SVM has many applications in investment management. It is particularly suited for small- to medium-size but complex high-dimensional data sets, such as corporate financial statements or bankruptcy databases. Investors seek to predict company failures for identifying stocks to avoid or to short sell, and SVM can generate a binary classification (e.g., bankruptcy likely vs. bankruptcy unlikely) using many fundamental and technical feature variables. SVM can quickly capture the characteristics of such data with many features while being resilient to outliers and correlated features. SVM can also be used to classify text from documents (e.g., news articles, company announcements, and company annual reports) into useful categories for investors (e.g., positive sentiment and negative sentiment).

5.3 K-Nearest Neighbor

K-nearest neighbor (KNN) is a supervised learning technique most often used for classification and sometimes for regression. The idea is to classify a new observation by finding similarities ("nearness") between this new observation and the existing data. Going back to the scatterplot in Exhibit 7, let us assume we have a new observation: The diamond in Exhibit 9 needs to be classified as belonging to either the cross or the triangle category. If k = 1, the diamond will be classified into the same category as its nearest neighbor (i.e., the triangle in the left panel). The right panel in Exhibit 9 presents the case where k = 5, so the algorithm will look at the diamond's 5 nearest neighbors, which are 3 triangles and 2 crosses. The decision rule is to choose the classification with the largest number of nearest neighbors out of the 5 being considered. So, the diamond is again classified as belonging to the triangle category.

Exhibit 9 K-Nearest Neighbor (KNN) Algorithm

A. KNN With New Observation, K=1 *B. KNN With New Observation, K=5*



Let us suppose we have a database of corporate bonds classified by credit rating that also contains detailed information on the characteristics of these bonds. Such features would include those of the issuing company (e.g., asset size, industry, leverage ratios, cash flow ratios) and of the bond issue itself (e.g., tenor, fixed/floating coupon, embedded options). Now, assume a new bond is about to be issued with no credit rating. By nature, corporate bonds with similar issuer and issue characteristics should be given a similar credit rating. So, by using KNN, we can predict the implied credit rating of the new bond based on the similarities of its characteristics to those of the bonds in our database.

KNN is a straightforward, intuitive model that is still very powerful because it is non-parametric; the model makes no assumptions about the distribution of the data. Moreover, it can be used directly for multi-class classification. A critical challenge of KNN, however, is defining what it means to be "similar" (or near). Besides the selection of features, an important decision relates to the distance metric used to model similarity because an inappropriate measure will generate poorly performing models. The choice of a correct distance measure may be even more subjective for ordinal or categorical data. For example, if an analyst is looking at the similarities in market performance of various equities, he/she may consider using the correlation between the stocks' historical returns as an appropriate measure of similarity.

Knowledge of the data and understanding of the business objectives of the analysis are critical steps in the process of defining similarity. KNN results can be sensitive to inclusion of irrelevant or correlated features, so it may be necessary to select features manually. By doing so, the analyst removes less-valuable information to keep the most relevant and pertinent information. If done correctly, this process should generate a more representative distance measure. KNN algorithms tend to work better with a small number of features.

Finally, the number *k*, the hyperparameter of the model, must be chosen with the understanding that different values of *k* can lead to different conclusions. For predicting the credit rating of an unrated bond, for example, should *k* be the 3, 15, or 50 most similar bonds to the unrated bond? If *k* is an even number, there may be ties and no clear classification. Choosing a value for *k* that is too small would result in a high error rate and sensitivity to local outliers, but choosing a value for *k* that is too large would dilute the concept of nearest neighbors by averaging too many outcomes. In practice, several different techniques can be used to determine an optimal value for *k*, taking into account the number of categories and their partitioning of the feature space.

The KNN algorithm has many applications in the investment industry, including bankruptcy prediction, stock price prediction, corporate bond credit rating assignment, and customized equity and bond index creation.

5.4 Classification and Regression Tree

Classification and Regression Tree (CART) is another common supervised machine learning technique that can be applied to predict either a categorical target variable, producing a classification tree, or a continuous target variable, producing a regression tree. Most commonly, CART is applied to binary classification or regression.

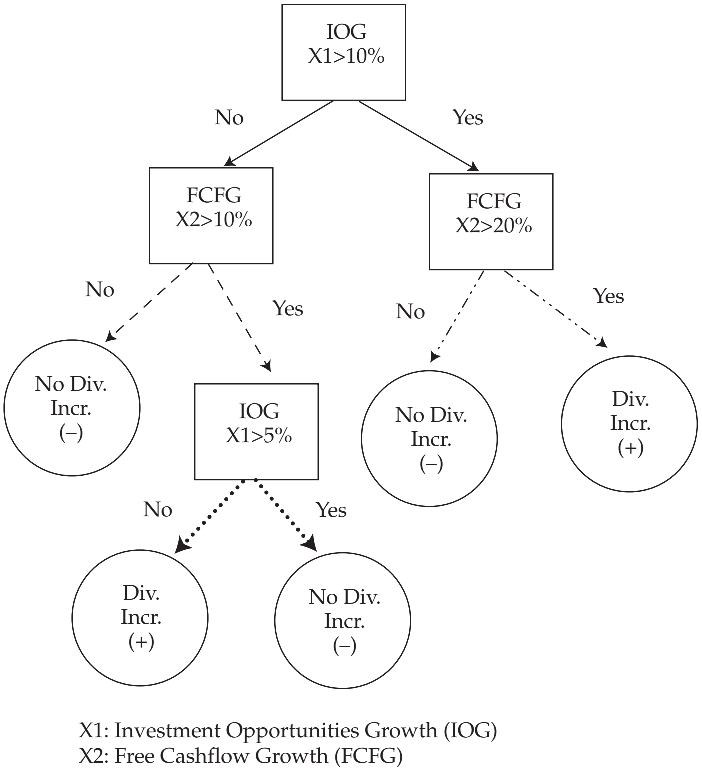
CART will be discussed in the context of a simplified model for classifying companies by whether or not they increase their dividends to shareholders. Such a classification requires a binary tree: a combination of an initial root node, decision nodes, and terminal nodes. The root node and each decision node represent a single feature (*f*) and a cutoff value (*c*) for that feature. As shown in Panel A of Exhibit 10, we start at the initial root node for a new data point. In this case, the initial root node represents the feature investment opportunities growth (IOG), designated as X1, with a cutoff value of 10%. From the initial root node, the data are partitioned at decision nodes into smaller and smaller subgroups until terminal nodes are formed that contain the predicted labels. In this case, the predicted labels are either dividend increase (the cross) or no dividend increase (the dash).

Also shown in Panel A of Exhibit 10, if the value of feature IOG (X1) is greater than 10% (Yes), then we proceed to the decision node for free cash flow growth (FCFG), designated as X2, which has a cutoff value of 20%. Now, if the value of FCFG is not greater than 20% (No), then CART will predict that data point belongs to the no dividend increase (dash) category, which represents a terminal node. Conversely, if the value of X2 is greater than 20% (Yes), then CART will predict that data point belongs to the dividend increase (cross) category, which represents another terminal node.

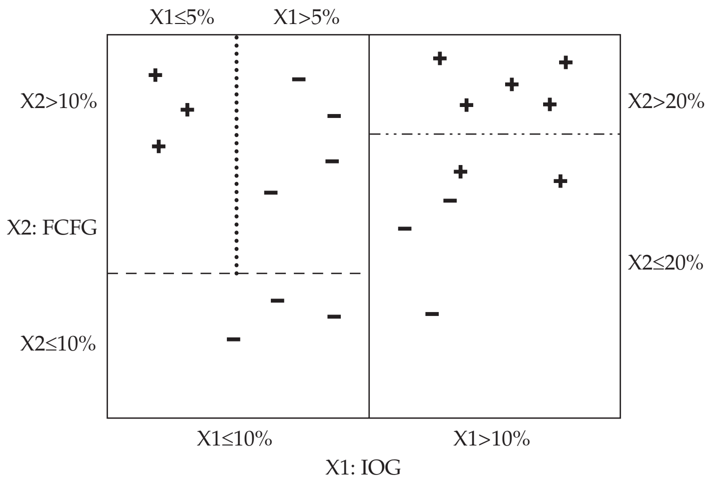
It is important to note that the same feature can appear several times in a tree in combination with other features. Moreover, some features may be relevant only if other conditions have been met. For example, going back to the initial root node, if IOG is not greater than 10% (X1 ≤10%) and FCFG is greater than 10%, then IOG appears again as another decision node, but this time it is lower down in the tree and has a cutoff value of 5%.

Exhibit 10 Classification and Regression Tree—Decision Tree and Partitioning of the Feature Space

A. Decision Tree



B. Partitioning of the Feature (X1,X2) Feature Space



We now turn to how the CART algorithm selects features and cutoff values for them. Initially, the classification model is trained from the labeled data, which in this hypothetical case are 10 instances of companies having dividend increase (the crosses) and 10 instances of companies with no dividend increase (the dashes). As shown in Panel B of Exhibit 10, at the initial root node and at each decision node the feature space (i.e., the plane defined by X1 and X2) is split into two rectangles for values above and below the cutoff value for the particular feature represented at that node. This can be seen by noting the distinct patterns of the lines that emanate from the decision nodes in Panel A. These same distinct patterns are used for partitioning the feature space in Panel B.

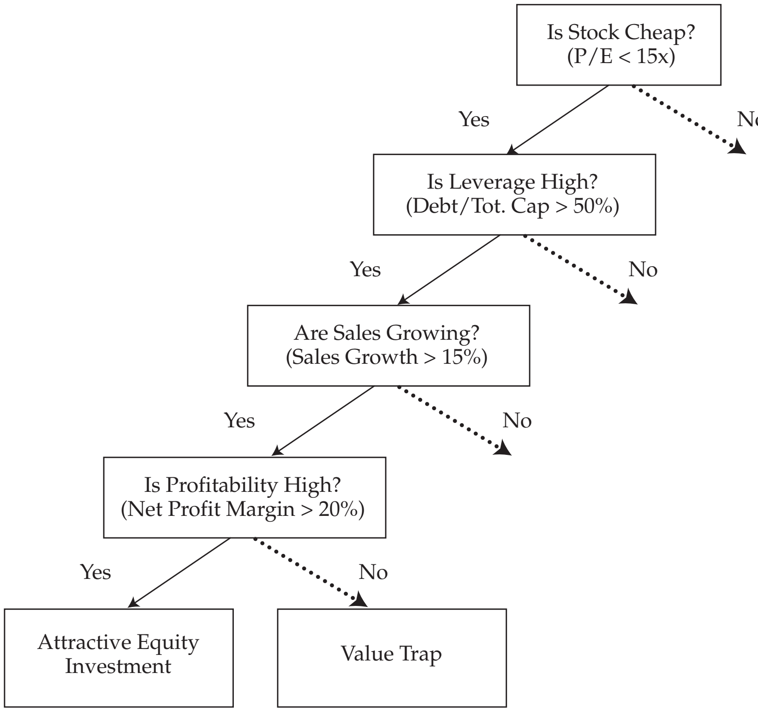
The CART algorithm chooses the feature and the cutoff value at each node that generates the widest separation of the labeled data to minimize classification error (e.g., by a criterion, such as mean-squared error). After each decision node, the partition of the feature space becomes smaller and smaller, so observations in each group have lower within-group error than before. At any level of the tree, when the classification error does not diminish much more from another split (bifurcation), the process stops, the node is a terminal node, and the category that is in the majority at that node is assigned to it. If the objective of the model is classification, then the prediction of the algorithm at each terminal node will be the category with the majority of data points. For example, in Panel B of Exhibit 10, the top right rectangle of the feature space, representing IOG (X1)>10% and FCFG (X2)>20%, contains 5 crosses, the most data points of any of the partitions. So, CART would predict that a new data point (i.e., a company) with such features belongs to the cross (dividend increase) category. However, if instead the new data point had IOG (X1)>10% and FCFG (X2) ≤20%, then it would be predicted to belong to the dash (no dividend increase) category—represented by the lower right rectangle with 2 crosses but with 3 dashes. Finally, if the goal is regression, then the prediction at each terminal node is the mean of the labeled values.

CART makes no assumptions about the characteristics of the training data, so if left unconstrained, potentially it can perfectly learn the training data. To avoid such overfitting, regularization parameters can be added, such as the maximum depth of the tree, the minimum population at a node, or the maximum number of decision nodes. The iterative process of building nodes is stopped once the regularization criterion has been reached. For example, in Panel B of Exhibit 10, the upper left rectangle of the feature space (determined by X1≤10%, X2>10%, and X1≤5% with three crosses) might represent a terminal node resulting from a regularization criterion with minimum population equal to 3. Alternatively, regularization can occur via a **pruning** technique that can be used afterward to reduce the size of the tree. Sections of the tree that provide little classifying power are pruned (i.e., removed).

By its iterative structure, CART can uncover complex dependencies between features that other models cannot reveal. As demonstrated in Exhibit 10, the same feature can appear several times in combination with other features and some features may be relevant only if other conditions have been met.

As shown in Exhibit 11, high profitability is a critical feature for predicting if a stock is an attractive investment or a value trap (i.e., an investment that, although apparently priced cheaply, is likely to be unprofitable). This feature is relevant only if the stock is cheap—for example, in this hypothetical case if P/E is less than 15, leverage is high (debt to total capital > 50%), and sales are expanding (sales growth > 15%). Said another way, high profitability is irrelevant in this context if the stock is not cheap, *and* if leverage is not high, *and* if sales are not expanding. Multiple linear regression typically fails in such situations where the relationship between the features and the outcome is non-linear.

Exhibit 11 Stylized Decision Tree—Attractive Investment or Value Trap?



CART is a popular supervised machine learning model because the tree provides a visual explanation for the prediction. This contrasts favorably with other algorithms that are often considered to be "black boxes" because it may be difficult to understand the reasoning behind their outcomes and thus to foster trust in them. CART is a powerful tool to build expert systems for decision-making processes. It can induce robust rules despite noisy data and complex relationships between high numbers of features. Typical applications of CART in investment management include, among others, enhancing detection of fraud in financial statements, generating consistent decision processes in equity and fixed-income selection, and simplifying communication of investment strategies to clients.

5.5 Ensemble Learning and Random Forest

Instead of basing predictions on the results of a single model as in the previous discussion, why not use the predictions of a group—or an ensemble—of models? Each single model will have a certain error rate and will make noisy predictions. But by taking the average result of many predictions from many models, we can expect to achieve a reduction in noise as the average result converges toward a more accurate prediction. This technique of combining the predictions from a collection of models is called **ensemble learning**, and the combination of multiple learning algorithms is known as the **ensemble method**. Ensemble learning typically produces more accurate and more stable predictions than the best single model. In fact, in many prestigious machine learning competitions, an ensemble method is often the winning solution.

Ensemble learning can be divided into two main categories: 1) an ensemble method can be an aggregation of heterogeneous learners (i.e., different types of algorithms combined together with a voting classifier); or 2) an ensemble method can be an aggregation of homogenous learners (i.e., a combination of the same algorithm, using different training data that are based, for example, on a bootstrap aggregating (i.e., bagging) technique as discussed later).

5.5.1 *Voting Classifiers*

Suppose you have been working on a machine learning project for some time and have trained and compared the results of several algorithms, such as SVM, KNN, and CART. A **majority-vote classifier** will assign to a new data point the predicted label with the most votes. For example, if the SVM and KNN models are both predicting the category "stock outperformance" and the CART model is predicting the category "stock underperformance," then the majority-vote classifier will choose "'stock outperformance." The more individual models you have trained, the higher the accuracy of the aggregated prediction up to a point. There is an optimal number of models beyond which performance would be expected to deteriorate from overfitting. The trick is to look for diversity in the choice of algorithms, modelling techniques, and hypotheses. The (extreme) assumption here is that if the predictions of the individual models are independent, then we can use the law of large numbers to achieve a more accurate prediction.

5.5.2 *Bootstrap Aggregating (Bagging)*

Alternatively, one can use the same machine learning algorithm but with different training data. **Bootstrap aggregating (or bagging)** is a technique whereby the original training data set is used to generate n new training data sets or bags of data. Each new bag of data is generated by random sampling with replacement from the initial training set. The algorithm can now be trained on n independent data sets that will generate n new models. Then, for each new observation, we can aggregate the n predictions using a majority-vote classifier for a classification or an average for a regression. Bagging is a very useful technique because it helps to improve the stability of predictions and protects against overfitting the model.

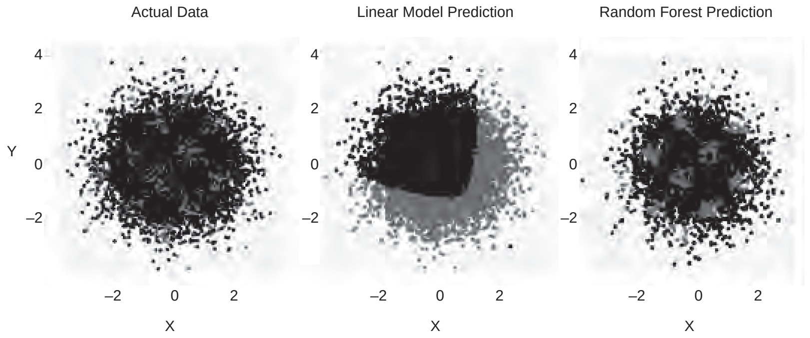
5.5.3 *Random Forest*

A **random forest classifier** is a collection of a large number of decision trees trained via a bagging method. For example, a CART algorithm would be trained using each of the n independent data sets (from the bagging process) to generate the multitude of different decision trees that make up the random forest classifier.

To derive even more individual predictions, added diversity can be generated in the trees by randomly reducing the number of features available during training. So, if each observation has n features, one can randomly select a subset of m features (where $m < n$) that will then be considered by the CART algorithm for splitting the data set at each of the decision nodes. The number of subset features (m), the number of trees to use, the minimum size (population) of each node (or leaf), and the maximum depth of each tree are all hyperparameters that can be tuned to improve overall model prediction accuracy. For any new observation, we let all the classifier trees (the "random forest") undertake classification by majority vote—implementing a machine learning version of the "wisdom of crowds." The process involved in random forest tends to protect against overfitting on the training data. It also reduces the ratio of noise to signal because errors cancel out across the collection of slightly different classification trees. However, an important drawback of random forest is that it lacks the ease of interpretability of individual trees; as a result, it is considered a relatively black box-type algorithm.

Exhibit 12 presents three scatterplots of actual and predicted defaults by small- and medium-sized businesses with respect to two features, X and Y—for example, firm profitability and leverage, respectively. The left plot shows the actual cases of default in light shade and no default in dark shade, while the middle and right plots present the predicted defaults and no defaults (also in light and dark shades, respectively). It is clear from the middle plot, which is based on a traditional linear regression model, that the model fails to predict the complex non-linear relationship between the features. Conversely, the right plot, which presents the prediction results of a random forest model, shows that this model performs very well in matching the actual distribution of the data.

Exhibit 12 Credit Defaults of Small- and Medium-Sized Borrowers



Source: Bacham and Zhao (2017).

Ensemble Learning with Random Forest

In making use of voting across classifier trees, random forest is an example of ensemble learning: Incorporating the output of a collection of models produces classifications that have better signal-to-noise ratios than the individual classifiers. A good example of a credit card fraud detection problem comes from an open source data set on Kaggle.[1] Here, the data contained several anonymized features that might be used to explain which transactions were fraudulent. The difficulty in the analysis arises from the fact that the rate of fraudulent transactions is very low; in a sample of 284,807 transactions, only 492 were fraudulent (0.17%). This is akin to finding a needle in a haystack. Applying a random forest classification algorithm with an oversampling technique—which involves increasing the proportional representation of fraudulent data in the training set—does extremely well. Despite the lopsided sample, it delivers precision (the percentage of cases marked fraudulent that were actually fraudulent) of 89% and recall (the percentage of fraudulent cases that were marked as fraudulent) of 82%.

Despite its relative simplicity, random forest is a powerful algorithm with many investment applications. These include, for example, use in factor-based investment strategies for asset allocation and investment selection or use in predicting whether an IPO will be successful (e.g., percent oversubscribed, first trading day close/IPO price) given the attributes of the IPO offering and the corporate issuer.

1 See https://www.kaggle.com/mlg-ulb/creditcardfraud (accessed 1 October 2018).

EXAMPLE 3

Support Vector Machine and K-Nearest Neighbor

Rachel Lee is a fixed-income portfolio manager analyst with Zeta Investment Management Company. Zeta manages an investment-grade bond portfolio for small, conservative institutions and a non-investment-grade (i.e., high-yield) bond portfolio for yield-seeking, high-net-worth individuals. Both portfolios can hold unrated bonds if the characteristics of the unrated bonds closely match those of the respective portfolio's average holding.

Lee is discussing an upcoming straight, 10-year fixed coupon bond issue with senior credit analyst Marc Watson. Watson comments that although the bond's issuer, Biotron Corporation, has not had this issue rated, his analysis of the company's profitability, cash flow, leverage, and coverage ratios places the issue near the borderline between low investment-grade (Baa3/BBB–) and high non-investment-grade (Ba1/BB+) bonds.

Lee decides to use machine learning methods to confirm the implied credit rating of Biotron Corporation.

1 State the type of problem being addressed by Lee.

2 State two ML algorithms that Lee could use to explore the implied credit rating of Biotron Corporation, and then describe how each algorithm could be applied.

Lee decides to apply the two identified ML algorithms. Both algorithms clearly support a high non-investment-grade rating. Watson states that because both ML algorithms agree on the rating, he has confidence in relying on the rating.

3 State one argument in support of Watson's viewpoint.

Solution to 1:

Lee is addressing a supervised learning classification problem because she must determine whether Biotron's upcoming bond issue would be classified as investment grade or non-investment grade.

Solution to 2:

One suitable ML algorithm is the SVM. The SVM algorithm is a linear classifier that aims to seek the optimal hyperplane—the one that separates observations into two distinct sets by the maximum margin. So, the SVM is well suited to binary classification problems, such as the one facing Lee (investment grade vs. non-investment grade). In this case, Lee could train the SVM algorithm on data—characteristics (features) and rating (target)—of low investment-grade (Baa3/BBB–) and high non-investment-grade (Ba1/BB+) bonds. Lee would then note on which side of the margin the new data point (Biotron's new bonds) lies.

The KNN algorithm is also well suited for classification problems because it classifies a new observation by finding similarities (or nearness) between the new observation and the existing data. Training the algorithm with data as for SVM, the decision rule for classifying Biotron's new bonds is which classification is in the majority among its k-nearest neighbors. Note that k must be pre-specified by Lee.

Solution to 3:

If the ML algorithms disagreed on the classification, the classification would be more likely to be sensitive to the algorithm's approach to classifying data. Because the classification of Biotron's new issue appears robust to the choice of ML algorithm (i.e., both algorithms agree on the rating), the resulting classification will likely be correct.

EXAMPLE 4

CART and Ensemble Learning

Laurie Kim is a portfolio manager at Hilux LLC, a high-yield bond investment firm. The economy has been in recession for several months, and high-yield bond prices have declined precipitously as credit spreads have widened in response to the weak macroeconomic environment. Kim, however, believes this is a good time to buy as she expects to profit as credit spreads narrow and high-yield bond prices rise in anticipation of economic recovery.

Based on her analysis, Kim believes that corporate high-yield bonds in the credit quality range of B/B2 to CCC/Caa2 are the most attractive. However, she must carefully select which bonds to buy and which bonds to avoid because of the elevated default risk caused by the currently weak economy.

To help with her bond selection, Kim turns to Hilux's data analytics team. Kim has supplied them with historical data consisting of 19 fundamental and 5 technical factors for several thousand high-yield bond issuers and issues labeled to indicate default or no default. Kim requests that the team develop an ML-based model that will make accurate classifications in two categories: default and no default. Exploratory data analysis suggests considerable non-linearities among the feature set.

1 State the type of problem being addressed by Kim.
2 Describe the dimensionality of the model that Kim requests her analytics team to develop.
3 Evaluate whether a CART model is appropriate for addressing her problem.
4 Describe how a CART model operates at each node of the tree.
5 Describe how the team might avoid overfitting and improve the predictive power of a CART model.
6 Describe how ensemble learning might be used by the team to develop even better predictions for Kim's selection of corporate high-yield bonds.

Solution to 1:

Kim is addressing a classification problem because she must determine whether bonds that she is considering purchasing in the credit quality range of B/B2 to CCC/Caa2 will default or not default.

Solution to 2:

With 19 fundamental and 5 technical factors (i.e., the features) the dimensionality of the model is 24.

Solution to 3:

The CART model is an available algorithm for addressing classification problems. Its ability to handle complex, non-linear relationships makes it a good choice to address the modelling problem at hand. An important advantage of CART is that its results are relatively straightforward to visualize and interpret, which should help Kim explain her recommendations based on the model to Hilux's investment committee and the firm's clients.

Solution to 4:

At each node in the decision tree, the algorithm will choose the feature and the cutoff value for the selected feature that generates the widest separation of the labeled data to minimize classification error.

Solution to 5:

The team can avoid overfitting and improve the predictive power of the CART model by adding regularization parameters. For example, the team could specify the maximum depth of the tree, the minimum population at a node, or the maximum number of decision nodes. The iterative process of building nodes will be stopped once the regularization criterion has been reached. Alternatively, a pruning technique can be used afterward to remove parts of the CART model that provide little power to correctly classify instances into default or no default categories.

Solution to 6:

The analytics team might use ensemble learning to combine the predictions from a collection of models, where the average result of many predictions leads to a reduction in noise and thus more accurate predictions. Ensemble learning can be achieved by an aggregation of either heterogeneous learners—different types of algorithms combined with a voting classifier—or homogenous learners—a combination of the same algorithm but using different training data based on the bootstrap aggregating (i.e., bagging) technique. The team may also consider developing a random forest classifier (i.e., a collection of many decision trees) trained via a bagging method.

6 UNSUPERVISED MACHINE LEARNING ALGORITHMS

As you may recall, unsupervised learning is machine learning that does not use labeled data (i.e., no target variable); thus, the algorithms are tasked with finding patterns within the data themselves. The two main types of unsupervised ML algorithms shown in Exhibits 2 and 3 are dimension reduction, using principal components analysis, and clustering, which includes k-means and hierarchical clustering. These will now be described in turn.

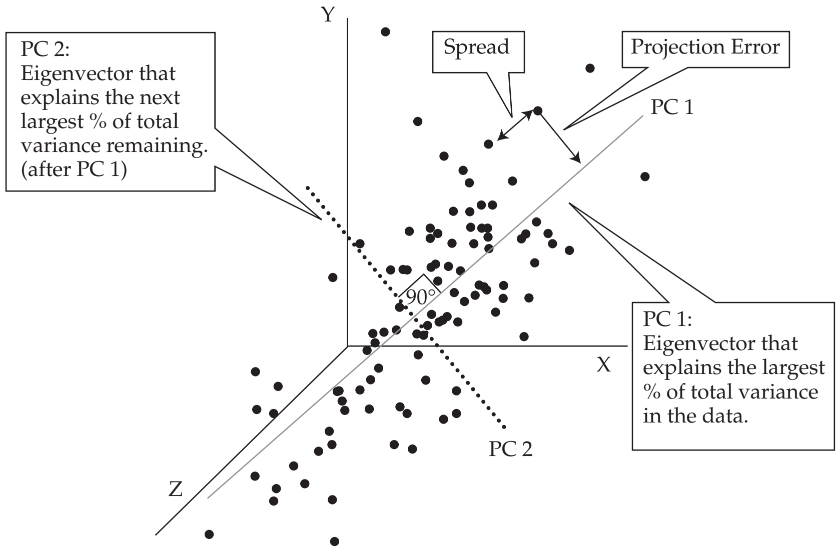
6.1 Principal Components Analysis

Dimension reduction is an important type of unsupervised learning that is used widely in practice. When many features are in a data set, representing the data visually or fitting models to the data may become extremely complex and "noisy" in the sense of reflecting random influences specific to a data set. In such cases, dimension reduction may be necessary. Dimension reduction aims to represent a data set with many, typically correlated, features by a smaller set of features that still do well in describing the data.

A long-established statistical method for dimension reduction is principal components analysis (PCA). PCA is used to summarize or reduce highly correlated features of data into a few main, uncorrelated composite variables. A **composite variable** is a variable that combines two or more variables that are statistically strongly related to each other. Informally, PCA involves transforming the covariance matrix of the features and involves two key concepts: eigenvectors and eigenvalues. The **eigenvectors** define new, mutually uncorrelated composite variables that are linear combinations of the original features. As a vector, an eigenvector also represents a direction. Associated with each eigenvector is an eigenvalue. An **eigenvalue** gives the proportion of total variance in the initial data that is explained by each eigenvector. The PCA algorithm orders the eigenvectors from highest to lowest according to their eigenvalues—that is, in terms of their usefulness in explaining the total variance in the initial data. PCA selects as the first principal component the eigenvector that explains the largest proportion of variation in the data set (the eigenvector with the largest eigenvalue). The second principal component explains the next largest proportion of variation remaining after the first principal component; it continues for the third, fourth, and subsequent principal components. As the principal components are linear combinations of the initial feature set, only a few principal components are typically required to explain most of the total variance in the initial feature covariance matrix.

Exhibit 13 shows a hypothetical data set with three features, so it is plotted in three dimensions along the x, y, and z axes. Each data point has a measurement (x, y, z), and the data should be standardized so that the mean of each series (x's, y's, and z's) is 0 and the standard deviation is 1. Assume PCA has been applied, revealing the first two principal components, PC1 and PC2. With respect to PC1, a perpendicular line dropped from each data point to PC1 shows the vertical distance between the data point and PC1, representing **projection error**. Moreover, the distance between each data point in the direction that is parallel to PC1 represents the spread or variation of the data along PC1. The PCA algorithm operates in such a way that it finds PC1 by selecting the line for which the sum of the projection errors for all data points is *minimized* and for which the sum of the spread between all the data is *maximized*. As a consequence of these selection criteria, PC1 is the unique vector that accounts for the largest proportion of the variance in the initial data. The next largest portion of the remaining variance is best explained by PC2, which is at right angles to PC1 and thus is uncorrelated with PC1. The data points can now be represented by the first two principal components. This example demonstrates the effectiveness of the PCA algorithm in summarizing the variability of the data and the resulting dimension reduction.

Exhibit 13 First and Second Principal Components of a Hypothetical 3-Dimensional Data Set

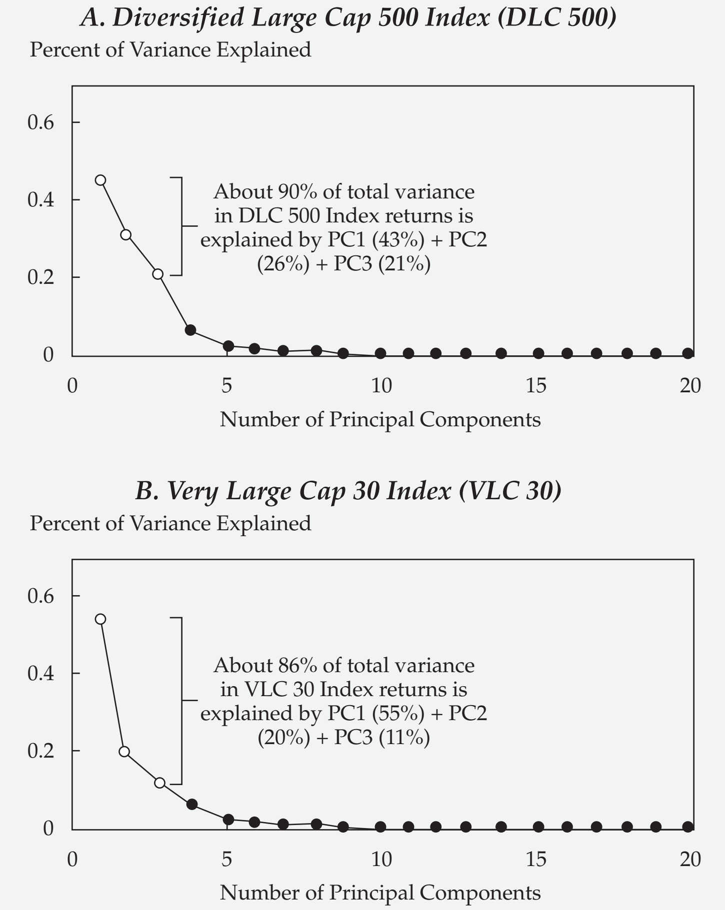


It is important to know how many principal components to retain because there is a trade-off between a lower-dimensional, more manageable view of a complex data set when a few are selected and some loss of information. **Scree plots**, which show the proportion of total variance in the data explained by each principal component, can be helpful in this regard (see the accompanying sidebar). In practice, the smallest number of principal components that should be retained is that which the scree plot shows as explaining 85% to 95% of total variance in the initial data set.

Scree Plots for the Principal Components of Returns to the Hypothetical DLC 500 and VLC 30 Equity Indexes

In this illustration, researchers use scree plots and decide that three principal components are sufficient for explaining the returns to the hypothetical DLC 500 and VLC 30 equity indexes over the last 10-year period. The DLC 500 can be thought of as a diversified index of large-cap companies covering all economic sectors, while the VLC 30 is a more concentrated index of the 30 largest publicly traded companies. The data set consists of index prices and more than 2,000 fundamental and technical features. Multi-collinearity among the features is a typical problem because that many features or combinations of features would tend to have overlaps. To mitigate the problem, PCA can be used to capture the information and variance in the data. The following scree plots show that of the 20 principal components generated, the first 3 together explain about 90% and 86% of the variance in the value of the DLC 500 and VLC 30 indexes, respectively. The scree plots indicate that for each of these indexes, the incremental contribution to explaining the variance structure of the data is quite small after about the fifth principal component. Therefore, these less useful principal components can be ignored without much loss of information.

Scree Plots of Percent of Total Variance Explained by Each Principal Component for Hypothetical DLC 500 and VLC 30 Equity Indexes



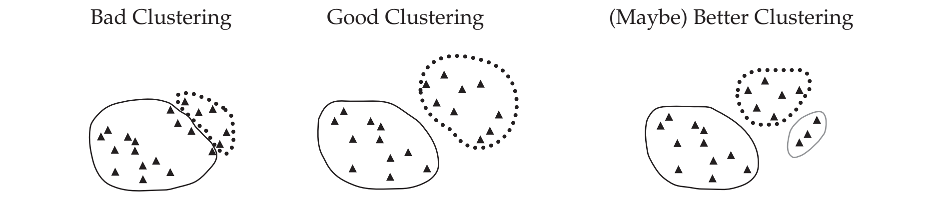
The main drawback of PCA is that since the principal components are combinations of the data set's initial features, they typically cannot be easily labeled or directly interpreted by the analyst. Compared to modelling data with variables that represent well-defined concepts, the end user of PCA may perceive PCA as something of a "black box."

Reducing the number of features to the most relevant predictors is very useful, even when working with data sets having as few as ten or so features. Notably, dimension reduction facilitates visually representing the data in two or three dimensions. It is typically performed as part of exploratory data analysis, before training another supervised or unsupervised learning model. Machine learning models are quicker to train, tend to reduce overfitting (by avoiding the curse of dimensionality), and are easier to interpret if provided with lower dimensional data sets.

6.2 Clustering

Clustering is another type of unsupervised machine learning that is used to organize data points into similar groups called clusters. A **cluster** contains a subset of observations from the data set such that all the observations within the same cluster are deemed "similar." The aim is to find a good clustering of the data—meaning that the observations inside each cluster are similar or close to each other (a property known as cohesion) and the observations in two different clusters are as far away from one another or are as dissimilar as possible (a property known as separation). Exhibit 14 depicts this intra-cluster cohesion and inter-cluster separation.

Exhibit 14 Evaluating Clustering—Intra-Cluster Cohesion and Inter-Cluster Separation



Clustering algorithms are particularly useful in the many investment problems and applications in which the concept of similarity is important. Applied to grouping companies, for example, clustering may uncover important similarities and differences among companies that are not captured by standard classifications of companies by industry and sector. In portfolio management, clustering methods have been used for improving portfolio diversification.

In practice, expert human judgment has a role in using clustering algorithms. In the first place, one must establish what it means to be "similar." Each company can be considered an observation with multiple features, including such financial statement items as total revenue and profit to shareholders, a wide array of financial ratios, or any other potential model inputs. Based on these features, a measure of similarity or "distance" between two observations (i.e., companies) can be defined. The smaller the distance, the more similar the observations; the larger the distance, the more dissimilar the observations.

A commonly used definition of distance is the Euclidian distance, the straight-line distance between two points. Roughly a dozen different distance measures are used regularly in ML. In practice, the choice of the distance measures depends on the nature of the data (numerical or not) and the business problem being investigated. Once the relevant distance measure is defined, similar observations can be grouped together. We now introduce two of the more popular clustering approaches: k-means and hierarchical clustering.

6.2.1 *K-Means Clustering*

K-means is a relatively old algorithm that repeatedly partitions observations into a fixed number, k, of non-overlapping clusters. The number of clusters, k, is a model hyperparameter—a parameter whose value must be set by the researcher before learning begins. Each cluster is characterized by its **centroid** (i.e., center), and each observation is assigned by the algorithm to the cluster with the centroid to which that observation is closest.

The k-means algorithm follows an iterative process. It is illustrated in Exhibit 15 for $k = 3$ and a set of observations on a variable that can be described by two features. In Exhibit 15, the horizontal and vertical axes represent, respectively, the first and second characteristics. For example, an investment analyst may want to group a set of firms into three groups according to two numerical measures of management quality. The algorithm groups the observations in the following steps:

1 K-means starts by determining the position of the k (here, 3) initial random centroids.

2 The algorithm then analyzes the features for each observation. Based on the distance measure that is utilized, k-means assigns each observation to its closest centroid, which defines a cluster.

3 Using the observations within each cluster, k-means then calculates the new (k) centroids for each cluster, where the centroid is the average value of their assigned observations.

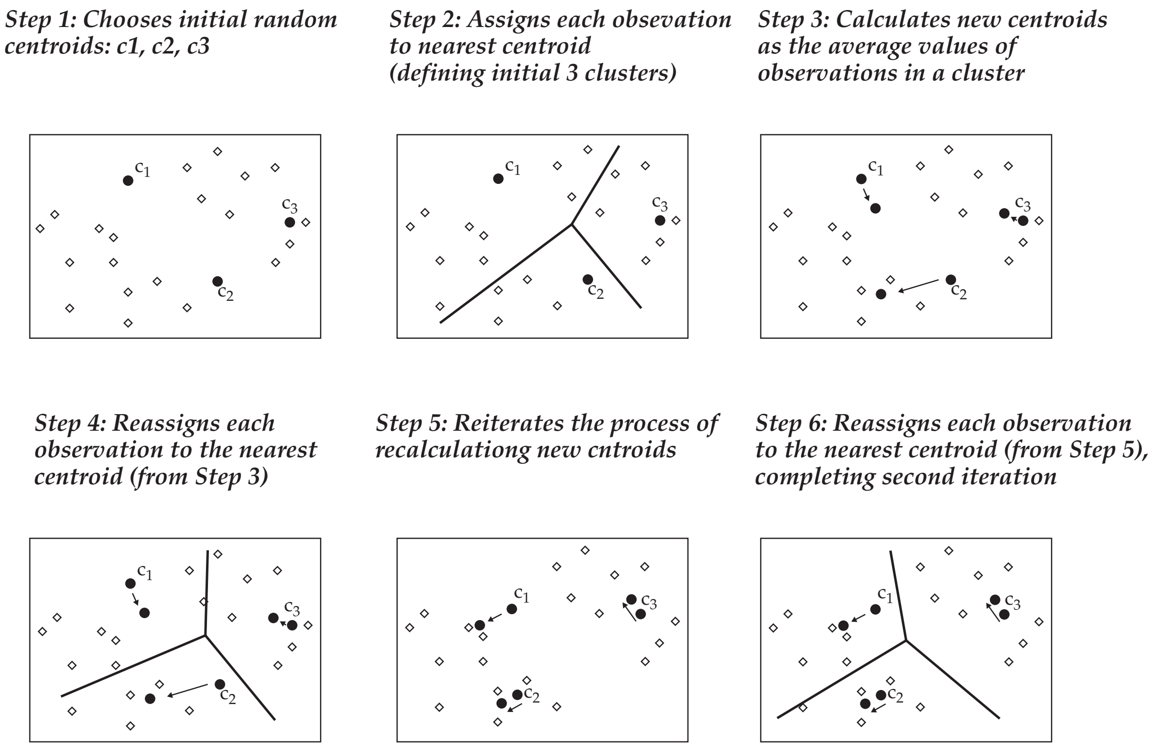
4 K-means then reassigns the observations to the new centroids, redefining the clusters in terms of included and excluded observations.

5 The process of recalculating the new (k) centroids for each cluster is reiterated.

6 K-means then reassigns the observations to the revised centroids, again redefining the clusters in terms of observations that are included and excluded.

The k-means algorithm will continue to iterate until no observation is reassigned to a new cluster (i.e., no need to recalculate new centroids). The algorithm has then converged and reveals the final k clusters with their member observations. The k-means algorithm has minimized intra-cluster distance (thereby maximizing cohesion) and has maximized inter-cluster distance (thereby maximizing separation) under the constraint that $k = 3$.

Exhibit 15 Example of 3-Means Algorithm



The k-means algorithm is fast and works well on very large data sets with hundreds of millions of observations. However, the final assignment of observations to clusters can depend on the initial location of the centroids. To address this problem, the algorithm can be run several times using different sets of initial centroids, and then one can choose the clustering that is most useful given the business purpose.

One limitation of this technique is that the hyperparameter, k, the number of clusters in which to partition the data, must be decided before k-means can be run. So, one needs to have a sense of how many clusters are reasonable for the problem under investigation and the data set being analyzed. Alternatively, one can run the algorithm using a range of values for *k* to find the optimal number of clusters—the *k* that minimizes intra-cluster distance and thereby maximizes intra-cluster similarity (i.e., cohesion) and that maximizes inter-cluster distance (i.e., separation). However, note that the final results can be subjective and dependent on the context of the problem and the particular training set.

For example, consider the Russell 3000 Index, which tracks the 3,000 highest market capitalization stocks in the United States. These 3,000 stocks can be grouped in 10, 50, or even more clusters based on their financial characteristics (e.g., total assets, total revenue, profitability, leverage) and operating characteristics (e.g., employee headcount, R&D intensity). As companies in the same standard industry classification can have very different financial and operating characteristics, using k-means to derive different clusters can provide insights and understanding into the nature of "peer" groups. As mentioned, the exact choice of the *k*, the number of clusters, will depend on the level of precision or segmentation desired. In a similar vein, clustering can be used to classify collective investment vehicles or hedge funds as an alternative to standard classifications. Clustering analysis can also help visualize the data and facilitate detecting trends or outliers.

In sum, the k-means algorithm is among the most used algorithms in investment practice, particularly in data exploration for discovering patterns in high dimensional data or as a method for deriving alternatives to existing static industry classifications.

6.2.2 *Hierarchical Clustering: Agglomerative and Divisive*

Hierarchical clustering is an iterative procedure used to build a hierarchy of clusters. In k-means clustering, the algorithm segments the data into a predetermined number of clusters; there is no defined relationship among the resulting clusters. In hierarchical clustering, however, the algorithms create intermediate rounds of clusters of increasing (in "agglomerative") or decreasing (in "divisive") size until a final clustering is reached. The process creates relationships among the rounds of clusters, as the word "hierarchical" suggests. Although more computationally intensive than k-means clustering, hierarchical clustering has the advantage of allowing the investment analyst to examine alternative segmentations of data of different granularity before deciding which one to use.

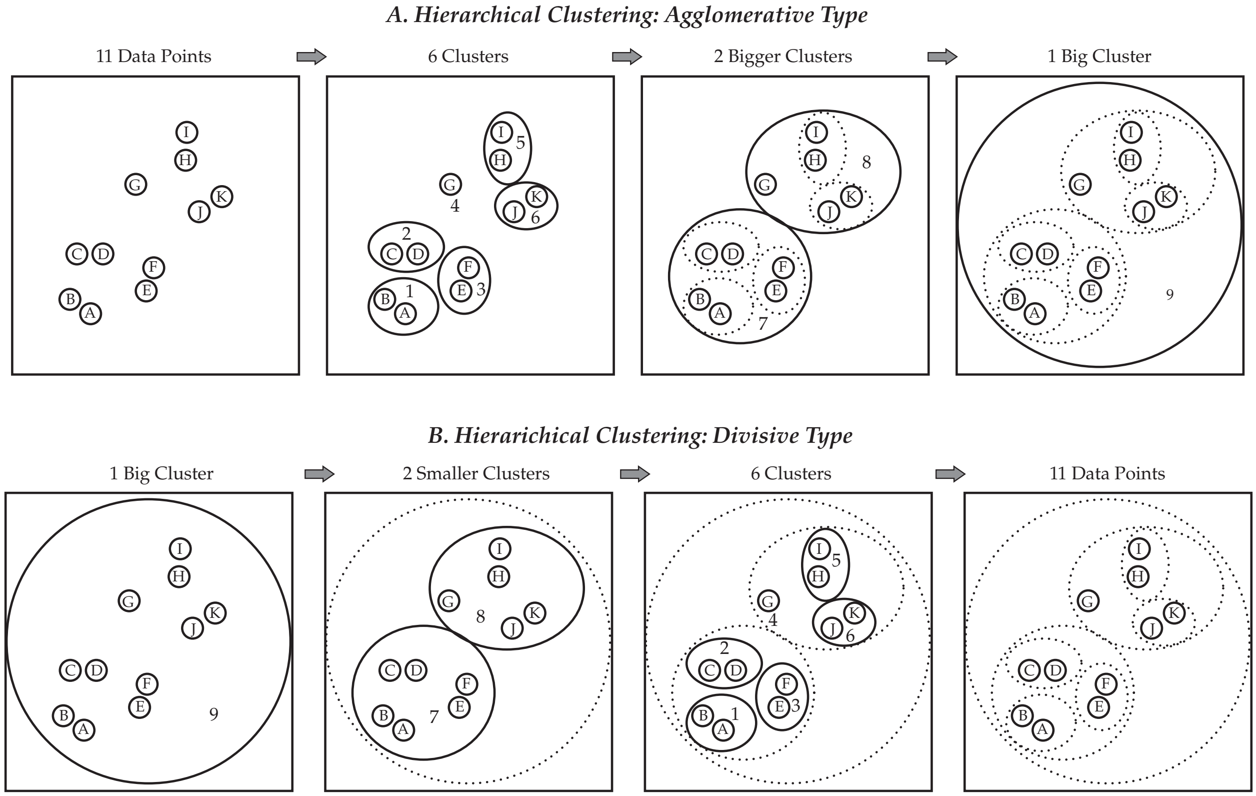
In more detail, **agglomerative clustering** (or bottom-up) hierarchical clustering begins with each observation being treated as its own cluster. Then, the algorithm finds the two closest clusters, defined by some measure of distance (similarity), and combines them into one new larger cluster. This process is repeated iteratively until all observations are clumped into a single cluster. A hypothetical example of how agglomerative clustering develops a hierarchical clustering scheme is depicted in the top part of Exhibit 16, where observations are lettered (A to K) and circles around observations denote clusters. The process begins with 11 individual clusters and then generates a sequence of groupings. The first sequence includes 5 clusters with 2 observations each and 1 cluster with a single observation, G, for a total of 6 clusters. It then generates 2 clusters, 1 cluster with 6 observations and the other with 5 observations. The final result is 1 large cluster containing all 11 observations. It is easily seen that this final large cluster includes the two main sub-clusters, with each containing three smaller sub-clusters.

By contrast, **divisive clustering** (or top-down) hierarchical clustering starts with all the observations belonging to a single cluster. The observations are then divided into two clusters based on some measure of distance (similarity). The algorithm then progressively partitions the intermediate clusters into smaller clusters until each cluster contains only 1 observation. Divisive clustering is depicted in the bottom part of Exhibit 16, which begins with all 11 observations in 1 large cluster. Next, the algorithm generates 2 smaller clusters, one with 6 observations and the other with

5 observations; then 6 clusters, with 2 observations each except for observation G, which is its own cluster. Finally, 11 clusters are generated, with each cluster containing only 1 observation.

In this hypothetical illustration, but not typically (as the two methods generally use different algorithms), the agglomerative and divisive clustering methods produced the same result: two main sub-clusters each having three smaller sub-clusters. The analyst could decide between using a 6- or a 2-cluster representation of the data. The agglomerative method is the approach typically used with large data sets because of the algorithm's fast computing speed. The agglomerative clustering algorithm makes clustering decisions based on local patterns without initially accounting for the global structure of the data. As such, the agglomerative method is well suited for identifying small clusters. However, as the divisive method starts with a holistic representation of the data, the divisive clustering algorithm is designed to account for the global structure of the data and thus is better suited for identifying large clusters.

Exhibit 16 Agglomerative and Divisive Hierarchical Clustering



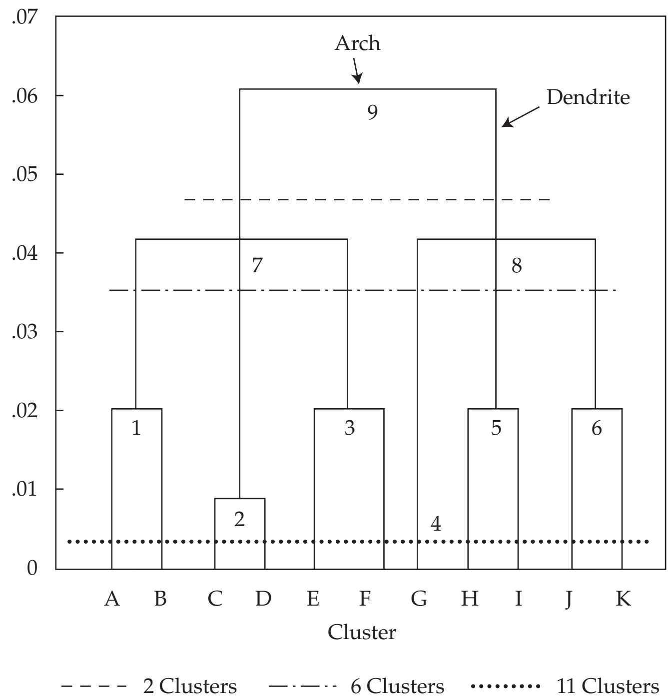
To decide on the closest clusters for combining in the agglomerative process or for dividing in the divisive process, an explicit definition for the distance between two clusters is required. Some commonly used definitions for the distance between two clusters involve finding the minimum, the maximum, or the average of the straight-line distances between all the pairs of observations in each cluster.

6.2.3 *Dendrograms*

A type of tree diagram for visualizing a hierarchical cluster analysis known as a **dendrogram** highlights the hierarchical relationships among the clusters. Exhibit 17 shows a dendrogram representation for the clustering shown in Exhibit 16. First, a few technical points on dendrograms bear mentioning—although they may not all be apparent in Exhibit 17. The x-axis shows the clusters, and the y-axis indicates some distance measure. Clusters are represented by a horizontal line, the arch, which connects two vertical lines, called dendrites, where the height of each arch represents the distance between the two clusters being considered. Shorter dendrites represent a shorter distance (and greater similarity) between clusters. The horizontal dashed lines cutting across the dendrites show the number of clusters into which the data are split at each stage.

The agglomerative algorithm starts at the bottom of the dendrite where each observation is its own cluster (A to K). Agglomerative clustering then generates the 6 larger clusters (1 to 6). For example, clusters A and B combine to form cluster 1, and observation G remains its own cluster, now cluster 4. Moving up the dendrogram, 2 larger clusters are formed, where, for example, cluster 7 includes clusters 1 to 3. Finally, at the top of the dendrogram is the single large cluster (9). The dendrogram readily shows how this largest cluster is composed of the two main sub-clusters (7 and 8), each having three smaller sub-clusters (1 to 3 and 4 to 6, respectively). The dendrogram also facilitates visualization of divisive clustering by starting at the top of the largest cluster and then working downward until the bottom is reached where all 11 single-observation clusters are shown.

Exhibit 17 Dendrogram of Agglomerative Hierarchical Clustering



Clustering has many applications in investment management. For example, portfolio diversification can be approached as a clustering problem with the aim of optimally diversifying risks by investing in assets from multiple different clusters. Because the clusters have maximum inter-cluster separation, diversifying among them helps ensure that the portfolio reflects a wide diversity of characteristics with well-diversified risk. Alternatively, information that investments are concentrated in a cluster indicates a high probability of concentrated risk. Finally, it is important to note that while the results of clustering algorithms are often difficult to evaluate (because the resulting clusters themselves are not explicitly defined), they are still very useful in practice for uncovering important underlying structure (namely, similarities among observations) in complex data sets.

EXAMPLE 5

Investment Uses of Clustering Algorithms

István Perényi is a portfolio manager of the Europe Diversified Equity Fund ("the Fund") within the Diversified Investment Management Company (DIMCO) fund family. The Fund is benchmarked to the STOXX Europe 600 Index, which spans 17 countries, 19 industry sectors, and 3 market capitalization groupings (large-, mid-, and small-cap).

Examining the Fund's most recent performance, Perényi is concerned that the Fund's holdings, although approximately aligned with the STOXX Europe 600 Index's country weights, may have unrecognized risk biases and concentrations. Perényi requests Elsa Lund, DIMCO's chief risk officer, to investigate the Fund's diversification. Lund asks her analysts for ideas on how Perényi's request can be addressed and receives three suggestions:

Suggestion 1 Estimate the country, industry, and market-cap exposures of each Fund holding, aggregate them, and compare the aggregate exposures to the benchmark's exposures. Then, examine mismatches for evidence of unexpected biases or concentrations.

Suggestion 2 Identify inherent groupings among fund holdings based on a broad set of eight numerical (operating and financial) measures related to the holdings' characteristics. Then, examine the groupings for evidence of unexpected biases or concentrations.

Suggestion 3 Regress the return of the Fund on a set of country equity market indexes and sector indexes based on the Fund's benchmark. Then, examine the regression coefficients for evidence of unexpected biases or concentrations.

Lund has several questions for analyst Greg Kane about using one or more clustering machine learning algorithms in relation to addressing Perényi's request.

Lund asks if any information needs to be specified for ML clustering algorithms no matter which one is used. Kane replies that only the distance measure that the algorithm will use and the hyperparameter, *k*, for k-means clustering need to be specified.

Lund further asks whether there would be an advantage to using k-means clustering as opposed to hierarchical clustering. Kane replies that in his opinion, hierarchical clustering is the more appropriate algorithm.

1 Which analyst suggestion is *most likely* to be implemented using machine learning?

A Suggestion 1.

B Suggestion 2.

C Suggestion 3.

2 Kane's reply to Lund's first question is:

A correct.

B not correct, because other hyperparameters must also be specified.

- **C** not correct, because the feature set for describing the holdings measure must also be specified.

3 In stating a preference for hierarchical clustering in his reply to Lund's second question, Kane *most likely* is giving consideration to:

- **A** the speed of the algorithms.
- **B** the dimensionality of the data set.
- **C** the need to specify the hyperparameter, k, in using a k-means algorithm.

Solution to 1:

B is correct. A machine learning clustering algorithm could be used to implement Suggestion 2. A and C are incorrect because Suggestions 1 and 3, respectively, can be addressed easily using traditional regression analysis.

Solution to 2:

C is correct. Beyond specifying a distance measure and the k for k-means, whichever clustering algorithm is selected, the feature set used to group holdings by similarities must also be specified. Operating and financial characteristics of the companies represented in the Fund's portfolio are examples of such features.

Solution to 3:

C is correct. The value of the hyperparameter, k, the number of distinct groups into which the STOXX Europe 600 Index can be segmented, is not known. Using a hierarchical algorithm, the sorting of observations into clusters will occur without any prior input on the analyst's part.

NEURAL NETWORKS, DEEP LEARNING NETS, AND REINFORCEMENT LEARNING 7

The artificial intelligence revolution has been driven in large part by advances in neural networks, deep learning algorithms, and reinforcement learning. These sophisticated algorithms can address highly complex machine learning tasks, such as image classification, face recognition, speech recognition, and natural language processing. These complicated tasks are characterized by non-linearities and interactions between large numbers of feature inputs. We now provide an overview of these algorithms and their investment applications.

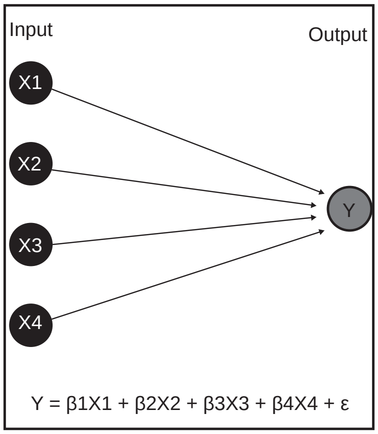
7.1 Neural Networks

Neural networks (also called *artificial neural networks*, or ANNs) are a highly flexible type of ML algorithm that have been successfully applied to a variety of tasks characterized by non-linearities and complex interactions among features. Neural networks are commonly used for classification and regression supervised learning but are also important in reinforcement learning, which can be unsupervised.

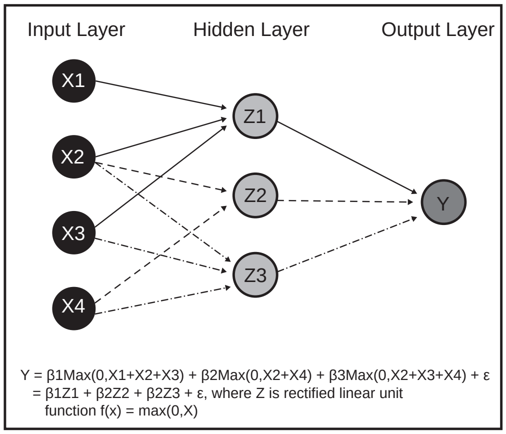
Exhibit 18 shows the connection between multiple regression and neural networks. Panel A represents a hypothetical regression for data using 4 inputs, the features x_1 to x_4, and 1 output, the predicted value of the target variable y. Panel B shows a schematic representation of a basic neural network, which consists of *nodes* (circles) connected by *links* (arrows connecting nodes). Neural networks have three types of layers: an input layer (here with a node for each of the 4 features); hidden layers, where learning occurs in training and inputs are processed on trained nets; and an output layer (here consisting of a single node for the target variable y), which passes information to outside the network. Besides the network structure, another important difference between multiple regression and neural networks is that the nodes in the neural network's hidden layer transform the inputs in a non-linear fashion into new values that are then combined into the target value. For example, in this case using the popular rectified linear unit (or ReLU) function, f(x) = max(0,x), y will be equal to β_1 times z_1, where z_1 is the maximum of $(x_1 + x_2 + x_3)$ or 0, plus β_2 times z_2, the maximum of $(x_2 + x_4)$ or 0, plus β_3 times z_3, the maximum of $(x_2 + x_3 + x_4)$ or 0, plus an error term.

Exhibit 18 Regression and Neural Networks (Regression with Transformed Features)

A. Conceptual Illustration of Regression



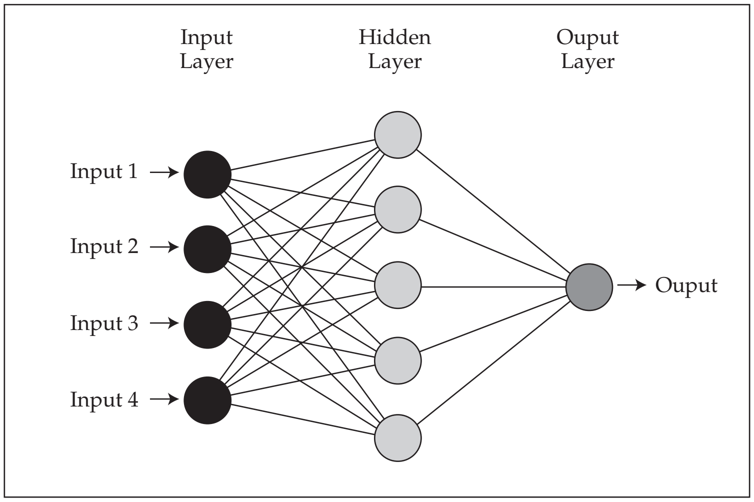
B. Conceptual Illustration of Hypothetical Neural Network



Note that for neural networks, the feature inputs would be scaled (i.e., standardized) to account for differences in the units of the data. For example, if the inputs were positive numbers, each could be scaled by its maximum value so that their values lie between 0 and 1.

Exhibit 19 shows a more complex neural network, with an input layer consisting of four nodes (i.e., four features), one hidden layer consisting of five hidden nodes, and an output node. These three numbers—4, 5, and 1—for the neural network are hyperparameters that determine the structure of the neural network.

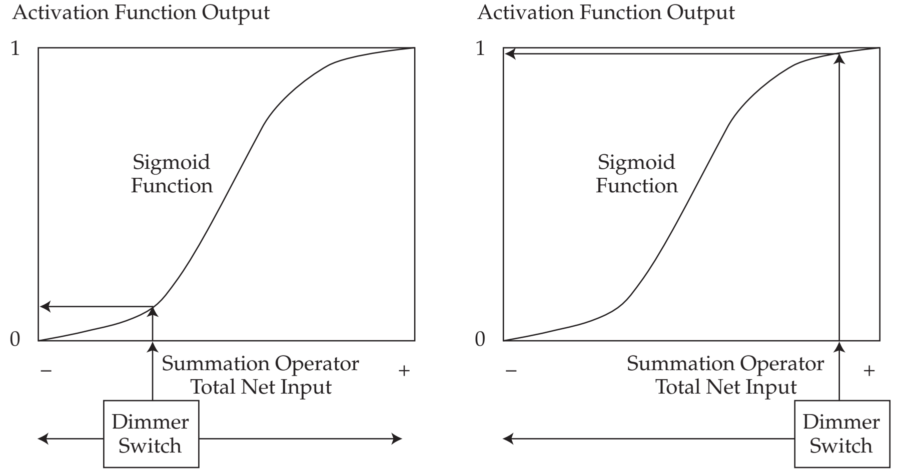
Exhibit 19 A More Complex (4-5-1) Neural Network with One Hidden Layer



Now consider any of the nodes to the right of the input layer. These nodes are sometimes called "neurons" because they process information received. Take the topmost hidden node. Four links connect to that node from the inputs, so the node gets four values transmitted by the links. Each node has, conceptually, two functional parts: a summation operator and an activation function. Once the node receives the four input values, the **summation operator** multiplies each value by a weight and sums the weighted values to form the total net input. The total net input is then passed to the **activation function**, which transforms this input into the final output of the node. Informally, the activation function operates like a light dimmer switch that decreases or increases the strength of the input. The activation function is characteristically non-linear, such as an S-shaped (sigmoidal) function (with output range of 0 to 1) or the rectified linear unit function shown in Panel B of Exhibit 18. Non-linearity implies that the rate of change of output differs at different levels of input.

This activation function is shown in Exhibit 20, where in the left graph a negative total net input is transformed via the S-shaped function into an output close to 0. This low output implies the node does not "fire," so there is nothing to pass to the next node. Conversely, in the right graph a positive total net input is transformed into an output close to 1, so the node does fire. The output of the activation function is then transmitted to the next set of nodes if there is a second hidden layer or, as in this case, to the output layer node as the predicted value. The process of transmission just described (think of forward pointing arrows in Exhibit 19) is referred to as **forward propagation**.

Exhibit 20 Activation Function as "Light Dimmer Switch" at Each Node in a Neural Network



Starting with an initialized set of random network weights, training a neural network in a supervised learning context is an iterative process in which predictions are compared to actual values of labeled data and evaluated by a specified performance measure (e.g., mean squared error). Then, network weights are adjusted to reduce total error of the network. (If the process of adjustment works backward through the layers of the network, this process is called **backward propagation**). Learning takes place through this process of adjustment to network weights with the aim of reducing total error. Without proliferating notation relating to nodes, the gist of the updating can be expressed informally as:

New weight = (Old weight) – (Learning rate) × (Partial derivative of the total error with respect to the old weight),

where "partial derivative" is a "gradient" or "rate of change of the total error with respect to the change in the old weight," and **learning rate** is a parameter that affects the magnitude of adjustments. When learning is complete, all the network weights have assigned values.

The structure of a network in which all the features are interconnected with non-linear activation functions allows neural networks to uncover and approximate complex non-linear relationships among features. Broadly speaking, when more nodes and more hidden layers are specified, a neural network's ability to handle complexity tends to increase (but so does the risk of overfitting).

Asset pricing is a noisy, stochastic process with potentially unstable relationships that challenge modelling processes, so researchers are asking if machine learning can improve our understanding of how markets work. Research comparing statistical and machine learning methods' abilities to explain and predict equity prices so far indicates that simple neural networks produce models of equity returns at the individual stock and portfolio level that are superior to models built using traditional statistical methods due to their ability to capture dynamic and interacting variables. This suggests that ML-based models, such as neural networks, may simply be better able to cope with the non-linear relationships inherent in security prices. However, the trade-offs in using them are the lack of interpretability and the amount of data needed to train such models.

7.2 Deep Learning Nets

Neural networks with many hidden layers—at least 3 but often more than 20 hidden layers—are known as **deep learning nets** (DLNs) and are the backbone of the artificial intelligence revolution. Advances in DLNs have driven developments in many complex activities, such as image, pattern, and speech recognition. To state the operation of DLNs succinctly, they take a set of inputs x from a feature set (the input layer), which are then passed to a layer of non-linear mathematical functions (neurons) with weights w_{ij} (for neuron i and input j), each of which usually produces a scaled number in the range (0, 1) or (–1, 1). These numbers are then passed to another layer of functions and into another and so on until the final layer produces a set of probabilities of the observation being in any of the target categories (each represented by a node in the output layer). The DLN assigns the category based on the category with the highest probability. The DLN is trained on large data sets; during training, the weights w_i are determined to minimize a specified loss function.

In practice, while the number of nodes in the input and the output layers are typically determined by the characteristics of the features and predicted output, many model hyperparameters still must be decided, particularly the number of hidden nodes and their connectivity and activation architecture. The objective is to choose them to achieve the best out-of-sample performance, but it is still a challenge with no simple solution. As such, a good starting point is a 'reasonable' guess for hyperparameters based on experience and literature. The researcher can then observe the result and adjust the hyperparameters incrementally until the model performance goal is reached. In practice, DLNs require substantial time to train, and systematically varying the hyperparameters may not be feasible. So, for many problems with relatively small data sets, one can start with just two or three hidden layers and a few hundred nodes before tuning the parameters until a model with acceptable predictive power is achieved.

DLNs have been shown to be useful in general for pattern recognition problems (e.g., character and image recognition), credit card fraud detection, vision and control problems in autonomous cars, natural language processing (such as machine translation), and other applications. DLNs have become hugely successful because of a confluence of three developments: 1) advances in analytical methods for fitting these models; 2) the availability of large quantities of machine readable data to train models; and 3) fast computers, especially new chips in the graphics processing unit (GPU) class, tailored for the type of calculations done on DLNs.

Several financial firms are experimenting with DLNs for trading as well as automating their internal processes. Researchers have described how DLNs were trained to price options, mimicking the Black–Scholes–Merton model. This research used the same six input parameters for the model as input layer features—spot price, strike, time to maturity, dividend yield, risk-free interest rate, and volatility—with four hidden layers of 100 neurons each and one output layer. The predicted option prices out-of-sample were very close to the actual option prices: A regression of predicted option prices on actual prices had an R^2 of 99.8%.

Some researchers have used DLNs known as multi-layer perceptrons (or feed-forward networks) with many nodes (sometimes over 1,000) split over several hidden layers to predict corporate fundamental factors and price-related technical factors. Such DLNs typically use dozens of input features with many from company financial statements and also technical indicators of stock price momentum over various prior periods (e.g., 1, 3, 6 months). Some investment strategies using the DLN-predicted factors (e.g., book value to market value, operating income to market capitalization, etc.) have achieved compounded annual returns of 270 basis points more than returns using standard factor models.

7.3 Reinforcement Learning

Reinforcement learning (RL) is an algorithm that made headlines in 2017 when DeepMind's AlphaGo program beat the reigning world champion at the ancient game of Go. The RL algorithm involves an agent that should perform actions that will maximize its rewards over time, taking into consideration the constraints of its environment. In the case of AlphaGo, a virtual gamer (the agent) uses his/her console commands (the actions) with the information on the screen (the environment) to maximize his/her score (the reward). Unlike supervised learning, reinforcement learning has neither direct labeled data for each observation nor instantaneous feedback. With RL, the algorithm needs to observe its environment, learn by testing new actions (some of which may not be immediately optimal), and reuse its previous experiences. The learning subsequently occurs through millions of trials and errors. Academics and practitioners are applying RL in a similar way in investment strategies where the agent could be a virtual trader who follows certain trading rules (the actions) in a specific market (the environment) to maximize its profits (its reward). The success of RL in dealing with the complexities of financial markets is still an open question.

EXAMPLE 6

Deep Learning Nets

Glen Mitsui is the CIO for a large Australian state's Public Employees' Pension Fund (PEPF), which currently has assets under management (AUM) of A$20 billion. The fund manages one-quarter of its assets internally, with A$5 billion mostly in domestic government and corporate fixed-income instruments and domestic equities. The remaining three-quarters of AUM, or A$15 billion, is managed by nearly 100 mostly active external asset managers and is invested in a wide range of asset classes, including foreign fixed income and equities, domestic and foreign hedge funds, REITs, commodities, and derivatives.

PEPF has a small staff of four investment professionals tasked with selecting and monitoring these external managers to whom it pays more than A$400 million in fees annually. Performance (compared to appropriate benchmarks) of many of PEPF's external managers has been lagging over the past several years. After studying the situation, Mitsui concludes that style drift may be an important factor in explaining such underperformance, for which PEPF is not happy to pay. Mitsui believes that machine learning may help and consults with Frank Monroe, professor of data analysis at Epsilon University.

Monroe suggests using a deep learning net model that collects and analyzes the real-time trading data of PEPF's external managers and compares them to well-known investment styles (e.g., high dividend, minimum volatility, momentum, growth, value) to detect potential style drift. Mitsui arranges for Monroe to meet with PEPF's investment committee (IC) to discuss the matter. As a junior data analyst working with Monroe, you must help him answer the following requests from the IC:

1 Define a deep learning net.
2 Evaluate Monroe's opinion on the applicability of deep learning nets to Mitsui's problem.
3 Describe the functions of the three groups of layers of a deep learning net.

Solution to 1:

A deep learning net (DLN) is a neural network (NN) with many hidden layers (at least 3 but often more than 20). NNs and DLNs have been successfully applied to a wide variety of complex tasks characterized by non-linearities and interactions among features, particularly pattern recognition problems.

Solution to 2:

Mitsui wants to detect patterns of potential style drift in the daily trading of nearly 100 external asset managers in many markets. This task will involve the fast processing of huge amounts of complicated data. Monroe is correct that a DLN is well suited to PEPF's needs.

Solution to 3:

The input layer, the hidden layers, and the output layer constitute the three groups of layers of DLNs. The input layer receives the inputs (i.e., features) and has as many nodes as there are dimensions of the feature set. The hidden layers consist of nodes, each comprised of a summation operator and an activation function that are connected by links. These hidden layers are, in effect, where the model is learned. The final layer, the output layer, produces a set of probabilities of an observation being in any of the target style categories (each represented by a node in the output layer). The DLN assigns the category based on the style category with the highest probability.

EXAMPLE 7

Summing Up the Major Types of Machine Learning

1 As used in supervised machine learning, regression problems involve:
   - A binary target variables.
   - B continuous target variables.
   - C categorical target variables.

2 Which of the following *best* describes penalized regression? Penalized regression:
   - A is unrelated to multiple linear regression.
   - B involves a penalty term that is added to the predicted target variable.
   - C is a category of general linear models used when the number of features and overfitting are concerns.

3 CART is *best* described as:
   - A an unsupervised ML algorithm.
   - B a clustering algorithm based on decision trees.
   - C a supervised ML algorithm that accounts for non-linear relationships among the features.

4 A neural network is *best* described as a technique for machine learning that is:
   - A exactly modeled on the human nervous system.
   - B based on layers of nodes connected by links when the relationships among the features are usually non-linear.

C based on a tree structure of nodes when the relationships among the features are linear.

5 Hierarchical clustering is *best* described as a technique in which:

A the grouping of observations is unsupervised.

B features are grouped into a pre-specified number, *k*, of clusters.

C observations are classified according to predetermined labels.

6 Dimension reduction techniques are *best* described as a means to reduce a set of features:

A to a manageable size without regard for the variation in the data.

B to a manageable size while increasing the variation in the data.

C to a manageable size while retaining as much of the variation in the data as possible.

Solution to 1:

B is correct. A and C are incorrect because when the target variable is binary or categorical, the problem is a *classification problem* rather than a regression problem.

Solution to 2:

C is correct. A is incorrect because penalized regression is related to multiple linear regression. B is incorrect because penalized regression involves adding a penalty term to the sum of the squared regression residuals.

Solution to 3:

C is correct. A is incorrect because CART is a supervised ML algorithm. B is incorrect because CART is a classification and regression algorithm, not a clustering algorithm.

Solution to 4:

B is correct. A is incorrect because neural networks are not exactly modeled on the human nervous system. C is incorrect because neural networks are not based on a tree structure of nodes when the relationships among the features are linear.

Solution to 5:

A is correct. B is incorrect because it refers to k-means clustering. C is incorrect because it refers to classification, which involves supervised learning.

Solution to 6:

C is correct because dimension reduction techniques, like PCA, are aimed at reducing the feature set to a manageable size while retaining as much of the variation in the data as possible.

SUMMARY

Machine learning methods are gaining usage at many stages in the investment management value chain. Among the major points made are the following:

- Machine learning aims at extracting knowledge from large amounts of data by learning from known examples to determine an underlying structure in the data. The emphasis is on generating structure or predictions without human intervention. An elementary way to think of ML algorithms is to "find the pattern, apply the pattern."
- Supervised learning depends on having labeled training data as well as matched sets of observed inputs (X's, or features) and the associated output (Y, or target). It can be divided into two categories: regression and classification. If the target variable to be predicted is continuous, then the task is one of regression. If the target variable is categorical or ordinal (e.g., determining a firm's rating), then it is a classification problem.
- With unsupervised learning, algorithms are trained with no labeled data, so they must infer relations between features, summarize them, or present an interesting underlying structure in their distributions that has not been explicitly provided. Two important types of problems well suited to unsupervised ML are dimension reduction and clustering.
- Another category of ML algorithm includes deep learning (based on neural networks) in which a computer learns from interacting with itself. Sophisticated algorithms address such highly complex tasks as image classification, face recognition, speech recognition and natural language processing, and reinforcement learning.
- Generalization describes the degree to which an ML model retains its explanatory power when predicting out-of-sample. Overfitting, a primary reason for lack of generalization, is the tendency of ML algorithms to tailor models to the training data at the expense of generalization to new data points.
- Bias error is the degree to which a model fits the training data. Variance error describes how much a model's results change in response to new data from validation and test samples. Base error is due to randomness in the data. Out-of-sample error equals bias error plus variance error plus base error.
- K-fold cross-validation is a technique for mitigating the holdout sample problem (excessive reduction of the training set size). The data (excluding test sample and fresh data) are shuffled randomly and then divided into k equal sub-samples, with $k - 1$ samples used as training samples and one sample, the kth, used as a validation sample.
- Regularization describes methods that reduce statistical variability in high dimensional data estimation or prediction problems.
- LASSO (least absolute shrinkage and selection operator) is a popular type of penalized regression where the penalty term involves summing the absolute values of the regression coefficients. The greater the number of included features, the larger the penalty. So, a feature must make a sufficient contribution to model fit to offset the penalty from including it.
- Support vector machine (SVM) is a linear classifier that aims to seek the optimal hyperplane—the one that separates the two sets of data points by the maximum margin (and thus is typically used for classification).

- K-nearest neighbor (KNN) is a supervised learning technique most often used for classification. The idea is to classify a new observation by finding similarities ("nearness") between it and its k-nearest neighbors in the existing data set.
- Classification and regression tree (CART) can be applied to predict either a categorical target variable, producing a classification tree, or a continuous target variable, producing a regression tree.
- A binary CART is a combination of an initial root node, decision nodes, and terminal nodes. The root node and each decision node represent a single feature (*f*) and a cutoff value (*c*) for that feature. The CART algorithm iteratively partitions the data into sub-groups until terminal nodes are formed that contain the predicted label.
- Ensemble learning is a technique of combining the predictions from a collection of models. It typically produces more accurate and more stable predictions than the best single model.
- A random forest classifier is a collection of many different decision trees generated by a bagging method or by randomly reducing the number of features available during training.
- Principal components analysis (PCA) is an unsupervised ML algorithm that reduces highly correlated features into fewer uncorrelated composite variables by transforming the feature covariance matrix. PCA produces eigenvectors that define the principal components (i.e., the new uncorrelated composite variables) and eigenvalues, which give the proportion of total variance in the initial data that is explained by each eigenvector and its associated principal component.
- K-means is an unsupervised ML algorithm that partitions observations into a fixed number (*k*) of non-overlapping clusters. Each cluster is characterized by its centroid, and each observation belongs to the cluster with the centroid to which that observation is closest.
- Hierarchical clustering is an unsupervised iterative algorithm that is used to build a hierarchy of clusters. Two main strategies are used to define the intermediary clusters (i.e., those clusters between the initial data set and the final set of clustered data).
- Agglomerative (bottom-up) hierarchical clustering begins with each observation being its own cluster. Then, the algorithm finds the two closest clusters, defined by some measure of distance, and combines them into a new, larger cluster. This process is repeated until all observations are clumped into a single cluster.
- Divisive (top-down) hierarchical clustering starts with all observations belonging to a single cluster. The observations are then divided into two clusters based on some measure of distance. The algorithm then progressively partitions the intermediate clusters into smaller clusters until each cluster contains only one observation.
- Neural networks consist of nodes connected by links. They have three types of layers: an input layer, hidden layers, and an output layer. Learning takes place in the hidden layer nodes, each of which consists of a summation operator and an activation function. Neural networks have been successfully applied to a variety of investment tasks characterized by non-linearities and complex interactions among variables.

- Neural networks with many hidden layers (at least 3 but often more than 20) are known as deep learning nets (DLNs) and are the backbone of the artificial intelligence revolution.
- The RL algorithm involves an agent that should perform actions that will maximize its rewards over time, taking into consideration the constraints of its environment.

REFERENCES

Bacham, Dinesh, and Janet Zhao. 2017. "Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling." *Moody's Analytics Risk Perspectives* 9:28–35.

PRACTICE PROBLEMS

The following information relates to Questions 1–10

Alef Associates manages a long-only fund specializing in global smallcap equities. Since its founding a decade ago, Alef maintains a portfolio of 100 stocks (out of an eligible universe of about 10,000 stocks). Some of these holdings are the result of screening the universe for attractive stocks based on several ratios that use readily available market and accounting data; others are the result of investment ideas generated by Alef's professional staff of five securities analysts and two portfolio managers.

Although Alef's investment performance has been good, its Chief Investment Officer, Paul Moresanu, is contemplating a change in the investment process aimed at achieving even better returns. After attending multiple workshops and being approached by data vendors, Moresanu feels that data science should play a role in the way Alef selects its investments. He has also noticed that much of Alef's past outperformance is due to stocks that became takeover targets. After some research and reflection, Moresanu writes the following email to the Alef's CEO.

Subject: Investment Process Reorganization

I have been thinking about modernizing the way we select stock investments. Given that our past success has put Alef Associates in an excellent financial position, now seems to be a good time to invest in our future. What I propose is that we continue managing a portfolio of 100 global small-cap stocks but restructure our process to benefit from machine learning (ML). Importantly, the new process will still allow a role for human insight, for example, in providing domain knowledge. In addition, I think we should make a special effort to identify companies that are likely to be acquired. Specifically, I suggest following the four steps which would be repeated every quarter.

Step 1 We apply ML techniques to a model including fundamental and technical variables (features) to predict next quarter's return for each of the 100 stocks currently in our portfolio. Then, the 20 stocks with the lowest estimated return are identified for replacement.

Step 2 We utilize ML techniques to divide our investable universe of about 10,000 stocks into 20 different groups, based on a wide variety of the most relevant financial and non-financial characteristics. The idea is to prevent unintended portfolio concentration by selecting stocks from each of these distinct groups.

Step 3 For each of the 20 different groups, we use labeled data to train a model that will predict the five stocks (in any given group) that are most likely to become acquisition targets in the next one year.

(Continued)

Step 4 Our five experienced securities analysts are each assigned four of the groups, and then each analyst selects their one best stock pick from each of their assigned groups. These 20 "high-conviction" stocks will be added to our portfolio (in replacement of the 20 relatively underperforming stocks to be sold in Step 1).

A couple of additional comments related to the above:

Comment 1 The ML algorithms will require large amounts of data. We would first need to explore using free or inexpensive historical datasets and then evaluate their usefulness for the ML-based stock selection processes before deciding on using data that requires subscription.

Comment 2 As time passes, we expect to find additional ways to apply ML techniques to refine Alef's investment processes.

What do you think?
Paul Moresanu

---

1 The machine learning techniques appropriate for executing Step 1 are *most* likely to be based on:

A regression

B classification

C clustering

2 Assuming regularization is utilized in the machine learning technique used for executing Step 1, which of the following ML models would be *least* appropriate:

A Regression tree with pruning.

B LASSO with lambda (λ) equal to 0.

C LASSO with lambda (λ) between 0.5 and 1.

3 Which of the following machine learning techniques is *most* appropriate for executing Step 2:

A K-Means Clustering

B Principal Components Analysis (PCA)

C Classification and Regression Trees (CART)

4 The hyperparameter in the ML model to be used for accomplishing Step 2 is?

A 100, the number of small-cap stocks in Alef's portfolio.

B 10,000, the eligible universe of small-cap stocks in which Alef can potentially invest.

C 20, the number of different groups (i.e. clusters) into which the eligible universe of small-cap stocks will be divided.

5 The target variable for the labelled training data to be used in Step 3 is *most* likely which one of the following?

A A continuous target variable.

B A categorical target variable.

C An ordinal target variable.

6 Comparing two ML models that could be used to accomplish Step 3, which statement(s) *best* describe(s) the advantages of using Classification and Regression Trees (CART) instead of K-Nearest Neighbor (KNN)?

Statement I	For CART there is no requirement to specify an initial hyperparameter (like K).
Statement II	For CART there is no requirement to specify a similarity (or distance) measure.
Statement III	For CART the output provides a visual explanation for the prediction.

A Statement I only.

B Statement III only.

C Statements I, II and III.

7 Assuming a Classification and Regression Tree (CART) model is used to accomplish Step 3, which of the following is *most* likely to result in model overfitting?

A Using the k-fold cross validation method

B Including an overfitting penalty (i.e., regularization term).

C Using a fitting curve to select a model with low bias error and high variance error.

8 Assuming a Classification and Regression Tree (CART) model is initially used to accomplish Step 3, as a further step which of the following techniques is most likely to result in more accurate predictions?

A Discarding CART and using the predictions of a Support Vector Machine (SVM) model instead.

B Discarding CART and using the predictions of a K-Nearest Neighbor (KNN) model instead.

C Combining the predictions of the CART model with the predictions of other models – such as logistic regression, SVM, and KNN – via ensemble learning.

9 Regarding Comment #2, Moresanu has been thinking about the applications of neural networks (NNs) and deep learning (DL) to investment management. Which statement(s) *best* describe(s) the tasks for which NNs and DL are well-suited?

Statement I	NNs and DL are well-suited for image and speech recognition, and natural language processing.
Statement II	NNs and DL are well-suited for developing single variable ordinary least squares regression models.
Statement III	NNs and DL are well-suited for modelling non-linearities and complex interactions among many features.

A Statement II only.

B Statements I and III.

C Statements I, II and III.

10 Regarding neural networks (NNs) that Alef might potentially implement, which of the following statements is *least* accurate?

A NNs must have at least 10 hidden layers to be considered deep learning nets.

B The activation function in a node operates like a light dimmer switch since it decreases or increases the strength of the total net input.

C The summation operator receives input values, multiplies each by a weight, sums up the weighted values into the total net input, and passes it to the activation function.

SOLUTIONS

1 A is correct. The target variable (quarterly return) is continuous, hence this calls for a supervised machine learning based regression model.

B is incorrect, since classification uses categorical or ordinal target variables, while in Step 1 the target variable (quarterly return) is continuous.

C is incorrect, since clustering involves unsupervised machine learning so does not have a target variable.

2 B is correct. It is least appropriate because with LASSO, when $\lambda = 0$ the penalty (i.e., regularization) term reduces to zero, so there is no regularization and the regression is equivalent to an ordinary least squares (OLS) regression.

A is incorrect. With Classification and Regression Trees (CART), one way that regularization can be implemented is via pruning which will reduce the size of the regression tree—sections that provide little explanatory power are pruned (i.e., removed).

C is incorrect. With LASSO, when λ is between 0.5 and 1 the relatively large penalty (i.e., regularization) term requires that a feature makes a sufficient contribution to model fit to offset the penalty from including it in the model.

3 A is correct. K-Means clustering is an unsupervised machine learning algorithm which repeatedly partitions observations into a fixed number, *k*, of non-overlapping clusters (i.e., groups).

B is incorrect. Principal Components Analysis is a long-established statistical method for dimension reduction, not clustering. PCA aims to summarize or reduce highly correlated features of data into a few main, uncorrelated composite variables.

C is incorrect. CART is a supervised machine learning technique that is most commonly applied to binary classification or regression.

4 C is correct. Here, 20 is a hyperparameter (in the K-Means algorithm), which is a parameter whose value must be set by the researcher before learning begins.

A is incorrect, because it is not a hyperparameter. It is just the size (number of stocks) of Alef's portfolio.

B is incorrect, because it is not a hyperparameter. It is just the size (number of stocks) of Alef's eligible universe.

5 B is correct. To predict which stocks are likely to become acquisition targets, the ML model would need to be trained on categorical labelled data having the following two categories: "0" for "not acquisition target", and "1" for "acquisition target".

A is incorrect, because the target variable is categorical, not continuous.

C is incorrect, because the target variable is categorical, not ordinal (i.e., 1st, 2nd, 3rd, etc.).

6 C is correct. The advantages of using CART over KNN to classify companies into two categories ("not acquisition target" and "acquisition target"), include all of the following: For CART there are no requirements to specify an initial hyperparameter (like K) or a similarity (or distance) measure as with KNN, and CART provides a visual explanation for the prediction (i.e., the feature variables and their cut-off values at each node).

A is incorrect, because CART provides all of the advantages indicated in Statements I, II and III.

B is incorrect, because CART provides all of the advantages indicated in Statements I, II and III.

7 C is correct. A fitting curve shows the trade-off between bias error and variance error for various potential models. A model with low bias error and high variance error is, by definition, overfitted.

A is incorrect, because there are two common methods to reduce overfitting, one of which is proper data sampling and cross-validation. K-fold cross validation is such a method for estimating out-of-sample error directly by determining the error in validation samples.

B is incorrect, because there are two common methods to reduce overfitting, one of which is preventing the algorithm from getting too complex during selection and training, which requires estimating an overfitting penalty.

8 C is correct. Ensemble learning is the technique of combining the predictions from a collection of models, and it typically produces more accurate and more stable predictions than the best single model.

A is incorrect, because a single model will have a certain error rate and will make noisy predictions. By taking the average result of many predictions from many models (i.e., ensemble learning) one can expect to achieve a reduction in noise as the average result converges towards a more accurate prediction.

B is incorrect, because a single model will have a certain error rate and will make noisy predictions. By taking the average result of many predictions from many models (i.e., ensemble learning) one can expect to achieve a reduction in noise as the average result converges towards a more accurate prediction.

9 B is correct. NNs and DL are well-suited for addressing highly complex machine learning tasks, such as image classification, face recognition, speech recognition and natural language processing. These complicated tasks are characterized by non-linearities and complex interactions between large numbers of feature inputs.

A is incorrect, because NNs and DL are well-suited for addressing highly complex machine learning tasks, not simple single variable OLS regression models.

C is incorrect, because NNs and DL are well-suited for addressing highly complex machine learning tasks, not simple single variable OLS regression models.

10 A is correct. It is the least accurate answer because neural networks with many hidden layers—at least 3, but often more than 20 hidden layers—are known as deep learning nets.

B is incorrect, because the node's activation function operates like a light dimmer switch which decreases or increases the strength of the (total net) input.

C is incorrect, because the node's summation operator multiplies each (input) value by a weight and sums up the weighted values to form the total net input. The total net input is then passed to the activation function.

READING

8

Big Data Projects

by Sreekanth Mallikarjun, PhD, and Ahmed Abbasi, PhD

Sreekanth Mallikarjun, PhD, is at Reorg (USA) and the University of Virginia, McIntire School of Commerce (USA). Ahmed Abbasi, PhD, is at the University of Virginia, McIntire School of Commerce (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	**a.** state and explain steps in a data analysis project;
☐	**b.** describe objectives, steps, and examples of preparing and wrangling data;
☐	**c.** describe objectives, methods, and examples of data exploration;
☐	**d.** describe objectives, steps, and techniques in model training;
☐	**e.** describe preparing, wrangling, and exploring text-based data for financial forecasting;
☐	**f.** describe methods for extracting, selecting and engineering features from textual data;
☐	**g.** evaluate the fit of a machine learning algorithm.

INTRODUCTION

1

Big data (also referred to as alternative data) encompasses data generated by financial markets (e.g., stock and bond prices), businesses (e.g., company financials, production volumes), governments (e.g., economic and trade data), individuals (e.g., credit card purchases, social media posts), sensors (e.g., satellite imagery, traffic patterns), and the Internet of Things, or IoT, (i.e., the network of interrelated digital devices that can transfer data among themselves without human interaction). A veritable explosion in big data has occurred over the past decade or so, especially in unstructured data generated from social media (e.g., posts, tweets, blogs), email and text communications, web traffic, online news sites, electronic images, and other electronic information sources. The prospects are for exponential growth in big data to continue.

Investment managers are increasingly using big data in their investment processes as they strive to discover signals embedded in such data that can provide them with an information edge. They seek to augment structured data with a plethora of unstructured data to develop improved forecasts of trends in asset prices, detect anomalies, etc. A typical example involves a fund manager using financial text data from 10-K reports for forecasting stock sentiment (i.e., positive or negative), which can then be used as an input to a more comprehensive forecasting model that includes corporate financial data.

Unlike structured data (numbers and values) that can be readily organized into data tables to be read and analyzed by computers, unstructured data typically require specific methods of preparation and refinement before being usable by machines (i.e., computers) and useful to investment professionals. Given the volume, variety, and velocity of available big data, it is important for portfolio managers and investment analysts to have a basic understanding of how unstructured data can be transformed into structured data suitable as inputs to machine learning (ML) methods (in fact, for any type of modeling methods) that can potentially improve their financial forecasts.

This reading describes the steps in using big data, both structured and unstructured, in financial forecasting. The concepts and methods are then demonstrated in a case study of an actual big data project. The project uses text-based data derived from financial documents to train an ML model to classify text into positive or negative sentiment classes for the respective stocks and then to predict sentiment.

Section 2 of the reading covers a description of the key characteristics of big data. Section 3 provides an overview of the steps in executing a financial forecasting project using big data. We then describe in Sections 4–6 key aspects of data preparation and wrangling, data exploration, and model training using structured data and unstructured (textual) data. In Section 7, we bring these pieces together by covering the execution of an actual big data project. A summary in Section 8 concludes the reading.

2 BIG DATA IN INVESTMENT MANAGEMENT

Big data differs from traditional data sources based on the presence of a set of characteristics commonly referred to as the 3Vs: volume, variety, and velocity.

Volume refers to the quantity of data. The US Library of Congress, which is tasked with archiving both digital and physical information artifacts in the United States, has collected hundreds of terabytes of data (one terabyte equals 1,024 gigabytes, which are equal to 1,048,576 megabytes). Several years ago, one of the authors managed an archival project for the Library of Congress in which many terabytes of online content were collected—a copious amount of data at the time. However, in most US industry sectors today, the average company collects more data than the Library of Congress! In big data conversations, terabytes have been replaced with petabytes and exabytes (one exabyte equals 1,024 petabytes, which are equal to 1,048,576 terabytes). The classic grains of sand analogy puts these volumes into perspective: If a megabyte is a tablespoon of sand, then a petabyte is a 1.6-kilometer-long beach and an exabyte is a beach extending about 1,600 kilometers.

Variety pertains to the array of available data sources. Organizations are now dealing with structured, semi-structured, and unstructured data from within and outside the enterprise. Variety includes traditional transactional data; user-generated text, images, and videos; social media; sensor-based data; web and mobile clickstreams; and spatial-temporal data. Effectively leveraging the variety of available data presents both opportunities and challenges, including such legal and ethical issues as data privacy.

Velocity is the speed at which data are created. Many large organizations collect several petabytes of data every hour. With respect to unstructured data, more than one billion new tweets (i.e., a message of 280 characters or less posted on the social media website Twitter) are generated every three days; five billion search queries occur daily. Such information has important implications for real-time predictive analytics in various financial applications. Analyzing such "data-in-motion" poses challenges since relevant patterns and insights might be moving targets relative to situations of "data-at-rest."

When using big data for inference or prediction, there is a "fourth V": *Veracity relates to the credibility and reliability of different data sources.* Determining the credibility and reliability of data sources is an important part of any empirical investigation. The issue of veracity becomes critically important for big data, however, because of the varied sources of these large datasets. Big data amplifies the age-old challenge of disentangling quality from quantity. Social media, including blogs, forums, and social networking sites, are plagued with spam; by some estimates, as much as 10%–15% of such content is completely fake. Similarly, according to our research, web spam accounts for more than 20% of all content on the worldwide web. Clickstreams from website and mobile traffic are equally susceptible to noise. Furthermore, deriving deep semantic knowledge from text remains challenging in certain instances despite significant advances in natural language processing (NLP).

These Vs have numerous implications for financial technology (commonly referred to as "fintech") pertaining to investment management. Machine learning assessments of creditworthiness, which have traditionally relied on structured financial metrics, are being enhanced by incorporating text derived from financial statements, news articles, and call transcripts. Customers in the financial industry are being segmented based not only on their transactional data but also on their views and preferences expressed on social media (to the degree permissible under applicable privacy agreements). Big data also affords opportunities for enhanced fraud detection and risk management.

3 STEPS IN EXECUTING A DATA ANALYSIS PROJECT: FINANCIAL FORECASTING WITH BIG DATA

In the era of big data, firms treat data like they do important assets. However, effective big data analytics are critical to allow appropriate data monetization. Let us take financial forecasting as an application area. Numerous forecasting tasks in this domain can benefit from predictive analytics models built using machine learning methods. One common example is predicting whether stock prices (for an individual stock or a portfolio) will go up or down in value at some specific point in the future. Traditionally, financial forecasting relied on various financial and accounting numbers, ratios, and metrics coupled with statistical or mathematical models. More recently, machine learning models have been commonly utilized. However, with the proliferation of textual big data (e.g., online news articles, internet financial forums, social networking platforms), such unstructured data have been shown to offer insights faster (as they are real-time) and have enhanced predictive power.

Textual big data provides several valuable types of information, including topics and sentiment. Topics are what people are talking about (e.g., a firm, an industry, a particular event). Sentiment is how people feel about what they are discussing. For instance, they might express positive, negative, or neutral views (i.e., sentiments) toward a topic of discussion. One study conducted in the United States found that positive sentiment on Twitter could predict the trend for the Dow Jones Industrial Average up to three days later with nearly 87% accuracy.

Deriving such insights requires supplementing traditional data with textual big data. As depicted in Exhibit 1, the inclusion of big data has immediate implications for building the machine learning model as well as downstream implications for financial forecasting and analysis. We begin with the top half of Exhibit 1, which shows the traditional (i.e., with structured data) *ML Model Building Steps:*

1 *Conceptualization of the modeling task*. This crucial first step entails determining what the output of the model should be (e.g., whether the price of a stock will go up/down one week from now), how this model will be used and by whom, and how it will be embedded in existing or new business processes.

2 *Data collection*. The data traditionally used for financial forecasting tasks are mostly numeric data derived from internal and external sources. Such data are typically already in a structured tabular format, with columns of features, rows of instances, and each cell representing a particular value.

3 *Data preparation and wrangling*. This step involves cleansing and preprocessing of the raw data. Cleansing may entail resolving missing values, out-of-range values, and the like. Preprocessing may involve extracting, aggregating, filtering, and selecting relevant data columns.

4 *Data exploration*. This step encompasses exploratory data analysis, feature selection, and feature engineering.

5 *Model training*. This step involves selecting the appropriate ML method (or methods), evaluating performance of the trained model, and tuning the model accordingly.

Note that these steps are iterative because model building is an iterative process. The insights gained from one iteration may inform the next iteration, beginning with reconceptualization. In contrast with structured data sources, textual big data originating in online news articles, social media, internal/external documents (such as public financial statements), and other openly available data sources are unstructured.

The *Text ML Model Building Steps* used for the unstructured data sources of big data are shown in the bottom half of Exhibit 1. They differ from those used for traditional data sources and are typically intended to create output information that is structured. The differences in steps between the text model and traditional model account for the characteristics of big data: volume, velocity, variety, and veracity. In this reading, we mostly focus on the variety and veracity dimensions of big data as they manifest themselves in text. The major differences in the *Text ML Model Building Steps* are in the first four steps:

1 *Text problem formulation*. Analysts begin by determining how to formulate the text classification problem, identifying the exact inputs and outputs for the model. Perhaps we are interested in computing sentiment scores (structured output) from text (unstructured input). Analysts must also decide how the text ML model's classification output will be utilized.

2 *Data (text) curation*. This step involves gathering relevant external text data via web services or **web spidering (scraping or crawling) programs** that extract raw content from a source, typically web pages. Annotation of the text data with high-quality, reliable target (dependent) variable labels might also be necessary for supervised learning and performance evaluation purposes. For instance, experts might need to label whether a given expert assessment of a stock is bearish or bullish.

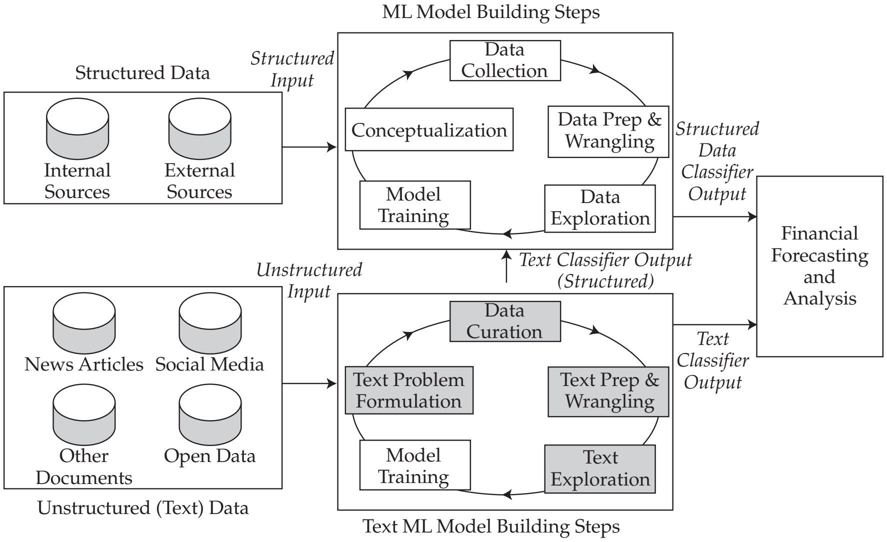
3 *Text preparation and wrangling.* This step involves critical cleansing and preprocessing tasks necessary to convert streams of unstructured data into a format that is usable by traditional modeling methods designed for structured inputs.

4 *Text exploration.* This step encompasses text visualization through techniques, such as word clouds, and text feature selection and engineering.

The resulting output (e.g., sentiment prediction scores) can either be combined with other structured variables or used directly for forecasting and/or analysis.

Next, we describe two key steps from the *ML Model Building Steps* depicted in Exhibit 1 that typically differ for structured data versus textual big data: data/text preparation and wrangling and data/text exploration. We then discuss model training. Finally, we focus on applying these steps to a case study related to classifying and predicting stock sentiment from financial texts.

Exhibit 1 Model Building for Financial Forecasting Using Big Data: Structured (Traditional) vs. Unstructured (Text)



EXAMPLE 1

Steps in ML Model Building

LendALot Corporation is a B2C (business-to-consumer) lender that has traditionally outsourced potential customers' creditworthiness scoring to a third-party firm. Given the recent advances in machine learning (ML)-based "fintech" that goes beyond traditional "repayment history" and "ability to repay" assessments derived from structured data, LendALot would like to develop in-house, ML-based credit scoring capabilities to enhance borrower risk assessment and differentiate itself in the B2C lending market. LendALot would like to follow a phased approach beginning with traditional (structured) data sources and then eventually incorporating textual (unstructured) big data sources. Paul Wang has been asked to lead a new analytics team at LendALot tasked with developing the ML-based creditworthiness scoring model. In the context of machine learning using structured data sources, address the following questions.

1 State and explain one decision Wang will need to make related to:

A conceptualizing the modeling task.

B data collection.

C data preparation and wrangling.

D data exploration.

E model training.

In a later phase of the project, LendALot attempts to improve its credit scoring processes by incorporating textual data in credit scoring. Wang tells his team, "Enhance the creditworthiness scoring model by incorporating insights from text provided by the prospective borrowers in the loan application free response fields."

2 Identify the process step that Wang's statement addresses.

3 State two potential needs of the LendAlot team in relation to text curation.

4 State two potential needs of the LendAlot team in relation to text preparation and wrangling.

Solution to 1:

A In the conceptualization step, Wang will need to decide how the output of the ML model will be specified (e.g., a binary classification of creditworthiness), how the model will be used and by whom, and how it will be embedded in LendALot's business processes.

B In the data collection phase, Wang must decide on what data—internal, external, or both—to use for credit scoring.

C In the data preparation and wrangling step, Wang will need to decide on data cleansing and preprocessing needs. Cleansing may entail resolving missing values, extreme values, etc. Preprocessing may involve extracting, aggregating, filtering, and selecting relevant data columns.

D In the data exploration phase, Wang will need to decide which exploratory data analysis methods are appropriate, which features to use in building a credit scoring model, and which features may need to be engineered.

E In the model training step, Wang must decide which ML algorithm(s) to use. Assuming labeled training data are available, the choice will be among supervised learning algorithms. Decisions will need to be made on how model fit is measured and how the model is validated and tuned.

Solution to 2:

Wang's statement relates to the initial step of text problem formulation.

Solution to 3:

Related to text curation, the team will be using internal data (from loan applications). They will need to ensure that the text comment fields on the loan applications have been correctly implemented and enabled. If these fields are not required, they need to ensure there is a sufficient response rate to analyze.

Solution to 4:

Related to text preparation and wrangling, the team will need to carry out the critical tasks of text cleansing and text preprocessing. These two tasks are necessary to convert an unstructured stream of data into structured values for use by traditional modeling methods.

DATA PREPARATION AND WRANGLING

4

Data preparation and wrangling involve cleansing and organizing raw data into a consolidated format. The resulting dataset is suitable to use for further analyses and training a machine learning (ML) model. This is a critical stage, the foundation, in big data projects. Most of the project time is spent on this step, and the quality of the data affects the training of the selected ML model. Domain knowledge—that is, the involvement of specialists in the particular field in which the data are obtained and used—is beneficial and often necessary to successfully execute this step. Data preparation is preceded by data collection, so we discuss the data collection process first.

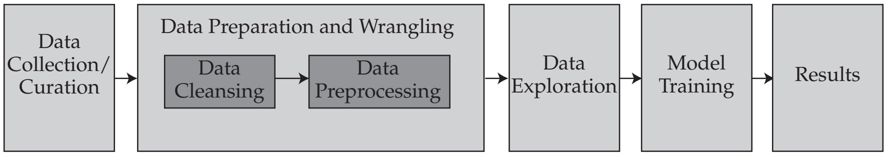
Before the data collection process even begins, it is important to state the problem, define objectives, identify useful data points, and conceptualize the model. Conceptualization is like a blueprint on a drawing board, a modifiable plan that is necessary to initiate the model building process. A project overview is established by determining the ML model type—supervised or unsupervised—and data sources/collection plans with respect to the needs of the project.

Data collection involves searching for and downloading the raw data from one or multiple sources. Data can be stored in different formats, sources, and locations. As databases are the most common primary sources, building necessary queries with the help of database administrators is critical. Database schemas are built with certain assumptions and exceptions, and it is safest to clarify the database architecture with an administrator or database architect before downloading the necessary data. Data also exist in the form of spreadsheets, comma-separated values (csv) files, text files, and other formats. Care must be taken before using such data, and documentation (often referred to as "Readme" files) must be referred to, if available. **Readme files** are text files provided with the raw data that contain information related to a data file. They are useful for understanding the data and how they can be interpreted correctly.

Alternatively, third-party data vendors can be sources of clean data. External data usually can be accessed through an **application programming interface (API)**—a set of well-defined methods of communication between various software components—or the vendors can deliver the required data in the form of csv files or other formats (as previously mentioned). Using external data can save time and resources that would otherwise go into data preparation and wrangling; however, vendor contracts come with a price. Depending on the big data project constraints, a decision must be made regarding the use of internal or external data based on the trade-offs between time, financial costs, and accuracy. For projects using internal user data, external data might not be suitable. For example, to understand user traffic on a company website, internally recorded site visits and click frequency may be captured and stored in the internal databases. External data are advantageous when a project requires generic data, such as demographics of a geographic area or traffic data of a public service. Another consideration in using external vendor provided data is that during the cleansing process, underlying trends in the data that are important for particular end-uses may be masked or even lost. This is where "alpha" is often found; so by simply buying a dataset from a vendor, you may lose your information edge. Of course, application of the data (e.g., merging and combining, putting through different types of models) will be different for everyone who uses it; there are always different ways to extract value.

Once the data are collected, the data preparation and wrangling stage begins. This stage involves two important tasks: cleansing and preprocessing, respectively. Exhibit 2 outlines data preparation and wrangling and defines the two component tasks. These tasks are explained in detail under the structured and unstructured subsections because the steps vary by the nature of data.

Exhibit 2 Data Preparation and Wrangling Stage



Data Preparation (Cleansing): This is the initial and most common task in data preparation that is performed on raw data. Data cleansing is the process of examining, identifying, and mitigating errors in raw data. Normally, the raw data are neither sufficiently complete nor sufficiently clean to directly train the ML model. Manually entered data can have incomplete, duplicated, erroneous, or inaccurate values. Automated data (recorded by systems) can have similar problems due to server failures and software bugs.

Data Wrangling (Preprocessing): This task performs transformations and critical processing steps on the cleansed data to make the data ready for ML model training. Raw data most commonly are not present in the appropriate format for model consumption. After the cleansing step, data need to be processed by dealing with outliers, extracting useful variables from existing data points, and scaling the data.

4.1 Structured Data

Data Preparation (Cleansing)

Structured data are organized in a systematic format that is readily searchable and readable by computer operations for processing and analyzing. In structured data, data errors can be in the form of incomplete, invalid, inaccurate, inconsistent, non-uniform, and duplicate data observations. The data cleansing process mainly deals with identifying and mitigating all such errors. Exhibit 3 shows a raw dataset before cleansing. The data have been collected from different sources and are organized in a data matrix (or data table) format. Each row contains observations of each customer of a US-based bank. Each column represents a variable (or feature) corresponding to each customer.

Exhibit 3 Raw Data Before Cleansing

1	ID	Name	Gender	Date of Birth	Salary	Other Income	State	Credit Card
2	1	Mr. ABC	M	12/5/1970	$50,200	$5,000	VA	Y
3	2	Ms. XYZ	M	15 Jan, 1975	$60,500	$0	NY	Yes
4	3	EFG		1/13/1979	$65,000	$1,000	CA	No
5	4	Ms. MNO	F	1/1/1900	—	—	FL	Don't Know

Exhibit 3 (Continued)

1	ID	Name	Gender	Date of Birth	Salary	Other Income	State	Credit Card
6	5	Ms. XYZ	F	15/1/1975	$60,500	$0		Y
7	6	Mr. GHI	M	9/10/1942	NA	$55,000	TX	N
8	7	Mr. TUV	M	2/27/1956	$300,000	$50,000	CT	Y
9	8	Ms. DEF	F	4/4/1980	$55,000	$0	British Columbia	N

The possible errors in a raw dataset include the following:

1 *Incompleteness error* is where the data are not present, resulting in missing data. This can be corrected by investigating alternate data sources. Missing values and NAs (not applicable or not available values) must be either omitted or replaced with "NA" for deletion or substitution with imputed values during the data exploration stage. The most common imputations are mean, median, or mode of the variable or simply assuming zero. In Exhibit 3, rows 4 (ID 3), 5 (ID 4), 6 (ID 5), and 7 (ID 6) are incomplete due to missing values in either Gender, Salary, Other Income, Name (Salutation), and State columns.

2 *Invalidity error* is where the data are outside of a meaningful range, resulting in invalid data. This can be corrected by verifying other administrative data records. In Exhibit 3, row 5 likely contains invalid data as the date of birth is out of the range of the expected human life span.

3 *Inaccuracy error* is where the data are not a measure of true value. This can be rectified with the help of business records and administrators. In Exhibit 3, row 5 is inaccurate (it shows "Don't Know"); in reality, every person either has a credit card or does not.

4 *Inconsistency error* is where the data conflict with the corresponding data points or reality. This contradiction should be eliminated by clarifying with another source. In Exhibit 3, row 3 (ID 2) is likely to be inconsistent as the Name column contains a female title and the Gender column contains male.

5 *Non-uniformity error* is where the data are not present in an identical format. This can be resolved by converting the data points into a preferable standard format. In Exhibit 3, the data under the Date of Birth column is present in various formats. The data under the Salary column may also be non-uniform as the monetary units are ambiguous; the dollar symbol can represent US dollar, Canadian dollar, or others.

6 *Duplication error* is where duplicate observations are present. This can be corrected by removing the duplicate entries. In Exhibit 3, row 6 is a duplicate as the data under Name and Date of Birth columns are identical to the ones in row 3, referring to the same customer.

Exhibit 4 shows the dataset after completion of the cleansing process.

Exhibit 4. Data After Cleansing

1	ID	Name	Gender	Date of Birth	Salary	Other Income	State	Credit Card
2	1	Mr. ABC	M	12/5/1970	USD 50200	USD 5000	VA	Y
3	2	Ms. XYZ	F	1/15/1975	USD 60500	USD 0	NY	Y
4	3	Mr. EFG	M	1/13/1979	USD 65000	USD 1000	CA	N
5	6	Mr. GHI	M	9/10/1942	USD 0	USD 55000	TX	N
6	7	Mr. TUV	M	2/27/1956	USD 300000	USD 50000	CT	Y
7	8	Ms. DEF	F	4/4/1980	CAD 55000	CAD 0	British Columbia	N

Data cleansing can be expensive and cumbersome because it involves the use of automated, rule-based, and pattern recognition tools coupled with manual human inspection to sequentially check for the aforementioned types of errors row by row and column by column. The process involves a detailed data analysis as an initial step in identifying various errors that are present in the data. In addition to a manual inspection and verification of the data, analysis software, such as SPSS, can be used to understand **metadata** (data that describes and gives information about other data) about the data properties to use as a starting point to investigate any errors in the data. The business value of the project determines the necessary quality of data cleansing and subsequently the amount of resources used in the cleansing process. In case the errors cannot be resolved due to lack of available resources, the data points with errors can simply be omitted depending on the size of the dataset. For instance, if a dataset is large with more than 10,000 rows, removing a few rows (approximately 100) may not have a significant impact on the project. If a dataset is small with less than 1,000 rows, every row might be important and deleting many rows thus harmful to the project.

Data Wrangling (Preprocessing)

To make structured data ready for analyses, the data should be preprocessed. Data preprocessing primarily includes transformations and scaling of the data. These processes are exercised on the cleansed dataset. The following transformations are common in practice:

1. *Extraction:* A new variable can be extracted from the current variable for ease of analyzing and using for training the ML model. In Exhibit 4, the Date of Birth column consists of dates that are not directly suitable for analyses. Thus, an additional variable called "Age" can be extracted by calculating the number of years between the present day and date of birth.
2. *Aggregation:* Two or more variables can be aggregated into one variable to consolidate similar variables. In Exhibit 4, the two forms of income, Salary and Other Income, can be summed into a single variable called Total Income.
3. *Filtration:* The data rows that are not needed for the project must be identified and filtered. In Exhibit 4, row 7 (ID 8) has a non-US state; however, this dataset is for the US-based bank customers where it is required to have a US address.

4 *Selection:* The data columns that are intuitively not needed for the project can be removed. This should not be confused with feature selection, which is explained later. In Exhibit 4, Name and Date of Birth columns are not required for training the ML model. The ID column is sufficient to identify the observations, and the new extracted variable Age replaces the Date of Birth column.

5 *Conversion:* The variables can be of different types: nominal, ordinal, continuous, and categorical. The variables in the dataset must be converted into appropriate types to further process and analyze them correctly. This is critical for ML model training. Before converting, values must be stripped out with prefixes and suffixes, such as currency symbols. In Exhibit 4, Name is nominal, Salary and Income are continuous, Gender and Credit Card are categorical with 2 classes, and State is ordinal. In case row 7 is not excluded, the Salary in row 7 must be converted into US dollars. Also, the conversion task applies to adjusting time value of money, time zones, and others when present.

Outliers may be present in the data, and domain knowledge is needed to deal with them. Any outliers that are present must first be identified. The outliers then should be examined and a decision made to either remove or replace them with values imputed using statistical techniques. In Exhibit 4, row 6 (ID 7) is an outlier because the Salary value is far above the upper quartile. Row 5 (ID 6) is also an outlier because the Salary value is far below the lower quartile. However, after the aggregation and formation of a new variable Total Income, as shown in Exhibit 5, row 5 (ID 6), it is no longer an outlier.

In practice, several techniques can be used to detect outliers in the data. Standard deviation can be used to identify outliers in normally distributed data. In general, a data value that is outside of 3 standard deviations from the mean may be considered an outlier. The interquartile range (IQR) can be used to identify outliers in data with any form of distribution. IQR is the difference between the 75th and the 25th percentile values of the data. The center of the IQR is the median (50th percentile). In general, data values outside of 1.5 IQR are considered as outliers and values outside of 3 IQR as extreme values.

There are several practical methods for handling outliers. When extreme values and outliers are simply removed from the dataset, it is known as **trimming** (also called truncation). For example, a 5% trimmed dataset is one for which the 5% highest and the 5% lowest values have been removed. When extreme values and outliers are replaced with the maximum (for large value outliers) and minimum (for small value outliers) values of data points that are not outliers, the process is known as **winsorization**.

Exhibit 5 Data After Applying Transformations

1	ID	Gender	Age	Total Income	State	Credit Card
2	1	M	48	55200	VA	Y
3	2	F	43	60500	NY	Y
4	3	M	39	66000	CA	N
5	6	M	76	55000	TX	N

Scaling is a process of adjusting the range of a feature by shifting and changing the scale of data. Variables, such as age and income, can have a diversity of ranges that result in a heterogeneous training dataset. For better ML model training when using such methods as support vector machines (SVMs) and artificial neural networks

(ANNs), all variables should have values in the same range to make the dataset homogeneous. It is important to remove outliers before scaling is performed. Here are two of the most common ways of scaling:

1 *Normalization* is the process of rescaling numeric variables in the range of [0, 1]. To normalize variable X, the minimum value (X_{min}) is subtracted from each observation (X_i), and then this value is divided by the difference between the maximum and minimum values of X ($X_{max} - X_{min}$) as follows:

$$X_{i\ \text{(normalized)}} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{1}$$

2 *Standardization* is the process of both centering and scaling the variables. Centering involves subtracting the mean (μ) of the variable from each observation (X_i) so the new mean is 0. Scaling adjusts the range of the data by dividing the centered values ($X_i - \mu$) by the standard deviation (σ) of feature X. The resultant standardized variable will have an arithmetic mean of 0 and standard deviation of 1.

$$X_{i\ \text{(standardized)}} = \frac{X_i - \mu}{\sigma} \tag{2}$$

Normalization is sensitive to outliers, so treatment of outliers is necessary before normalization is performed. Normalization can be used when the distribution of the data is not known. Standardization is relatively less sensitive to outliers as it depends on the mean and standard deviation of the data. However, the data must be normally distributed to use standardization.

EXAMPLE 2

Preparing and Wrangling Structured Data

Paul Wang's analytics team at LendALot Corporation is working to develop its first ML model for classifying prospective borrowers' creditworthiness. Wang has asked one of his data scientists, Lynn Lee, to perform a preliminary assessment of the data cleansing and preprocessing tasks the team will need to perform. As part of this assessment, Lee pulled the following sample of data for manual examination, which she brings to Wang to discuss.

1	ID	Name	Loan Outcome	Income (USD)	Loan Amount (USD)	Credit Score	Loan Type
2	1	Mr. Alpha	No Default	34,000	10,000	685	Mortgage
3	2	Ms. Beta	No Default	–63,050	49,000	770	Student Loan
4	3	Mr. Gamma	Defaulted	20,565	35,000	730	
5	4	Ms. Delta	No Default	50,021	unknown	664	Mortgage
6	5	Mr. Epsilon	Defaulted	100,350	129,000	705	Car Loan
7	6	Mr. Zeta	No Default	800,000	300,000	800	Boat Loan
8	6	Mr. Zeta	No Default	800,000	300,000	800	Boat Loan

After sharing a concern that the data should be thoroughly cleansed, Wang makes the following statements:

Statement 1 "Let's keep the ID column and remove the column for Name from the dataset."

Statement 2 "Let's create a new feature, "Loan Amount as a Percent of Income," to use as an additional feature."

1 The data shown for Ms. Beta contain what is *best described* as an:

A invalidity error.

B inaccuracy error.

C incompleteness error.

2 The data shown for Mr. Gamma contain what is *best described* as an:

A invalidity error.

B duplication error.

C incompleteness error.

3 The data shown for Ms. Delta contain what is *best described* as an:

A invalidity error.

B inaccuracy error.

C duplication error.

4 The data shown for Mr. Zeta contain what is *best described* as an:

A invalidity error.

B inaccuracy error.

C duplication error.

5 The process mentioned in Wang's first statement is *best described* as:

A feature selection.

B feature extraction.

C feature engineering

6 Wang's second statement is *best described* as:

A feature selection.

B feature extraction.

C feature engineering.

Solution to 1:

A is correct. This is an invalidity error because the data are outside of a meaningful range. Income cannot be negative.

Solution to 2:

C is correct. This is an incompleteness error as the loan type is missing.

Solution to 3:

B is correct. This is an inaccuracy error because LendALot must know how much they have lent to that particular borrower (who eventually repaid the loan as indicated by the loan outcome of no default).

Solution to 4:

C is correct. Row 8 duplicates row 7: This is a duplication error.

Solution to 5:

A is correct. The process mentioned involves selecting the features to use. The proposal makes sense; with "ID," "Name" is not needed to identify an observation.

Solution to 6:

B is correct. The proposed feature is a ratio of two existing features. *Feature extraction* is the process of creating (i.e., extracting) new variables from existing ones in the data.

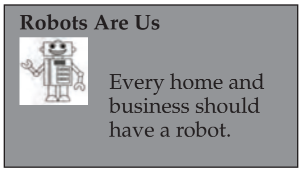
4.2 Unstructured (Text) Data

Unstructured data are not organized into any systematic format that can be processed by computers directly. They are available in formats meant for human usage rather than computer processing. Unstructured data constitute approximately 80% of the total data available today. They can be in the form of text, images, videos, and audio files. Unlike in structured data, preparing and wrangling unstructured data are both more challenging. For analysis and use to train the ML model, the unstructured data must be transformed into structured data. In this section, text data will be used to demonstrate unstructured data preparation and wrangling. The cleansing and preprocessing of text data is called *text processing.* Text processing is essentially cleansing and transforming the unstructured text data into a structured format. Text processing can be divided into two tasks: cleansing and preprocessing. The following content is related to text data in the English language.

Text Preparation (Cleansing)

Raw text data are a sequence of characters and contain other non-useful elements, including html tags, punctuations, and white spaces (including tabs, line breaks, and new lines). It is important to clean the text data before preprocessing. Exhibit 6 shows a sample text from the home page for the hypothetical company Robots Are Us website. The text appears to be clean visually and is designed for human readability.

Exhibit 6 Sample Text from Robots Are Us Home Page



However, the source text that can be downloaded is not as clean. The raw text contains html tags and formatting elements along with the actual text. Exhibit 7 shows the raw text from the source.

Exhibit 7 Raw Text from the Source

<h1 class="text-left mb-3">Robots Are Us</h1>
<h2> Every home and business shoudl have a robot </h2>

The initial step in text processing is cleansing, which involves basic operations to clean the text by removing unnecessary elements from the raw text. Text operations often use regular expressions. A **regular expression (regex)** is a series that contains characters in a particular order. Regex is used to search for patterns of interest in a given text. For example, a regex "<.*?>" can be used to find all the html tags that are present in the form of <…> in text.[1] GREP (global regular expression print) is a commonly available utility in programming languages for searching patterns using regex. Once a pattern is found, it can be removed or replaced. Additionally, advanced html parsers and packages are available in the popular programming languages, such as R and Python, to deal with this task.

The following steps describe the basic operations in the text cleansing process.

1. *Remove html tags*: Most of the text data are acquired from web pages, and the text inherits html markup tags with the actual content. The initial task is to remove (or strip) the html tags that are not part of the actual text using programming functions or using regular expressions. In Exhibit 7, </h2> is an html tag that can be identified by a regex and be removed. Note that it is not uncommon to keep some generic html tags to maintain certain formatting meaning in the text.

2. *Remove Punctuations*: Most punctuations are not necessary for text analysis and should be removed. However, some punctuations, such as percentage signs, currency symbols, and question marks, may be useful for ML model training. These punctuations should be substituted with such annotations as /percentSign/, /dollarSign/, and /questionMark/ to preserve their grammatical meaning in the text. Such annotations preserve the semantic meaning of important characters in the text for further text processing and analysis stages. It is important to note that periods (dots) in the text need to be processed carefully. There are different circumstances for periods to be present in text—characteristically used for abbreviations, sentence boundaries, and decimal points. The periods and the context in which they are used need to be identified and must be appropriately replaced or removed. In general, periods after abbreviations can be removed, but the periods separating sentences should be replaced by the annotation /endSentence/. Some punctuations, such as hyphens and underscores, can be kept in the text to keep the consecutive words intact as a single term (e.g., e-mail). Regex are often used to remove or replace punctuations.

3. *Remove Numbers*: When numbers (or digits) are present in the text, they should be removed or substituted with an annotation /number/. This helps inform the computer that a number is present, but the actual value of the number itself is not helpful for categorizing/analyzing the text. Such operations are critical for ML model training. Otherwise, the computers will treat each number as a separate word, which may complicate the analyses or add noise. Regex are often used to remove or replace numbers. However, the number and any decimals must be retained where the outputs of interest are the actual values of the number. One such text application is information extraction (IE), where the goal is to extract relevant information from a given text. An IE task could be extracting monetary values from financial reports, where the actual number values are critical.

4. *Remove white spaces*: It is possible to have extra white spaces, tab spaces, and leading and ending spaces in the text. The extra white spaces may be introduced after executing the previously mentioned operations. These should be identified and removed to keep the text intact and clean. Certain functions in programming languages can be used to remove unnecessary white spaces from the text. For example, the text mining package in R offers a *stripwhitespace* function.

---

1 A regex of the form "<.*?>" will identify all html tags with anything (*) of any length (?) between the brackets (< >).

Exhibit 8 uses a sample financial text to show the transformations occurring after applying each operation of the text cleansing process. The four steps are applied on a mock financial text after scraping from a source. As noted previously, scraping (or web scraping) is a technique to extract raw content from a source, typically web pages. It is important to note that the sequence and choice of cleansing operations does matter. For instance, after removing punctuation, the "1.2 million" becomes "12 million." This is acceptable here since a subsequent operation replaces all numbers with a "/number/" tag. However, if numbers were not replaced with such tags, the punctuation removal operation could affect the data.

Exhibit 8 Text Cleansing Process Example

Original text from a financial statement as shown on a webpage

CapEx on the normal operations remained stable on historicallylow levels, $800,000 compared to $1.2 million last year.

Quarter 3, so far, is 5% sales growth quarter-to-date, and year-to-date, we have a 4% local currency sales development.

Raw text after scraping from the source

<p> CapEx on the normal operations remained stable on historically low levels, $800,000 compared to $1.2 million last year. <b/><b/> Quarter 3, so far, is 5% sales growth quarter-to-date, and year-to-date, we have a 4% local currency sales development.</p>

Text after removing html tags

CapEx on the normal operations remained stable on historically low levels, $800,000 compared to $1.2 million last year.
Quarter 3, so far, is 5% sales growth quarter-to-date, and year-to-date, we have a 4% local currency sales development.

Text after removing and replacing punctuations

CapEx on the normal operations remained stable on historically low levels /dollarSign/800000 compared to /dollarSign/12 million last year /endSentence/ Quarter 3 so far is 5 /percentSign/ sales growth quarter-to-date and year-to-date we have a 4 /percentSign/ local currency sales development /endSentence/

Text after replacing numbers

CapEx on the normal operations remained stable on historically low levels /dollarSign//number / compared to/dollarSign//number/ million last year /endSentence/ Quarter/number/ so far is /number/ /percentSign/sales growth quarter-to-date and year-to-date we have a /number/ / percentSign/ local currency sales development /endSentence/

Text after removing extra white spaces

CapEx on the normal operations remained stable on historically low levels/dollarSign//number /compared to/dollarSign//number/million last year/endSentence/ Quarter/number/so far is /number//percentSign/sales growth quarter-to-date and year-to-date we have a/number// percentSign/local currency sales development/endSentence/

Text Wrangling (Preprocessing)

To further understand text processing, tokens and tokenization need to be defined. A **token** is equivalent to a word, and **tokenization** is the process of splitting a given text into separate tokens. In other words, a text is considered to be a collection of tokens. Tokenization can be performed at word or character level, but it is most commonly performed at word level. Exhibit 9 shows a sample dataset of four cleansed texts and their word tokens.

Exhibit 9 Tokenization of Four Texts

	Cleaned Texts	Tokens
Text 1	The man went to the market today	The · man · went · to · the · market · today
Text 2	Market values are increasing	Market · values · are · increasing
Text 3	Increased marketing is needed	Increased · marketing · is · needed
Text 4	There is no market for the product	There · is · no · market · for · the · product

Similar to structured data, text data also require normalization. The normalization process in text processing involves the following:

1 *Lowercasing* the alphabet removes distinctions among the same words due to upper and lower cases. This action helps the computers to process the same words appropriately (e.g., "The" and "the").

2 *Stop words* are such commonly used words as "the," "is," and "a." Stop words do not carry a semantic meaning for the purpose of text analyses and ML training. However, depending on the end-use of text processing, for advance text applications it may be critical to keep the stop words in the text in order to understand the context of adjacent words. For ML training purposes, stop words typically are removed to reduce the number of tokens involved in the training set. A predefined list of stop words is available in programming languages to help with this task. In some cases, additional stop words can be added to the list based on the content. For example, the word "exhibit" may occur often in financial filings, which in general is not a stop word but in the context of the filings can be treated as a stop word.

3 *Stemming* is the process of converting inflected forms of a word into its base word (known as stem). Stemming is a rule-based approach, and the results need not necessarily be linguistically sensible. Stems may not be the same as the morphological root of the word. Porter's algorithm is the most popular method for stemming. For example, the stem of the words "analyzed" and "analyzing" is "analyz." Similarly, the British English variant "analysing" would become "analys." Stemming is available in R and Python. The text mining package in R provides a *stemDocument* function that uses this algorithm.

4 *Lemmatization* is the process of converting inflected forms of a word into its morphological root (known as lemma). Lemmatization is an algorithmic approach and depends on the knowledge of the word and language structure. For example, the lemma of the words "analyzed" and "analyzing" is "analyze." Lemmatization is computationally more expensive and advanced.

Stemming or lemmatization will reduce the repetition of words occurring in various forms and maintain the semantic structure of the text data. Stemming is more common than lemmatization in the English language since it is simpler to perform. In text data, data sparseness refers to words that appear very infrequently, resulting in data consisting of many unique, low frequency tokens. Both techniques decrease data sparseness by aggregating many sparsely occurring words in relatively less sparse stems or lemmas, thereby aiding in training less complex ML models.

After the cleansed text is normalized, a bag-of-words is created. **Bag-of-words (BOW)** representation is a basic procedure used to analyze text. It is essentially a collection of a distinct set of tokens from all the texts in a sample dataset. BOW is simply a set of words and does not capture the position or sequence of words present in the text. However, it is memory efficient and easy to handle for text analyses.

Exhibit 10 shows the BOW and transformations occurring in each step of normalization on the cleansed texts from Exhibit 9. Note that the number of words decreases as the normalizing steps are applied, making the resulting BOW smaller and simpler.

Exhibit 10 Bag-of-Words Representation of Four Texts Before and After Normalization Process

BOW before normalizing

"The"	"man"	"went"	"to"	"the"	"market"
"today"	"Market"	"values"	"are"	"increasing"	"Increased"
"marketing"	"is"	"needed"	"There"	"no"	"for"
"product"					

BOW after removing uppercase letters

"the"	"man"	"went"	"to"	"market"	"today"
"values"	"are"	"increasing"	"increased"	"marketing"	"is"
"needed"	"there"	"no"	"for"	"product"	

BOW after removing stop words

"man"	"went"	"market"	"today"	"values"	"increasing"
"increased"	"marketing"	"needed"	"product"		

BOW after stemming

"man"	"went"	"market"	"today"	"valu"	"increas"	"need"	"product"

The last step of text preprocessing is using the final BOW after normalizing to build a **document term matrix (DTM)**. DTM is a matrix that is similar to a data table for structured data and is widely used for text data. Each row of the matrix belongs to a document (or text file), and each column represents a token (or term). The number of rows of DTM is equal to the number of documents (or text files) in a sample dataset. The number of columns is equal to the number of tokens from the BOW that is built using all the documents in a sample dataset. The cells can contain the counts of the number of times a token is present in each document. The matrix cells can be filled with other values that will be explained in the financial forecasting project section of this reading; a large dataset is helpful in understanding the concepts. At this point, the unstructured text data are converted to structured data that can be processed further and used to train the ML model. Exhibit 11 shows a DTM constructed from the resultant BOW of the four texts from Exhibit 10.

Exhibit 11 DTM of Four Texts and Using Normalized BOW Filled with Counts of Occurrence

	man	went	market	today	valu	increas	need	product
Text 1	1	1	1	1	0	0	0	0
Text 2	0	0	1	0	1	1	0	0
Text 3	0	0	1	0	0	1	1	0
Text 4	0	0	1	0	0	0	0	1

As seen in Exhibit 10, BOW does not represent the word sequences or positions, which limits its use for some advanced ML training applications. In the example, the word "no" is treated as a single token and has been removed during the normalization because it is a stop word. Consequently, this fails to signify the negative meaning ("no market") of the text (i.e., Text 4). To overcome such problems, a technique called n-grams can be employed. **N-grams** is a representation of word sequences. The length of a sequence can vary from 1 to n. When one word is used, it is a unigram; a two-word sequence is a bigram; and a 3-word sequence is a trigram; and so on. Exhibit 10, for example, shows a unigram ($n = 1$) BOW. The advantage of n-grams is that they can be used in the same way as unigrams to build a BOW. In practice, different n-grams can be combined to form a BOW and eventually be used to build a DTM. Exhibit 12 shows unigrams, bigrams, and trigrams. Exhibit 12 also shows a combined unigram-to-trigram BOW for the particular text. Stemming can be applied on the cleansed text before building n-grams and BOW (not shown in Exhibit 12).

Exhibit 12 N-Grams and N-Grams BOW

Clean text

The man went to the market today

Unigrams

"The" "man" "went" "to" "the" "market" "today"

Bigrams

"The_man" "man_went" "went_to" "to_the" "the_market" "market_today"

Trigrams

"The_man_went" "man_went_to" "went_to_the" "to_the_market" "the_market_today"

BOW before normalizing

"The"	"man"	"went"	"to"	"the"	"market"	"today"
"The_man"	"man_went"	"went_to"	"to_the"	"the_market"	"market_today"	"The_man_went"
"man_went_to"	"went_to_the"	"to_the_market"	"the_market_today"			

BOW after removing upper case letters

"the"	"man"	"went"	"to"	"market"	"today"	"the_man"
"man_went"	"went_to"	"to_the"	"the_market"	"market_today"	"the_man_went"	"man_went_to"
"went_to_the"	"to_the_market"	"the_market_today"				

BOW after removing stop words

"man"	"went"	"market"	"today"	"the_man"	"man_went"	"went_to"
"to_the"	"the_market"	"market_today"	"the_man_went"	"man_went_to"	"went_to_the"	"to_the_market"
"the_market_today"						

The n-grams implementation will vary the impact of normalization on the BOW. Even after removing isolated stop words, stop words tend to persist when they are attached to their adjacent words. For instance, "to_the" (Exhibit 12) is a single bigram token consisting of stop words and will not be removed by the predetermined list of stop words.

EXAMPLE 3

Unstructured Data Preparation and Wrangling

1 The output produced by preparing and wrangling textual data is best described as a:

A data table.

B confusion matrix.

C document term matrix.

2 In text cleansing, situations in which one may need to add an annotation include the removal of:

A html tags.

B white spaces.

C punctuations.

3 A column of a document term matrix is *best* described as representing:

A a token.

B a regularization term.

C an instance.

4 A cell of a document term matrix is *best* described as containing:

A a token.

B a count of tokens.

C a count of instances.

5 Points to cover in normalizing textual data include:

A removing numbers.

B removing white spaces.

C lowercasing the alphabet.

6 When some words appear very infrequently in a textual dataset, techniques that may address the risk of training highly complex models include:

A stemming.

B scaling.

C data cleansing.

7 Which of the following statements concerning tokenization is *most* accurate?

A Tokenization is part of the text cleansing process.

B Tokenization is most commonly performed at the character level.

C Tokenization is the process of splitting a given text into separate tokens.

Solution to 1:

C is correct. The objective of data preparation and wrangling of textual data is to transform the unstructured data into structured data. The output of these processes is a document term matrix that can be read by computers. The document term matrix is similar to a data table for structured data.

Solution to 2:

C is correct. Some punctuations, such as percentage signs, currency symbols, and question marks, may be useful for ML model training, so when such punctuations are removed annotations should be added.

Solution to 3:

A is correct. Each column of a document term matrix represents a token from the bag-of-words that is built using all the documents in a sample dataset.

Solution to 4:

B is correct. A cell in a document term matrix contains a count of the number of tokens of the kind indicated in the column heading.

Solution to 5:

C is correct. The other choices are related to text cleansing.

Solution to 6:

A is correct. Stemming, the process of converting inflected word forms into a base word (or stem), is one technique that can address the problem described.

Solution to 7:

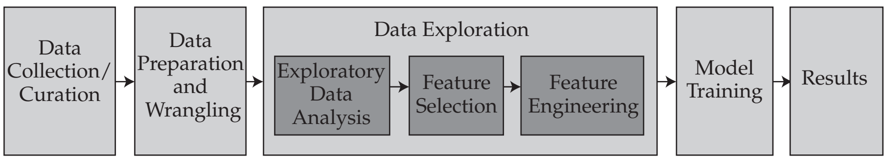
C is correct, by definition. The other choices are not true.

5 DATA EXPLORATION OBJECTIVES AND METHODS

Data exploration is a crucial part of big data projects. The prepared data are explored to investigate and comprehend data distributions and relationships. The knowledge that is gained about the data in this stage is used throughout the project. The outcome and quality of exploration strongly affects ML model training results. Domain knowledge plays a vital role in exploratory analysis as this stage should involve cooperation between analysts, model designers, and experts in the particular data domain. Data exploration without domain knowledge can result in ascertaining spurious relationships among the variables in the data that can mislead the analyses. The data exploration stage follows the data preparation stage and leads to the model training stage.

Data exploration involves three important tasks: exploratory data analysis, feature selection, and feature engineering. These three tasks are outlined in Exhibit 13 and are defined and further explained under the structured and unstructured data subsections.

Exhibit 13 Data Exploration Stage



Exploratory data analysis (EDA) is the preliminary step in data exploration. Exploratory graphs, charts, and other visualizations, such as heat maps and word clouds, are designed to summarize and observe data. In practice, many exploratory graphs are made for investigation and can be made swiftly using statistical programming and generic spreadsheet software tools. Data can also be summarized and examined using quantitative methods, such as descriptive statistics and central tendency measures. An important objective of EDA is to serve as a communication medium among project stakeholders, including business users, domain experts, and analysts. Relatively quick and easy exploratory visualizations help stakeholders connect and ensure the prepared data are sensible. Other objectives of EDA include:

- understanding data properties,
- finding patterns and relationships in data,
- inspecting basic questions and hypotheses,
- documenting data distributions and other characteristics, and
- planning modeling strategies for the next steps.

Feature selection is a process whereby only pertinent features from the dataset are selected for ML model training. Selecting fewer features decreases ML model complexity and training time. **Feature engineering** is a process of creating new features by changing or transforming existing features. Model performance heavily depends on feature selection and engineering.

5.1 Structured Data

Exploratory Data Analysis

For structured data, each data table row contains an observation and each column contains a feature. EDA can be performed on a single feature (one-dimension) or on multiple features (multi-dimension). For high-dimension data with many features, EDA can be facilitated by using a dimension reduction technique, such as principal components analysis (PCA). Based on the number of dimensions, the exploratory techniques will vary.

For one-dimensional data, summary statistics, such as mean, median, quartiles, ranges, standard deviations, skewness, and kurtosis, of a feature can be computed. One-dimension visualization summarizes each feature in the dataset. The basic one-dimension exploratory visualizations are as follows:

- Histograms
- Bar charts
- Box plots
- Density plots

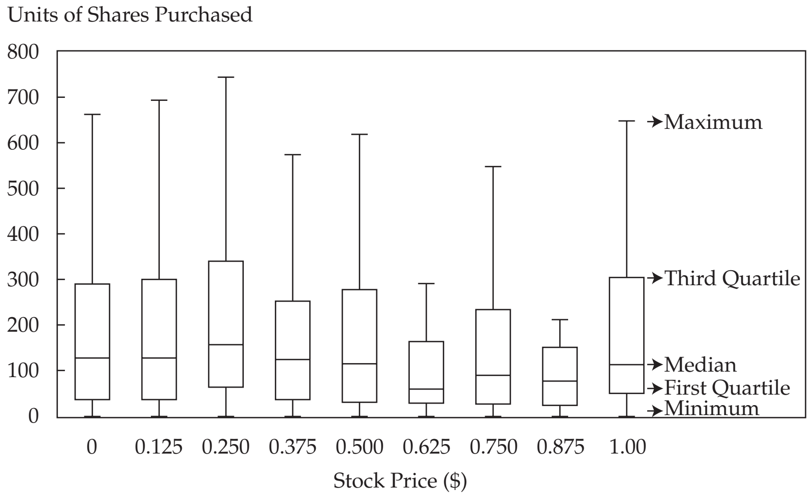
Histograms represent equal bins of data and their respective frequencies. They can be used to understand the high-level distribution of the data. Bar charts summarize the frequencies of categorical variables. Box plots show the distribution of continuous data by highlighting the median, quartiles, and outliers of a feature that is normally distributed. Density plots are another effective way to understand the distribution of continuous data. Density plots are smoothed histograms and are commonly laid on top of histograms, as shown in Exhibit 14. This histogram shows a hypothetical annual salary distribution (in £) of entry-level analyst positions at UK banks. The data represent a normal distribution with an approximate mean of £68,500.

Exhibit 14 Histogram with Superimposed Density Plot

For data with two or more dimensions, summary statistics of relationships, such as a correlation matrix, can be calculated. Two- or more-dimensional visualization explores interactions between different features in the dataset. Common methods include scatterplots and line graphs. In multi-dimensional visualization, one-dimensional plots are overlaid to summarize each feature, thus enabling comparison between features. Additionally, attributes (e.g., color, shape, and size) and legends can be used creatively to pack more information about the data into fewer graphs.

For multivariate data, commonly utilized exploratory visualization designs include stacked bar and line charts, multiple box plots, and scatterplots showing multivariate data that use different colors or shapes for each feature. Multiple box plots can be arranged in a single chart, where each individual box plot represents a feature. Such a multi-box plot chart assesses the relationship between each feature (x-axis) in the dataset and the target variable of interest (y-axis). The multi-box plot chart in Exhibit 15 represents units of shares purchased versus stock price for a hypothetical stock. The x-axis shows the stock price in increments of $0.125, and the y-axis shows units of shares purchased. The individual box plots indicate the distribution of shares purchased at the different stock prices. When the stock price is $0.25, the median number of shares purchased is the highest; when the stock price is $0.625, the median number of shares purchased is the lowest. However, visually it appears that the number of shares purchased at different stock prices is not significantly different.

Exhibit 15 Multiple Box Plots in One Chart



Two-dimensional charts can summarize and approximately measure relationships between two or more features. An example scatterplot in Exhibit 16 shows the interaction of two hypothetical features: age (x-axis) and annual salary (y-axis). The feature on the y-axis tends to increase as the feature on the x-axis increases. This pattern appears true visually; however, it may not be a statistically significant relationship. A scatterplot provides a starting point where relationships can be examined visually. These potential relationships should be tested further using statistical tests. Common parametric statistical tests include ANOVA, *t*-test, and Pearson correlation. Common non-parametric statistical tests include chi-square and the Spearman rank-order correlation.

Exhibit 16 Scatterplot Showing a Linear Relationship Between Two Features

In addition to visualization, descriptive statistics are a good means to summarize data. Central tendency measures as well as minimum and maximum values for continuous data are useful. Counts and frequencies for categorical data are commonly employed to gain insight regarding the distribution of possible values.

EDA is not only useful for revealing possible relationships among features or general trends in the data; it is also beneficial during the feature selection and engineering stages. These possible relationships and trends in the data may be used to suggest new features that, when incorporated into a model, may improve model training.

Feature Selection

Structured data consist of features, represented by different columns of data in a table or matrix. After using EDA to discover relevant patterns in the data, it is essential to identify and remove unneeded, irrelevant, and redundant features. Basic diagnostic testing should also be performed on features to identify redundancy, heteroscedasticity, and multi-collinearity. The objective of the feature selection process is to assist in identifying significant features that when used in a model retain the important patterns and complexities of the larger dataset while requiring fewer data overall. This last point is important since computing power is not free (i.e., explicit costs and processing time).

Typically, structured data even after the data preparation step can contain features that do not contribute to the accuracy of an ML model or that negatively affect the quality of ML training. The most desirable outcome is a parsimonious model with fewer features that provides the maximum predictive power out-of-sample.

Feature selection must not be confused with the data preprocessing steps during data preparation. Good feature selection requires an understanding of the data and statistics, and comprehensive EDA must be performed to assist with this step. Data preprocessing needs clarification only from data administrators and basic intuition (e.g., salary vs. income) during data preparation.

Feature selection on structured data is a methodical and iterative process. Statistical measures can be used to assign a score gauging the importance of each feature. The features can then be ranked using this score and either retained or eliminated from the dataset. The statistical methods utilized for this task are usually univariate and consider each feature independently or with regard to the target variable. Methods include chi-square test, correlation coefficients, and information-gain measures (i.e., R-squared values from regression analysis). All of these statistical methods can be combined in a manner that uses each method individually on each feature, automatically performing backward and forward passes over features to improve feature selection. Prebuilt feature selection functions are available in popular programming languages used to build and train ML models.

Dimensionality reduction assists in identifying the features in the data that account for the greatest variance between observations and allows for the processing of a reduced volume of data. Dimensionality reduction may be implemented to reduce a large number of features, which helps reduce the memory needed and speed up learning algorithms. Feature selection is different from dimensionality reduction, but both methods seek to reduce the number of features in the dataset. The dimensionality reduction method creates new combinations of features that are uncorrelated, whereas feature selection includes and excludes features present in the data without altering them.

Feature Engineering

After the appropriate features are selected, feature engineering helps further optimize and improve the features. The success of ML model training depends on how well the data are presented to the model. The feature engineering process attempts to produce good features that describe the structures inherent in the dataset. This process depends on the context of the project, domain of the data, and nature of the problem. Structured data are likely to contain quantities, which can be engineered to better present relevant patterns in the dataset. This action involves engineering an existing feature into a new feature or decomposing it into multiple features.

For continuous data, a new feature may be created—for example, by taking the logarithm of the product of two or more features. As another example, when considering a salary or income feature, it may be important to recognize that different salary brackets impose a different taxation rate. Domain knowledge can be used to decompose an income feature into different tax brackets, resulting in a new feature: "income_above_100k," with possible values 0 and 1. The value 1 under the new feature captures the fact that a subject has an annual salary of more than $100,000. By grouping subjects into income categories, assumptions about income tax can be made and utilized in a model that uses the income tax implications of higher and lower salaries to make financial predictions.

For categorical data, for example, a new feature can be a combination (e.g., sum or product) of two features or a decomposition of one feature into many. If a single categorical feature represents education level with five possible values—high school, associates, bachelor's, master's, and doctorate—then these values can be decomposed into five new features, one for each possible value (e.g., is_highSchool, is_doctorate) filled with 0s (for false) and 1s (for true). The process in which categorical variables are converted into binary form (0 or 1) for machine reading is called **one hot encoding**. It is one of the most common methods for handling categorical features in text data. When date-time is present in the data, such features as "second of the hour," "hour of the day," and "day of the date" can be engineered to capture critical information about temporal data attributes—which are important, for example, in modeling trading algorithms.

Feature engineering techniques systemically alter, decompose, or combine existing features to produce more meaningful features. More meaningful features allow an ML model to train more swiftly and easily. Different feature engineering strategies can lead to the generation of dramatically different results from the same ML model. The impact of feature selection and engineering on ML training is discussed further in the next section.

5.2 Unstructured Data: Text Exploration

Exploratory Data Analysis

Just like with structured data, it is important to gain insight into existing patterns in the unstructured data for further analysis. In this section, text data will be discussed. Text analytics has various applications. The most common applications are text classification, topic modeling, fraud detection, and sentiment analysis. Text classification uses supervised ML approaches to classify texts into different classes. Topic modeling uses unsupervised ML approaches to group the texts in the dataset into topic clusters. Sentiment analysis predicts sentiment (negative, neutral, or positive) of the texts in a dataset using both supervised and unsupervised approaches.

Various statistics are used to explore, summarize, and analyze text data. Text data include a collection of texts (also known as a corpus) that are sequences of tokens. It is useful to perform EDA of text data by computing on the tokens such basic text statistics as **term frequency (TF)**, the ratio of the number of times a given token occurs in all the texts in the dataset to the total number of tokens in the dataset (e.g., word associations, average word and sentence length, and word and syllable counts).

Text statistics reveal patterns in the co-occurrence of words. There are many applications of text analytics, and necessary text statistics vary according to the context of the application. Topic modeling is a text data application in which the words that are most informative are identified by calculating the TF of each word. For example, the word "soccer" can be informative for the topic "sports." The words with high TF values are eliminated as they are likely to be stop words or other common vocabulary words, making the resulting BOW compact and more likely to be relevant to topics within the texts. In sentiment analysis and text classification applications, the chi-square measure of word association can be useful for understanding the significant word appearances in negative and positive sentences in the text or in different documents. The chi-square measure is further explained under feature selection. Such EDA plays a vital role in executing the feature selection step.

Text statistics can be visually comprehended by using the same methods as explained in the structured data section. For example, bar charts can be used to show word counts or frequency. Words clouds are common visualizations when working with text data as they can be made to visualize the most informative words and their TF values. The most commonly occurring words in the dataset can be shown by varying font size, and color is used to add more dimensions, such as frequency and length of words. Exhibit 17 shows a word cloud constructed from a sample dataset of generic financial news wires after text processing. Word cloud building functions and packages are available in several popular programming languages. A detailed demonstration of text data EDA will be presented in Section 7, where we work with actual text data in a financial forecasting project.

Exhibit 17 Word Cloud of Generic Financial Newsfeed Data Sample

Feature Selection

For text data, feature selection involves selecting a subset of the terms or tokens occurring in the dataset. The tokens serve as features for ML model training. Feature selection in text data effectively decreases the size of the vocabulary or BOW. This helps the ML model be more efficient and less complex. Another benefit is to eliminate noisy features from the dataset. Noisy features are tokens that do not contribute to ML model training and actually might detract from the ML model accuracy.

Noisy features are both the most frequent and most sparse (or rare) tokens in the dataset. On one end, noisy features can be stop words that are typically present frequently in all the texts across the dataset. On the other end, noisy features can be sparse terms that are present in only a few text cases. Text classification involves dividing text documents into assigned classes (a class is a category; examples include "relevant" and "irrelevant" text documents or "bearish" and "bullish" sentences). The *frequent* tokens strain the ML model to choose a decision boundary among the texts as the terms are present across all the texts, an example of model *underfitting*. The *rare* tokens mislead the ML model into classifying texts containing the rare terms into a specific class, an example of model *overfitting*. Identifying and removing noise features is very critical for text classification applications. The general feature selection methods in text data are as follows:

1. *Frequency* measures can be used for vocabulary pruning to remove noise features by filtering the tokens with very high and low TF values across all the texts. **Document frequency (DF)** is another frequency measure that helps to discard the noise features that carry no specific information about the text class and are present across all texts. The DF of a token is defined as the number of documents (texts) that contain the respective token divided by the total number of documents. It is the simplest feature selection method and often performs well when many thousands of tokens are present.
2. *Chi-square* test can be useful for feature selection in text data. The chi-square test is applied to test the independence of two events: occurrence of the token and occurrence of the class. The test ranks the tokens by their usefulness to each class in text classification problems. Tokens with the highest chi-square test statistic values occur more frequently in texts associated with a particular class and therefore can be selected for use as features for ML model training due to higher discriminatory potential.
3. *Mutual information* (MI) measures how much information is contributed by a token to a class of texts. The **mutual information** value will be equal to 0 if the token's distribution in all text classes is the same. The MI value approaches 1 as the token in any one class tends to occur more often in only that particular class of text. Exhibit 18 shows a simple depiction of some tokens with high MI scores for their corresponding text classes. Note how the tokens (or features) with the highest MI values narrowly relate to their corresponding text class name.

Exhibit 18 Tokens with Mutual Information (MI) Values for Two Given Text Classes

Text Classes: Sports or Politics			
Sports		Politics	
Token	MI Value	Token	MI Value
soccer	0.0781	election	0.0612
cup	0.0525	president	0.0511
match	0.0456	polls	0.0341
play	0.0387	vote	0.0288
game	0.0299	party	0.0202
team	0.0265	candidate	0.0201
win	0.0189	campaign	0.0201

Feature Engineering

As with structured data, feature engineering can greatly improve ML model training and remains a combination of art and science. The following are some techniques for feature engineering, which may overlap with text processing techniques.

1 *Numbers*: In text processing, numbers are converted into a token, such as "/number/." However, numbers can be of different lengths of digits representing different kinds of numbers, so it may be useful to convert different numbers into different tokens. For example, numbers with four digits may indicate years, and numbers with many digits could be an identification number. Four-digit numbers can be replaced with "/number4/," 10-digit numbers with "/number10/," and so forth.

2 *N-grams*: Multi-word patterns that are particularly discriminative can be identified and their connection kept intact. For example, "market" is a common word that can be indicative of many subjects or classes; the words "stock market" are used in a particular context and may be helpful to distinguish general texts from finance-related texts. Here, a bigram would be useful as it treats the two adjacent words as a single token (e.g., stock_market).

3 *Name entity recognition (NER)*: NER is an extensive procedure available as a library or package in many programming languages. The **name entity recognition** algorithm analyzes the individual tokens and their surrounding semantics while referring to its dictionary to tag an object class to the token. Exhibit 19 shows the NER tags of the text "*CFA Institute was formed in 1947 and is headquartered in Virginia*." Additional object classes are, for example, MONEY, TIME, and PERCENT, which are not present in the example text. The NER tags, when applicable, can be used as features for ML model training for better model performance. NER tags can also help identify critical tokens on which such operations as lowercasing and stemming then can be avoided (e.g., Institute here refers to an organization rather than a verb). Such techniques make the features more discriminative.

Exhibit 19 Name Entity Recognition and Parts of Speech (POS) on Example Text

Token	NER Tag	POS Tag	POS Description
CFA	ORGANIZATION	NNP	Proper noun
Institute	ORGANIZATION	NNP	Proper noun
was		VBD	Verb, past tense
formed		VBN	Verb, past participle
in		IN	Preposition
1947	DATE	CD	Cardinal number
and		CC	Coordinating conjunction
is		VBZ	Verb, 3rd person singular present
headquartered		VBN	Verb, past participle

(continued)

Exhibit 19 (Continued)

Token	NER Tag	POS Tag	POS Description
in		IN	Preposition
Virginia	LOCATION	NNP	Proper noun

4 *Parts of speech (POS)*: Similar to NER, **parts of speech** uses language structure and dictionaries to tag every token in the text with a corresponding part of speech. Some common POS tags are noun, verb, adjective, and proper noun. Exhibit 19 shows the POS tags and descriptions of tags for the example text. POS tags can be used as features for ML model training and to identify the number of tokens that belong to each POS tag. If a given text contains many proper nouns, it means that it may be related to people and organizations and may be a business topic. POS tags can be useful for separating verbs and nouns for text analytics. For example, the word "market" can be a verb when used as "to market …" or noun when used as "in the market." Differentiating such tokens can help further clarify the meaning of the text. The use of "market" as a verb could indicate that the text relates to the topic of marketing and might discuss marketing a product or service. The use of "market" as a noun could suggest that the text relates to a physical or stock market and might discuss stock trading. Also for POS tagging, such compound nouns as "CFA Institute" can be treated as a single token. POS tagging can be performed using libraries or packages in programming languages.

In addition, many more creative techniques convey text information in a structured way to the ML training process. The goal of feature engineering is to maintain the semantic essence of the text while simplifying and converting it into structured data for ML.

EXAMPLE 4

Data Exploration

Paul Wang's analytics team at LendALot Corporation has completed its initial data preparation and wrangling related to their creditworthiness classification ML model building efforts. As a next step, Wang has asked one of the team members, Eric Kim, to examine the available structured data sources to see what types of exploratory data analysis might make sense. Kim has been tasked with reporting to the team on high-level patterns and trends in the data and which variables seem interesting. Greater situational awareness about the data can inform the team's decisions regarding model training and whether (and how) to incorporate textual big data in conjunction with the structured data inputs. Use the following sample of columns and rows Kim pulled for manual examination to answer the next questions.

1	ID	Loan Outcome	Income (USD)	Loan Amount (USD)	Credit Score	Loan Type	Free Responses to "Explain Credit Score" (excerpts from full text)
2	1	No Default	34,000	10,000	685	Mortgage	I am embarrassed that my score is below 700, but it was due to mitigating circumstances. I have developed a plan to improve my score.
3	2	No Default	63,050	49,000	770	Student Loan	I have a good credit score and am constantly looking to further improve it…
4	3	Defaulted	20,565	35,000	730	Student Loan	I think I have great credit. I don't think there are any issues. Having to provide a written response to these questions is kind of annoying…
5	4	No Default	50,021	10,000	664	Mortgage	I have a decent credit score. I regret not being as responsible in the past but feel I have worked hard to improve my score recently…
6	5	Defaulted	100,350	129,000	705	Car Loan	Honestly, my score probably would have been higher if I had worked harder. But it is probably good enough…
7	6	No Default	800,000	300,000	800	Boat Loan	I have worked hard to maintain a good credit rating. I am very responsible. I maintain a payment schedule and always stick to the payment plan…

1 Evaluate whether data visualization techniques, such as histograms, box plots, and scatterplots, could be relevant to exploratory data analysis.

2 State one visualization technique that could be used in relation to the free responses.

3 Describe how ranking methods can be used to select potentially interesting features to report back to the team.

4 State an example of a bigram from the free response texts that could be used to discriminate among loan outcomes.

Solution to 1:

The data provided include structured features (ID, Loan Outcome, Income, Loan Amount, Credit Score) and unstructured data. Histograms, box plots, and scatterplots are relevant visualization methods for structured data features. Histograms and box plots could be used by Kim to see how income, loan amount, and credit score are distributed. Moreover, these visualizations can be performed across all historical borrowing instances in the dataset as well as within the sets of defaulted loans versus non-defaulted loans. Scatterplots of income versus loan amount, income versus credit score, and loan amount versus credit score, both overall and within defaulted and non-defaulted datasets, can shed light on relationships between potentially important continuous variables.

Solution to 2:

For the text in the free response field, word clouds offer an appropriate starting point for exploratory analysis. A word cloud can enable a quick glimpse into the most frequently occurring words (i.e., term frequency). While some obvious words (e.g., "credit" and "score") may be valuable, other frequently occurring words (e.g., "worked," "hard," "probably," "embarrassed," "regret," "good," "decent," and "great") might have potential use for creditworthiness prediction.

Solution to 3:

Kim can use feature selection methods to rank all features. Since the target variable of interest (loan outcome) is discrete in this case, such techniques as chi-square and information gain would be well suited. These are univariate techniques that can score feature variables individually. In addition to the structured features, these univariate ranking methods can also be applied to word count-related features, such as term frequency and document frequency, that are derived from the text using frequently occurring words. Such frequently occurring words (e.g., "worked" and "hard") can be identified from the word cloud.

Solution to 4:

The bigrams "credit_score" and "worked_hard" from the text in the free response section may have potential to discriminate among loan outcomes.

EXAMPLE 5

Textual Feature Representations for ML Model Building

Having completed their exploration of the data, Paul Wang's analytics team at LendALot Corporation recognizes the importance of incorporating features derived from text data in their ML models for classifying creditworthiness. Wang has asked his colleagues, Lynn Lee and Eric Kim, to propose textual feature representations that might be well suited to constructing features for their task. As a starting point, Lee and Kim review the following sample of data:

1	ID	Loan Outcome	Income (USD)	Loan Amount (USD)	Credit Score	Loan Type	Free Responses to "Explain Credit Score" (excerpts from full text)
2	1	No Default	34,000	10,000	685	Mortgage	I am embarrassed that my score is below 700, but it was due to mitigating circumstances. I have developed a plan to improve my score.
3	2	No Default	63,050	49,000	770	Student Loan	I have a good credit score and am constantly looking to further improve it…
4	3	Defaulted	20,565	35,000	730	Student Loan	I think I have great credit. I don't think there are any issues. Having to provide a written response to these questions is kind of annoying…
5	4	No Default	50,021	10,000	664	Mortgage	I have a decent credit score. I regret not being as responsible in the past but feel I have worked hard to improve my score recently…
6	5	Defaulted	100,350	129,000	705	Car Loan	Honestly, my score probably would have been higher if I had worked harder. But it is probably good enough…
7	6	No Default	800,000	300,000	800	Boat Loan	I have worked hard to maintain a good credit rating. I am very responsible. I maintain a payment schedule and always stick to the payment plan…

Based on the information given, address the following questions.

1 Describe three textual feature representations that Lee and Kim should consider for their text data.

2 Describe a rationale for adopting each of the three textual feature representations identified in Question 1.

Solution 1:

Lee and Kim should consider bag-of-words (BOW), n-grams, and parts-of-speech (POS) as key textual feature representations for their text data. Conversely, name entity recognition (NER) might not be as applicable in this context because the data on prospective borrowers does not include any explicit references to people, locations, dates, or organizations.

Solution 2:

All three textual feature representations have the potential to add value.

Bag-of-words (BOW) is typically applicable in most contexts involving text features derived from languages where token boundaries are explicitly present (e.g., English) or can be inferred through processing (e.g., a different language, such as Spanish). BOW is generally the best starting point for most projects exploring text feature representations.

N-grams, representations of word or token sequences, are also applicable. N-grams can offer invaluable contextual information that can complement and enrich a BOW. In this specific credit-worthiness context, we examine the BOW token "worked." It appears three times (rows 5–7), twice in no-default loan texts and once in a defaulted loan text. This finding suggests that "worked" is being used to refer to the borrower's work ethic and may be a good predictor of credit worthiness. Digging deeper and looking at several trigrams (i.e., three-token sequences) involving "worked," we see that "have_worked_hard" appears in the two no-default loan related texts (referring to borrower accomplishments and plans) and "had_worked_harder" appears in the defaulted loan text (referring to what could have been done). This example illustrates how n-grams can provide richer contextualization capabilities for the creditworthiness prediction ML models.

Parts-of-speech tags can add value because they identify the composition of the texts. For example, POS provides information on whether the prospective borrowers are including many action words (verbs) or descriptors (adjectives) and whether this is being done differently in instances of no-default versus instances of defaulted loans.

MODEL TRAINING

6

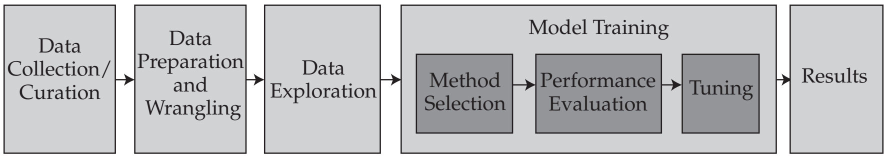
Machine learning model training is a systematic, iterative, and recursive process. The number of iterations required to reach optimum results depends on:

- the nature of the problem and input data and
- the level of model performance needed for practical application.

Machine learning models combine multiple principles and operations to provide predictions. As seen in the last two sections, typical ML model building requires data preparation and wrangling (cleansing and preprocessing) and data exploration (exploratory data analysis as well as feature selection and engineering). In addition, domain knowledge related to the nature of the data is required for good model building and training. For instance, knowledge of investment management and securities trading is important when using financial data to train a model for predicting costs of trading stocks. It is crucial for ML engineers and domain experts to work together in building and training robust ML models.

The three tasks of ML model training are method selection, performance evaluation, and tuning. Exhibit 20 outlines model training and its three component tasks. Method selection is the art and science of deciding which ML method(s) to incorporate and is guided by such considerations as the classification task, type of data, and size of data. Performance evaluation entails using an array of complementary techniques and measures to quantify and understand a model's performance. Tuning is the process of undertaking decisions and actions to improve model performance. These steps may be repeated multiple times until the desired level of ML model performance is attained. Although no standard rulebook for training an ML model exists, having a fundamental understanding of domain-specific training data and ML algorithm principles plays a vital role in good model training.

Exhibit 20 Model Training Stage



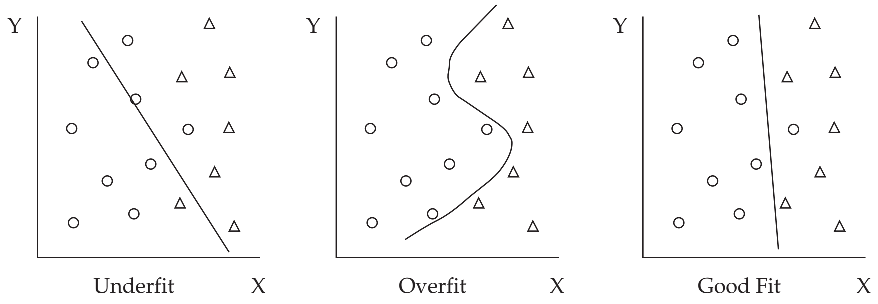
Before training a model, it is important to state the problem, define objectives, identify useful data points, and conceptualize the model. Conceptualization is like a blueprint on a drawing board, a modifiable plan that is necessary to initiate the model training process. Because modeling is an iterative process, many changes and refinements will be made to the model plan as the process evolves.

6.1 Structured and Unstructured Data

The ML model training process for structured and unstructured data is typically the same. Most ML models are intended to train on structured data, so unstructured data in the data preparation stage are processed and organized into a structured format. The systematic processing of unstructured text data so that they can be structured in the form of a data matrix has been previously covered. Similarly, other forms of unstructured data can also be prepared and formed into data matrixes or tables for ML training.

The fundamental idea of ML model training is fitting a system of rules on a training dataset to reveal a pattern in the data. In other words, fitting describes the degree to which (or how well) an ML model can be generalized to new data. A good model fit results in good model performance and can be validated using new data outside of the training dataset (i.e., out-of-sample). Exhibit 21 shows model decision boundaries in three possible model fitting scenarios for a classification task comprising two different classes of data (i.e., circles and triangles). The model on the left is underfit; it does not fit the training data well enough since it results in four misclassification errors (three circles and one triangle). Although the center model that generates the "S"-shaped line has the best accuracy (no errors) on the training data, it is overfit (i.e., fits the training data too well) and thus unlikely to perform well on future test cases. The model on the right (with one classification error, a circle) is a model with good fit (i.e., it fits the training data well but not so well that it cannot be generalized to out-of-sample data).

Exhibit 21 Model Fitting Scenarios: Underfit, Overfit, and Good Fit



Model fitting errors are caused by several factors—the main ones being dataset size and number of features in the dataset.

- *Dataset Size*: Small datasets can lead to underfitting of the model since small datasets often are not sufficient to expose patterns in the data. Restricted by a small dataset, an ML model may not recognize important patterns.
- *Number of Features*: A dataset with a small number of features can lead to underfitting, and a dataset with a large number of features can lead to overfitting. As with small dataset size, a small number of features may not carry all the characteristics that explain relationships between the target variable and the features. Conversely, a large number of features can complicate the model and potentially distort patterns in the data due to low degrees of freedom, causing overfitting. Therefore, appropriate feature selection using the types of techniques described earlier (e.g., chi-square, mutual information) is a key factor in minimizing such model overfitting.

Feature engineering tends to prevent underfitting in the training of the model. New features, when engineered properly, can elevate the underlying data points that better explain the interactions of features. Thus, feature engineering can be critical to overcome underfitting. Method-related factors that affect model fitting are explained shortly under tuning.

Method Selection

ML model training is a craft (part art and part science); it has no strict guidelines. Selecting and applying a method or an algorithm is the first step of the training process. Method selection is governed by the following factors:

1. *Supervised or unsupervised learning.* The data for training and testing supervised ML models contain **ground truth**, the known outcome (i.e., target variable) of each observation in these datasets. Unsupervised ML modeling is relatively challenging because of the absence of ground truth (i.e., no target variable). Supervised models bring a structure that may or may not be supported by the data. Unsupervised models bring no structure beyond that which arises from the given data. For supervised learning (with labeled training data), typical methods of choice are regression, ensemble trees, support vector machines (SVMs), and neural networks (NNs). Supervised learning would be used, for example, for default prediction based on high-yield corporate bond issuer data. For unsupervised learning, common methods are dimensionality

reduction, clustering, and anomaly detection. Unsupervised learning, for example, would be used for clustering financial institutions into different groups based on their financial attributes.

2 *Type of data*. For numerical data (e.g., predicting stock prices using historical stock market values), classification and regression tree (CART) methods may be suitable. For text data (for example, predicting the topic of a financial news article by reading the headline of the article), such methods as generalized linear models (GLMs) and SVMs are commonly used. For image data (e.g., identifying objects in a satellite image, such as tanker ships moving in and out of port), NNs and deep learning methods tend to perform better than others. For speech data (e.g., predicting financial sentiment from quarterly earnings' conference call recordings), deep learning methods can offer promising results.

3 *Size of data.* A typical dataset has two basic characteristics: number of instances (i.e., observations) and number of features. The combination of these two characteristics can govern which method is most suitable for model training. For instance, SVMs have been found to work well on "wider" datasets with 10,000 to 100,000 features and with fewer instances. Conversely, NNs often work better on "longer" datasets, where the number of instances is much larger than the number of features.

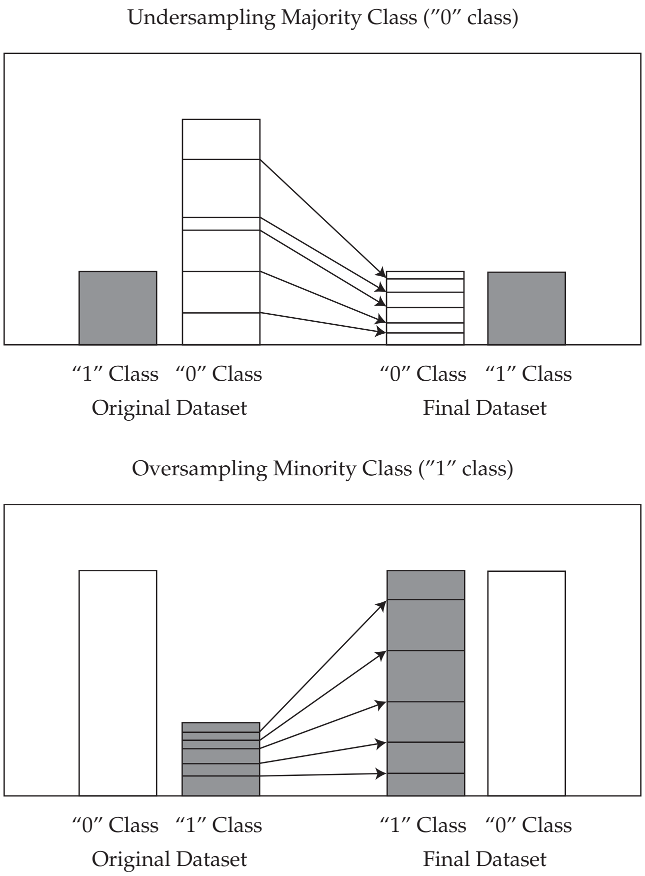
Once a method is selected, certain method-related decisions (e.g., on hyperparameters) need to be made. These decisions include the number of hidden layers in a neural network and the number of trees in ensemble methods (discussed later in the sub-section on tuning). In practice, datasets can be a combination of numerical and text data. To deal with mixed data, the results from more than one method can be combined. Sometimes, the predictions from one method can be used as predictors (features) by another. For example, unstructured financial text data can be used with logistic regression to classify stock sentiment as either positive or negative. Then, this sentiment classification cam be used as a predictor in a larger model, say CART, that also uses structured financial data as predictors for the purpose of stock selection. Finally, more than one method can be used and the results combined with quantitative or subjective weighing to exploit the advantages of each method.

Before model training begins, in the case of supervised learning the master dataset is split into three subsets used for model training and testing purposes. The first subset, a training set used to train the model, should constitute approximately 60% of the master dataset. The second subset, a cross-validation set (or validation set) used to tune and validate the model, should constitute approximately 20% of the master dataset. The third subset is a test set for testing the model and uses the remaining data. The data are split using a random sampling technique, such as the k-fold method. A commonly recommended split ratio is 60:20:20, as detailed above; however, the split percentages can vary. For unsupervised learning, no splitting is needed due to the absence of labeled training data.

Class imbalance, where the number of instances for a particular class is significantly larger than for other classes, may be a problem for data used in supervised learning because the ML classification method's objective is to train a high-accuracy model. In a high-yield bond default prediction example, say for corporate issuers in the BB+/Ba1 to B+/B1 credit quality range, issuers who defaulted (positive or "1" class) would be very few compared to issuers who did not default (negative or "0" class). Hence, on such training data, a naive model that simply assumes no corporate issuer will default may achieve good accuracy—albeit with all default cases misclassified. Balancing the training data can help alleviate such problems. In cases of unbalanced data, the "0" class (majority class) can be randomly undersampled or the "1" class (minority class) randomly oversampled. The random sampling can be done with or without replacement because they both work the same in general probability theory.

Exhibit 22 depicts the idea of undersampling of the majority class and oversampling of the minority class. In practice, the choice of whether to undersample or oversample depends on the specific problem context. Advanced techniques can also reproduce synthetic observations from the existing data, and the new observations can be added to the dataset to balance the minority class.

Exhibit 22 Undersampling and Oversampling



Performance Evaluation

It is important to measure the model training performance or goodness of fit for validation of the model. We shall cover several techniques to measure model performance that are well suited specifically for binary classification models.

1 *Error analysis.* For classification problems, error analysis involves computing four basic evaluation metrics: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) metrics. FP is also called a Type I error, and FN is also called a Type II error. Exhibit 23 shows a **confusion matrix**, a grid that is used to summarize values of these four metrics.

Exhibit 23 Confusion Matrix for Error Analysis

		Actual Training Labels	
		Class "1"	Class "0"
Predicted Results	Class "1"	True Positives (TP)	False Positives (FP) Type I Error
	Class "0"	False Negatives (FN) Type II Error	True Negatives (TN)

Additional metrics, such as precision and recall, can be computed. Assume in the following explanation that Class "0" is "not defective" and Class "1" is "defective." **Precision** is the ratio of correctly predicted positive classes to all predicted positive classes. Precision is useful in situations where the cost of FP, or Type I error, is high—or example, when an expensive product fails quality inspection (predicted Class "1") and is scrapped, but it is actually perfectly good (actual Class "0"). **Recall** (also known as *sensitivity*) is the ratio of correctly predicted positive classes to all actual positive classes. Recall is useful in situations where the cost of FN or Type II error is high—for example, when an expensive product passes quality inspection (predicted Class "0") and is sent to the valued customer, but it is actually quite defective (actual Class "1"). The formulas for precision and recall are:

$$\text{Precision (P)} = \text{TP}/(\text{TP} + \text{FP}). \quad (3)$$

$$\text{Recall (R)} = \text{TP}/(\text{TP} + \text{FN}). \quad (4)$$

Trading off precision and recall is subject to business decisions and model application. Therefore, additional evaluation metrics that provide the overall performance of the model are generally used. The two overall performance metrics are accuracy and F1 score. **Accuracy** is the percentage of correctly predicted classes out of total predictions. **F1 score** is the harmonic mean of precision and recall. F1 score is more appropriate (than accuracy) when unequal class distribution is in the dataset and it is necessary to measure the equilibrium of precision and recall. High scores on both of these metrics suggest good model performance. The formulas for accuracy and F1 score are as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}). \quad (5)$$

$$\text{F1 score} = (2 * \text{P} * \text{R})/(\text{P} + \text{R}). \quad (6)$$

Exhibit 24 illustrates computations of model evaluation metrics and performance scores on a sample dataset.

Exhibit 24 Performance Metrics and Scores Computation

Sample Dataset with Classification Results

Observation	Actual Training Labels	Predicted Results	Classification
1	1	1	TP
2	0	0	TN
3	1	1	TP
4	1	0	FN
5	1	1	TP
6	1	0	FN
7	0	0	TN
8	0	0	TN
9	0	0	TN
10	0	1	FP

Confusion Matrix

		Actual Training Labels	
		Class "1"	Class "0"
Predicted Results	Class "1"	3 (TP)	1 (FP)
	Class "0"	2 (FN)	4 (TN)

Performance Metrics

TP = 3, FP = 1, FN = 2, TN = 4

P = 3 / (3+1) = 0.75

R = 3 / (3+2) = 0.60

F1 Score = (2 × 0.75 × 0.60) / (0.75 + 0.60) = 0.67

Accuracy = (3 + 4) / (3 + 1 + 4 + 2) = 0.70

In Exhibit 24, if all "1" classes were predicted correctly (no FPs), the precision would have been equal to 1. If all "0" classes were predicted correctly (no FNs), the recall would have been equal to 1. Thus, the resulting F1 score would have been equal to 1. The precision of 0.75 and recall of 0.60 indicate that the model is better at minimizing FPs than FNs. To find the equilibrium between precision and recall, F1 score is calculated, which is equal to 0.67. The F1 score is closer to the smaller value among both precision and recall, giving the model a more appropriate score rather than just an arithmetic mean. Accuracy, the percentage of correct predictions (for both classes) made by the model, is equal to 0.70. Accuracy would be equal to 1 if all predictions were correct. As the number of "1" and "0" classes is equal in the dataset (i.e., a balanced dataset), accuracy can be considered an appropriate performance measure in this case. If the number of classes in a dataset is unequal; however, then F1 score should be used as the overall performance measure for the model.

2 *Receiver Operating Characteristic (ROC).* This technique for assessing model performance involves the plot of a curve showing the trade-off between the false positive rate (x-axis) and true positive rate (y-axis) for various cutoff points—for example, for the predicted probability (p) in a logistic regression. The formulas for false positive rate and true positive rate (note that true positive rate is the same as recall) are:

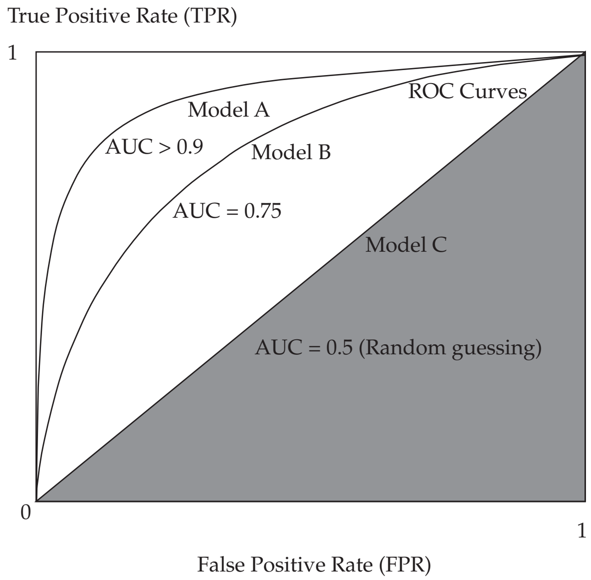
$$\text{False positive rate (FPR)} = \text{FP}/(\text{TN} + \text{FP}) \text{ and} \quad (7)$$

$$\text{True positive rate (TPR)} = \text{TP}/(\text{TP} + \text{FN}). \quad (8)$$

If p from a logistic regression model for a given observation is greater than the cutoff point (or threshold), then the observation is classified as class = 1. Otherwise, the observation will be classified as class = 0.

The shape of the ROC curve provides insight into the model's performance. A more convex curve indicates better model performance. Area under the curve (AUC) is the metric that measures the area under the ROC curve. An AUC close to 1.0 indicates near perfect prediction, while an AUC of 0.5 signifies random guessing. Exhibit 25 displays three ROC curves and indicates their respective AUC values. It is clear from observing the shapes of the ROC curves and their AUCs that Model A—with the most convex ROC curve with AUC of more than 0.9 (or 90%)—is the best performing among the three models.

Exhibit 25 ROC Curves and AUCs



3 *Root Mean Squared Error (RMSE).* This measure is appropriate for continuous data prediction and is mostly used for regression methods. It is a single metric that captures all the prediction errors in the data (n). The root mean squared error is computed by finding the square root of the mean of the squared differences between the actual values and the model's predicted values (error). A small RMSE indicates potentially better model performance. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n}\frac{\left(\text{Predicted}_i - \text{Actual}_i\right)^2}{n}} \quad (9)$$

Tuning

Once the model is evaluated, certain decisions and actions must be taken based on the findings to improve the performance of the model. If the prediction error on the training set is high, the model is underfitting. If the prediction error on the cross-validation (CV) set is significantly higher than on the training set, the model is overfitting. Model fitting has two types of error: bias and variance. Bias error is associated with underfitting, and variance error is associated with overfitting. Bias error is high when a model is overly simplified and does not sufficiently learn from the patterns in the training data. Variance error is high when the model is overly complicated and memorizes the training data so much that it will likely perform poorly on new data. It is not possible to completely eliminate both types of errors. However, both errors can be minimized so the total aggregate error (bias error + variance error) is at a minimum. The bias–variance trade-off is critical to finding an optimum balance where a model neither underfits nor overfits.

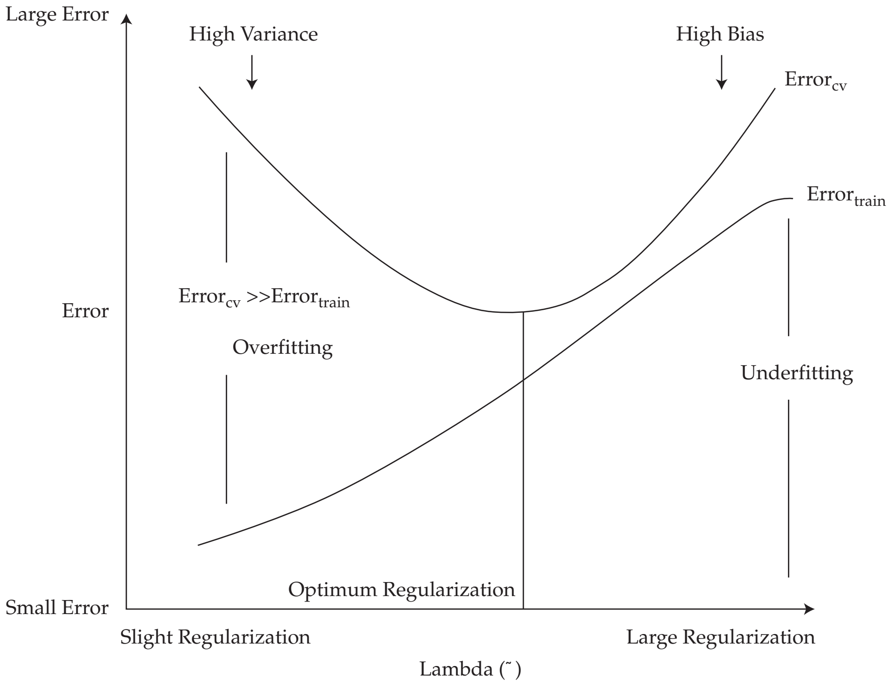
1 *Parameters* are critical for a model and are dependent on the training data. Parameters are learned from the training data as part of the training process by an optimization technique. Examples of parameters include coefficients in regression, weights in NN, and support vectors in SVM.

2 *Hyperparameters* are used for estimating model parameters and are not dependent on the training data. Examples of hyperparameters include the regularization term (λ) in supervised models, activation function and number of hidden layers in NN, number of trees and tree depth in ensemble methods, *k* in k-nearest neighbor classification and k-means clustering, and p-threshold in logistic regression. Hyperparameters are manually set and tuned.

For example, if a researcher is using a logistic regression model to classify sentences from financial statements into positive or negative stock sentiment, the initial cutoff point for the trained model might be a p-threshold of 0.50 (50%). Therefore, any sentence for which the model produces a probability >50% is classified as having positive sentiment. The researcher can create a confusion matrix from the classification results (of running the CV dataset) to determine such model performance metrics as accuracy and F1 score. Next, the researcher can vary the logistic regression's p-threshold—say to 0.55 (55%), 0.60 (60%), or even 0.65 (65%)—and then re-run the CV set, create new confusion matrixes from the new classification results, and compare accuracy and F1 scores. Ultimately, the researcher would select the logistic regression model with a p-threshold value that produces classification results generating the highest accuracy and F1 scores. Note that the process just outlined will be demonstrated in Section 7.

There is no general formula to estimate hyperparameters. Thus, tuning heuristics and such techniques as grid search are used to obtain the optimum values of hyperparameters. **Grid search** is a method of systematically training an ML model by using various combinations of hyperparameter values, cross validating each model, and determining which combination of hyperparameter values ensures the best model performance. The model is trained using different combinations of hyperparameter values until the optimum set of values are found. Optimum values must result in similar performance of the model on training and CV datasets, meaning that the training error and CV error are close. This ensures that the model can be generalized to test data or to new data and thus is less likely to overfit. The plot of training errors for each value of a hyperparameter (i.e., changing model complexity) is called a fitting curve. Fitting curves provide visual insight on the model's performance (for the given hyperparameter and level of model complexity) on the training and CV datasets and are visually helpful to tune hyperparameters. Exhibit 26 shows the bias–variance error trade-off by plotting a generic fitting curve for a regularization hyperparameter (λ).

Exhibit 26 Fitting Curve for Regularization Hyperparameter (λ)



Slight regularization lightly penalizes model complexity, thereby allowing most or all of the features to be included in the model and thus potentially enabling the model to "memorize" the data. Typically with no or slight regularization, the prediction error on the training dataset is small while the prediction error on the CV dataset is significantly larger. This difference in error is variance. High variance error, which typically results from too many features and model complexity, results in model overfitting. When high variance error and low bias error exist, the model performs well on the training dataset but generates many FP and FN errors on the CV dataset; in other words, the model is overfitted and does not generalize to new data well.

Large regularization excessively penalizes model complexity, thereby allowing too few of the features to be included in the model and causing the model to learn less from the data. The model may lack the necessary predictor variables and complexity needed to discern underlying patterns in the data. Typically with large regularization, the prediction errors on the training and CV datasets are both large. Large prediction errors on the training dataset indicate high bias, and high bias error results from model underfitting. When high bias error exists, the model does not perform well on either training or CV datasets because it is typically lacking important predictor variables.

Optimum regularization minimizes both variance and bias errors in a balanced fashion. It penalizes model complexity just enough so that only the most important features are included in the model. This process prevents the model from memorizing the data while enabling the model to learn enough from the data to distinguish important patterns. This results in prediction errors in both training and CV datasets that are similar and also minimal. The range of optimum regularization values can be found heuristically using such techniques as grid search.

If high bias or variance exists after the tuning of hyperparameters, either a larger number of training examples (instances) may be needed or the number of features included in the model may need to be decreased (in the case of high variance) or increased (in the case of high bias). The model then needs to be re-trained and re-tuned using the new training dataset. In the case of a complex model, where a large model is comprised of sub-model(s), ceiling analysis can be performed. **Ceiling analysis** is a systematic process of evaluating different components in the pipeline of model building. It helps to understand what part of the pipeline can potentially improve in performance by further tuning. For example, a stock market prediction model needs historical data from the stock market and perhaps news articles related to the stocks. The sub-model will extract relevant information from the news articles or classify the sentiment of the news articles. The results of the sub-model will feed into the larger model as features. Thus, the performance of the larger model depends on performance of the sub-model(s). Ceiling analysis can help determine which sub-model needs to be tuned to improve the overall accuracy of the larger model.

7 FINANCIAL FORECASTING PROJECT: CLASSIFYING AND PREDICTING SENTIMENT FOR STOCKS

Robo-readers are automated programs used to analyze large quantities of text, including news articles and social media. In the financial services space, robo-readers are being used by investors to examine how views expressed in text relate to future company performance. One important dimension that robo-readers look to analyze is sentiment polarity—which means how positive, negative, or neutral a particular phrase or statement is regarding a "target." For example, in the statement "XYZ Corporation is doing terrific things with its new product innovation," positive sentiment (i.e., the polarity) is being expressed regarding XYZ Corporation (i.e., the target of the sentiment). Such sentiment can provide invaluable predictive power, both alone and when coupled with structured financial data, for predicting stock price movements for individual firms and for portfolios of companies.

To provide a practical application, we use a financial forecasting project to examine how effectively sentiment—expressed in English news articles on LexisNexis (a searchable database of news articles) related to all companies listed on the NASDAQ OMX Helsinki (Finland)—can be classified. To accomplish this task, we followed the text ML model building steps presented in Sections 3 to 6 of this reading.

7.1 Text Curation, Preparation, and Wrangling

Text Curation

The text data used in this financial forecasting project are a collection of English language sentences from financial and economic news sources. The text data are acquired from the Financial Phrase Bank located on the website Researchgate.net.[2] The compressed folder contains six text files. The first two files are license and readme files. The other four files contain the text data. The data are presented in a text document format (.txt), which can be opened and viewed using any text editor. Note that this is cross-sectional data (not time series data).

A total of 14,780 sentences are in the four files. The sentiment of each sentence has already been labeled with one of three sentiment classes: positive, neutral, or negative. The sentiment classes are provided from an investor's perspective and may be useful for predicting whether a sentence may have a corresponding positive, neutral, or negative influence on the respective company's stock price.

2 https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10.

This project uses sentences from two of the text files (Sentences_AllAgree and Sentences_75Agree), labeled as either in the positive or negative sentiment class, for a total of 2,180 sentences. There are 1,457 positive sentiment class sentences and 723 negative sentiment class sentences. A supervised ML model is trained, validated, and tested using these data. The final ML model can be used to predict the sentiment classes of sentences present in similar financial news statements. Exhibit 27 shows a sample of 10 rows of raw text from the Sentences_AllAgree text file. Note the sentiment annotations at the end of each sentence with prefix character "@."

Exhibit 27 Ten Sample Sentences and Sentiment from Raw Text File (Sentences_AllAgree.txt)

```
Profit before taxes amounted to EUR 56.5 mn , down from EUR 232.9 mn a year ago .@negative
Profit before taxes decreased by 9 % to EUR 187.8 mn in the first nine months of 2008 , compared to EUR 207.1 mn a year earlier .@negative
Profit before taxes decreased to EUR 31.6 mn from EUR 50.0 mn the year before .@negative
Profit before taxes was EUR 4.0 mn , down from EUR 4.9 mn .@negative
The company 's profit before taxes fell to EUR 21.1 mn in the third quarter of 2008 , compared to EUR 35.8 mn in the corresponding period in 2007 .@negative
In August-October 2010 , the company 's result before taxes totalled EUR 9.6 mn , up from EUR 0.5 mn in the corresponding period in 2009 .@positive
Finnish Bore that is owned by the Rettig family has grown recently through the acquisition of smaller shipping companies .@positive
The plan is estimated to generate some EUR 5 million ( USD 6.5 m ) in cost savings on an annual basis .@positive
Finnish pharmaceuticals company Orion reports profit before taxes of EUR 70.0 mn in the third quarter of 2010 , up from EUR 54.9 mn in the corresponding period in 2009 .@positive
Finnish Sampo Bank , of Danish Danske Bank group , reports profit before taxes of EUR 152.3 mn in 2010 , up from EUR 32.7 mn in 2009 .@positive
```

Text Preparation (Cleansing)

The raw text data (i.e., sentences) are initially organized into a data table. The data table contains two columns: The first column (sentence) is for the text, and the second column (sentiment) is for the corresponding sentiment class. The separator character, which is "@" in this case, is used to split the data into text and sentiment class columns. A collection of text data in any form, including list, matrix, or data table forms, is called a **corpus**. Exhibit 28 shows a sample of 10 sentences from the data table corpus.

Exhibit 28 Ten Sample Rows of the Data Table (Corpus)

Sentence	Sentiment
Profit before taxes amounted to EUR 56.5 mn , down from EUR 232.9 mn a year ago .	negative
Profit before taxes decreased by 9 % to EUR 187.8 mn in the first nine months of 2008 , compared to EUR 207.1 mn a year earlier .	negative
Profit before taxes decreased to EUR 31.6 mn from EUR 50.0 mn the year before .	negative
Profit before taxes was EUR 4.0 mn , down from EUR 4.9 mn .	negative
The company 's profit before taxes fell to EUR 21.1 mn in the third quarter of 2008 , compared to EUR 35.8 mn in the corresponding period in 2007 .	negative
In August-October 2010 , the company 's result before taxes totalled EUR 9.6 mn , up from EUR 0.5 mn in the corresponding period in 2009 .	positive
Finnish Bore that is owned by the Rettig family has grown recently through the acquisition of smaller shipping companies .	positive
The plan is estimated to generate some EUR 5 million (USD 6.5 m) in cost savings on an annual basis .	positive
Finnish pharmaceuticals company Orion reports profit before taxes of EUR 70.0 mn in the third quarter of 2010 , up from EUR 54.9 mn in the corresponding period in 2009 .	positive
Finnish Sampo Bank , of Danish Danske Bank group , reports profit before taxes of EUR 152.3 mn in 2010 , up from EUR 32.7 mn in 2009 .	positive

The raw text contains punctuations, numbers, and white spaces that may not be necessary for model training. Text cleansing involves removing, or incorporating appropriate substitutions for, potentially extraneous information present in the text. Operations to remove html tags are unnecessary because none are present in the text

Punctuations: Before stripping out punctuations, percentage and dollar symbols are substituted with word annotations to retain their essence in the financial texts. Such word annotation substitutions convey that percentage and currency-related tokens were involved in the text. As the sentences have already been identified within and extracted from the source text, punctuation helpful for identifying discrete sentences—such as periods, semi-colons, and commas—are removed. Some special characters, such as "+" and "©," are also removed. It is a good practice to implement word annotation substitutions before removing the rest of the punctuations.

Numbers: Numerical values of numbers in the text have no significant utility for sentiment prediction in this project because sentiment primarily depends on the words in a sentence. Here is an example sentence: "*Ragutis, which is based in Lithuania's second-largest city, Kaunas, boosted its sales last year 22.3 percent to 36.4 million litas*." The word "boosted" implies that there was growth in sales, so analysis of this sentiment does not need to rely on interpretation of numerical text data. Sentiment analysis typically does not involve extracting, interpreting, and calculating relevant numbers but instead seeks to understand the context in which the numbers are used. Other commonly occurring numbers are dates and years, which are also not required to predict sentence sentiment. Thus, all numbers present in the text are removed for this financial sentiment project. However, prior to removing numbers, abbreviations representing orders of magnitude, such as million (commonly represented by "m," "mln," or "mn"), billion, or trillion, are replaced with the complete word. Retaining these orders of magnitude-identifying words in the text preserves the original text meaning and can be useful in predicting sentence sentiment.

Whitespaces: White spaces are present in the raw text. Additional white spaces occur after performing the above operations to remove extraneous characters. The white spaces must be removed to keep the text intact. Exhibit 29 shows the sample text after cleansing. The cleansed text is free of punctuations and numbers, with useful substitutions.

Exhibit 29 Ten Sample Rows After Cleansing Process

Sentence	Sentiment
Profit before taxes amounted to EUR million down from EUR million a year ago	negative
Profit before taxes decreased by percentSign to EUR million in the first nine months of compared to EUR million a year earlier	negative
Profit before taxes decreased to EUR million from EUR million the year before	negative
Profit before taxes was EUR million down from EUR million	negative
The companys profit before taxes fell to EUR million in the third quarter of compared to EUR million in the corresponding period in	negative
In August October the companys result before taxes totalled EUR million up from EUR million in the corresponding period in	positive
Finnish Bore that is owned by the Rettig family has grown recently through the acquisition of smaller shipping companies	positive
The plan is estimated to generate some EUR million USD million in cost savings on an annual basis	positive

(continued)

Exhibit 29 (Continued)

Sentence	Sentiment
Finnish pharmaceuticals company Orion reports profit before taxes of EUR million in the third quarter of up from EUR million in the corresponding period in	positive
Finnish Sampo Bank of Danish Danske Bank group reports profit before taxes of EUR million in up from EUR million in	positive

Text Wrangling (Preprocessing)

The cleansed text needs to be normalized using the following normalization procedures:

1 *Lowercasing* of all text to consolidate duplicate words (example, "THE," "The," and "the").

2 *Stop words* are not removed because some stop words (e.g., not, more, very, and few) carry significant meaning in the financial texts that is useful for sentiment prediction. Some stop words, such as articles (a, an, the), may be removed. Nevertheless, to avoid confusion no words are removed at this point. This issue will be revisited during the data exploration stage, which will carefully examine the text using frequency analysis and find custom stop words (common words) for these particular text data.

3 *Stemming*, the converting of inflected forms of a word into its base word (stem), is performed on the text as it is simple to perform and is appropriate for training an ML model for sentiment prediction.

White spaces are stripped after performing these operations. As part of text normalization, different currency abbreviations, such as EUR and USD, can be converted into a single token, such as "currencysign." As we are dealing with financial domain text, the earlier substitution of dollarsign can be replaced with currencysign as well. This step will remove tokens that are different but redundant in nature while maintaining their meaning. Through careful examination of the text and use of domain knowledge, similar substitutions of redundant tokens can be performed. Exhibit 30 shows how the sample text appears after normalization.

Exhibit 30 Ten Sample Rows After Normalization Process

Sentence	Sentiment
profit befor tax amount to currencysign million down from currencysign million a year ago	negative
profit befor tax decreas by percentsign to currencysign million in the first nine month of compar to currencysign million a year earlier	negative
profit before tax decreas to currencysign million from currencysign million the year befor	negative
profit befor tax was currencysign million down from currencysign million	negative
the compani profit befor tax fell to currencysign million in the third quarter of compar to currencysign million in the correspond period in	negative
in august octob the compani result befor tax total currencysign million up from currencysign million in the correspond period in	positive
finnish bore that is own by the rettig famili has grown recent through the acquisit of smaller shipping company	positive
the plan is estim to generat some currencysign million currencysign million in cost save on an annual basi	positive

Exhibit 30 (Continued)

Sentence	Sentiment
finnish pharmaceut compani orion report profit befor tax of currencysign million in the third quarter of up from currencysign million in the correspond period in	positive
finnish sampo bank of danish danske bank group report profit befor tax of currencysign million in up from currencysign million in	positive

The normalized text is tokenized, resulting in 2,673 unique tokens. Altogether, these unique tokens comprise the bag-of-words (BOW) of the text corpus. Exhibit 31 shows a sample of 100 tokens from the BOW. This preliminary unigram BOW can be used to construct a document term matrix (DTM) for ML training.

Exhibit 31 One Hundred Sample Tokens from Preliminary Unigram BOW

"for"	"foundri"	"quarter"	"shop"	"net"	"share"	"to"
"currencysign"	"nokia"	"same"	"plan"	"year"	"sanyo"	"it"
"move"	"nokian"	"tax"	"earn"	"in"	"expect"	"by"
"percentsign"	"director"	"rose"	"dividned"	"total"	"megafon"	"talentum"
"report"	"as"	"chain"	"number"	"consolid"	"accord"	"compar"
"prior"	"last"	"machin"	"componenta"	"afx"	"doubl"	"higher"
"led"	"from"	"announc"	"a"	"with"	"while"	"g"
"handset"	"pre"	"fourth"	"loss"	"analyst"	"increas"	"said"
"board"	"oper"	"propos"	"repres"	"paid"	"finnish"	"base"
"user"	"retail"	"market"	"is"	"late"	"amount"	"estim"
"the"	"divis"	"of"	"helsinki"	"sale"	"close"	
"million"	"after"	"period"	"team"	"earlier"	"manufactur"	
"zero"	"tyre"	"profit"	"beat"	"third"	"dealer"	
"and"	"will"	"correspond"	"per"	"up"	"subscrib"	
"cloth"	"decemb"	"sepp"	"custom"	"reach"	"teliasonera"	

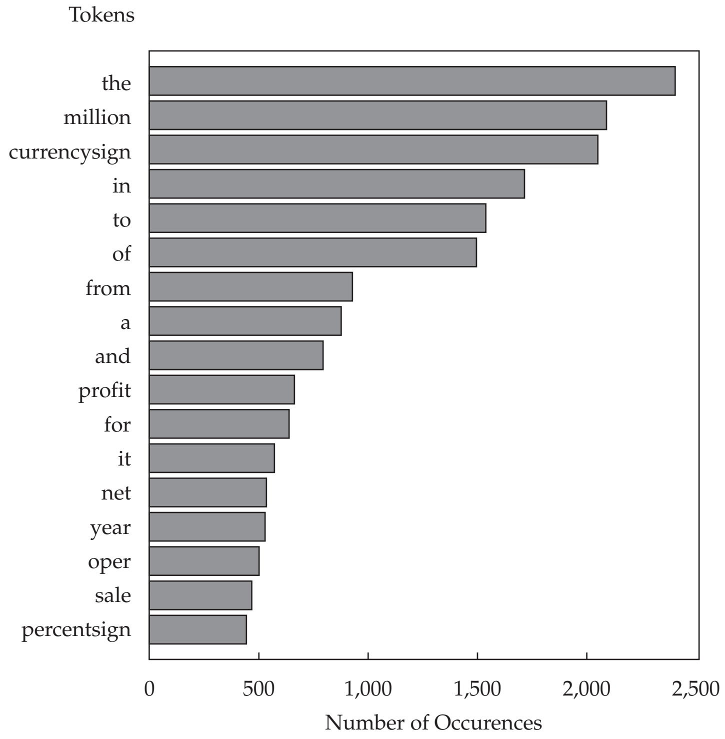
The final DTM for ML model training will be prepared after the data exploration stage. Data exploration may reveal unnecessary tokens or anomalies in the data. Any unnecessary tokens that are not informative must be removed, which will also impact the creation of n-grams. Thus, the final DTM must be made after further analyses and operations, such as exploratory data analysis and feature selection.

7.2 Data Exploration

Exploratory Data Analysis

Exploratory data analysis (EDA) performed on text data provides insights on word distribution in the text. Word counts from all the sentences are computed. These word counts can be used to examine outlier tokens—words that are most commonly and least commonly present in the texts. The most frequent word occurrences in all sentences from the dataset are shown in Exhibit 32. These common words will be removed during the feature selection step. Notably, the tokens "million" and "currencysign" occur frequently due to the financial nature of the data.

Exhibit 32 Most Frequently Used Tokens in the Corpus



The most frequent word occurrences in the sentences in the negative sentiment and the positive sentiment classes are shown in Exhibit 33. The most commonly occurring words are similar for both sentiment classes, meaning that they are not useful in discriminating between the two sentiment classes. This finding demonstrates the utility of removing the most commonly used tokens from the BOW.

Exhibit 33 Most Frequently Used Tokens in Two Sentiment Classes of the Corpus

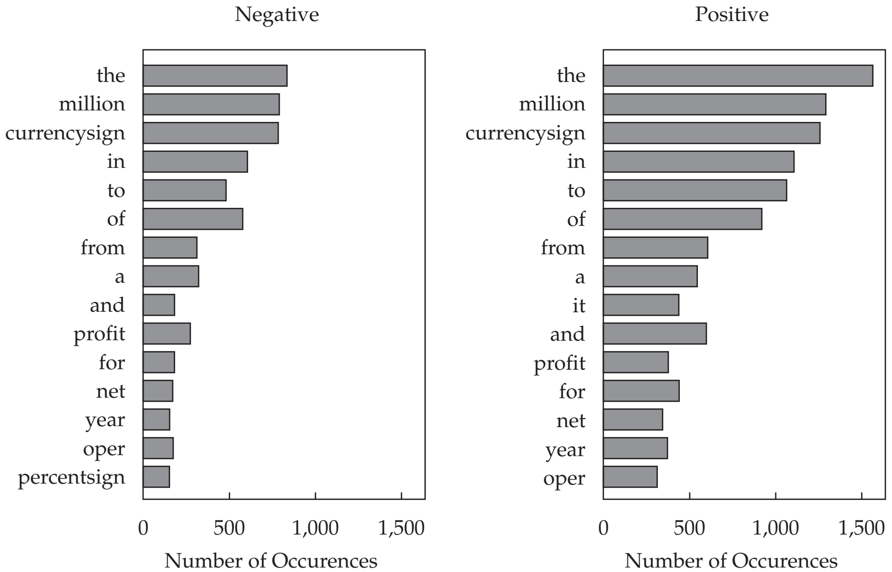
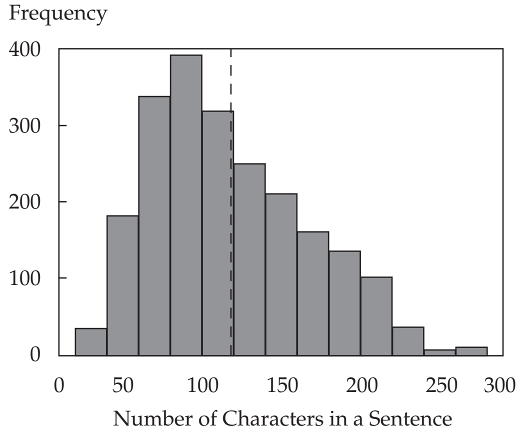


Exhibit 34 shows a histogram of sentence length distribution. **Sentence length** is defined as the number of characters, including spaces, in a sentence. The longest sentence has 273 characters; the shortest sentence has 26 characters; and the average number of characters is about 120 (indicated by the vertical line). Although this distribution does not have any direct impact on model training, this histogram visually demonstrates the range of sentence lengths and helps identify any extremely long or short sentences. This histogram does not appear unusual, so no outlier sentences need to be removed.

Exhibit 34 Histogram of Sentence Lengths with Mean Sentence Length



Word clouds are a convenient method of visualizing the text data because they enable rapid comprehension of a large number of tokens and their corresponding weights. Exhibit 35 shows a word cloud for all the sentences in the corpus. The font sizes of the words are proportionate to the number of occurrences of each word in the corpus. Similarly, Exhibit 36 shows the word cloud divided into two halves: one half representing negative sentiment class sentences (upper half); one half representing positive sentiment class sentences (lower half). Notably, some highly discriminative stems and words, such as “decreas” and “down” in the negative half and “increas” and “rose” in the positive half, are present. The feature selection process will eliminate common words and highlight useful words for better model training.

Exhibit 35 Word Cloud of Entire Corpus

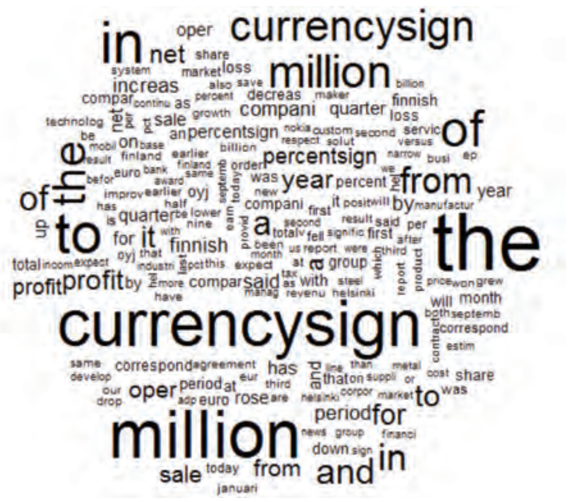


Exhibit 36 Word Cloud Divided by Two Sub-Groups of the Corpus

Feature Selection

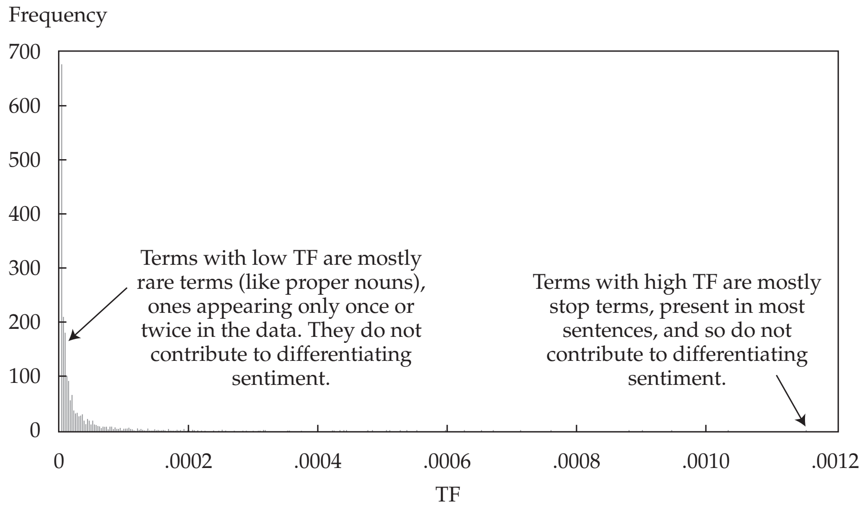
Exploratory data analysis revealed the most frequent tokens in the texts that could potentially add noise to this ML model training process. In addition to common tokens, many rarely occurring tokens, often proper nouns (i.e., names), are not informative for understanding the sentiment of the sentence. Further analyses must be conducted to decide which words to eliminate. Feature selection for text data involves keeping the useful tokens in the BOW that are informative and help to discriminate different classes of texts—those with positive sentiment and those with negative sentiment. At this point, a total of 44,151 non-unique tokens are in the 2,180 sentences.

Frequency analysis on the processed text data helps in filtering unnecessary tokens (or features) by quantifying how important tokens are in a sentence and in the corpus as a whole. Term frequency (TF) at the corpus level—also known as **collection frequency (CF)**—is the number of times a given word appears in the whole corpus (i.e., collection of sentences) divided by the total number of words in the corpus. Term frequency can be calculated and examined to identify outlier words. Exhibit 37 shows the descriptive statistics of term frequency for the words at the collection level. The statistics of TF range between 0 and 1 because TF values are ratios of total occurrences of a particular word to total number of words in the collection. A sample of words with the highest TF and lowest TF values is also shown to gain insight into what kinds of words occur at these extreme frequencies.

Exhibit 37 Summary Statistics of TF for Words at the Collection Level, Sample Words with High and Low TF Values, and Histogram of TF Values

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.265e-05	2.265e-05	4.530e-05	3.741e-04	1.585e-04	5.429e-02

word	TF	word	TF
<chr>	<dbl>	<chr>	<dbl>
the	0.05429096	yet	2.264954e-05
million	0.04722430	yihn	2.264954e-05
currencysign	0.04627302	young	2.264954e-05
in	0.03870807	zahariev	2.264954e-05
to	0.03476705	zone	2.264954e-05
of	0.03377047	zoo	2.264954e-05



Calculating highest and lowest TFs at the collection level is a general strategy to identify noisy terms. The histogram in Exhibit 37 shows a long tail to the right, which represents common terms that must be removed. The high frequency bars on the left show that there are also many rare terms (e.g., ones appearing only once or twice across the data). Such rare terms do not appear enough to be used as meaningful features and are often removed. The words with the highest TF are mostly stop words that are not useful because they are present in most of the sentences and thus do not contribute to differentiating the sentiment embedded in the text. The words with the lowest TF values are mostly proper nouns or sparse terms that are also not important to the meaning of the text. In this example, after careful examination of words with extreme frequencies, the words with high TF values (>99.5th percentile, 14 words) and low TF values (<30th percentile, 714 words) are removed before forming the final document term matrix (DTM). Exhibit 38 shows the 14 words with the highest TF values (>99.5th percentile) that are the custom stop words for this project.

Exhibit 38 Fourteen Custom Stop Words for the Project

"the"	"million"	"currencysign"	"in"	"to"	"of"	"from"
"and"	"profit"	"for"	"it"	"not"	"year"	"a"

To construct a DTM for ML training, different TF measures need to be computed to fill in the cells of the DTM. Exhibit 39 displays part of a TF measures table that is computed for the text data before the removal of custom stop words.

Exhibit 39 Sample Output of Term Frequency (TF) Measures Table

SentenceNo	TotalWordsInSentence	Word	TotalWordCount	WordCountInSentence	SentenceCountWithWord	TF	DF	IDF	TFIDF
<int>	<int>	<chr>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
624	34	a	873	6	687	0.1764706	0.3151376	1.1547459	0.20377868
701	39	the	2397	6	1453	0.1538462	0.6665138	0.4056945	0.06241454
1826	34	a	873	6	687	0.1764706	0.3151376	1.1547459	0.20377868
1963	39	the	2397	6	1453	0.1538462	0.6665138	0.4056945	0.06241454
128	30	of	1491	5	984	0.1666667	0.4513761	0.7954543	0.13257571
223	37	the	2397	5	1453	0.1351351	0.6665138	0.4056945	0.05482358

The columns of the term frequency measures table are as follows:

1 *SentenceNo*: A unique identification number assigned to each sentence in the order they are present in the original dataset. For example, sentence number 701 is a sentence in row 701 from the data table: "*the airlin estim that the cancel of it flight due to the closur of european airspac and the process of recommenc traffic have caus a the compani a loss of currencysign million includ the cost of strand passeng accommod*."

2 *TotalWordsInSentence*: Count of total number of words present in the sentence. For example, sentence number 701 has a total of 39 words.

3 *Word*: A word token that is present in the corresponding sentence.

4 *TotalWordCount*: Total number of occurrences of the word in the entire corpus or collection. For example, the token "the" occurs 2,397 times in the whole collection of sentences. The following equation can be used to compute TF at the collection level:

TF (Collection Level) =
TotalWordCount/Total number of words in collection. **(10)**

The TF of the word "the" at the collection level is calculated as 2,397/44,151 = 0.05429096. Note that this result was seen previously in Exhibit 37.

5 *WordCountInSentence*: Number of times the token is present in the corresponding sentence. For example, token "the" is present six times in sentence number 701.

6 *SentenceCountWithWord*: Number of sentences in which the word is present. For example, the token "the" is present in 1,453 sentences.

7 *TF (Term Frequency) at Sentence Level*: Number of times a word is present in a sentence divided by the total number of words in that sentence. The following equation can be used to compute TF at the sentence level:

TF (Sentence Level) =
WordCountInSentence/TotalWordsInSentence. **(11)**

For example, TF at the sentence level for the word "the" in sentences number 701 and 223 is calculated as 6/39 = 0.1538462 and 5/37 = 0.1351351, respectively.

8 *DF (Document Frequency)*: Defined as the number of documents (i.e., sentences) that contain a given word divided by the total number of sentences (here, 2,180). Document frequency is important since words frequently occurring across sentences provide no differentiating information in each sentence. The following equation can be used to compute DF:

$$\text{DF} = \text{SentenceCountWithWord/Total number of sentences.} \quad (12)$$

For example, DF of the word "the" is 1,453/2,180 = 0.6665138; so, 66.7% of the sentences contain the word "the." A high DF indicates high word frequency in the text.

9 *IDF (Inverse Document Frequency)*: A relative measure of how unique a term is across the entire corpus. Its meaning is not directly related to the size of the corpus. The following equation can be used to compute IDF:

$$\text{IDF} = \log(1/\text{DF}). \quad (13)$$

For example, IDF of the word "the" is log(1/0.6665138) = 0.4056945. A low IDF indicates high word frequency in the text.

10 *TF–IDF*: To get a complete representation of the value of each word, TF at the *sentence level* is multiplied by the IDF of a word across the entire dataset. Higher TF–IDF values indicate words that appear more frequently within a smaller number of documents. This signifies relatively more unique terms that are important. Conversely, a low TF–IDF value indicates terms that appear in many documents. TF–IDF values can be useful in measuring the key terms across a compilation of documents and can serve as word feature values for training an ML model. The following equation can be used to compute TF–IDF:

$$\text{TF–IDF} = \text{TF} \times \text{IDF}. \quad (14)$$

For example, TF–IDF of the token "of" is calculated as 0.1666667 × 0.7954543 = 0.13257571.

Similarly, Exhibit 40 shows high TF–IDF words for the text data before the removal of custom stop words.

Exhibit 40 Sample Output of High TF–IDF Words

SentenceNo	TotalWordsInSentence	Word	TotalWordCount	WordCountInSentence	SentenceCountWithWord	TF	DF	IDF	TFIDF
<int>	<int>	<chr>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
28	7	risen	3	1	3	0.1428571	0.0013761468	6.588468	0.9412097
830	7	diminish	2	1	2	0.1428571	0.0009174312	6.993933	0.9991333
1368	9	great	4	1	4	0.1111111	0.0018348624	6.300786	0.7000873
1848	8	injuri	1	1	1	0.1250000	0.0004587156	7.687080	0.9608850
1912	7	cheaper	1	1	1	0.1428571	0.0004587156	7.687080	1.0981543
1952	6	argument	1	1	1	0.1666667	0.0004587156	7.687080	1.2811800

TF or TF–IDF values are placed at the intersection of sentences (rows) and terms (columns) of the document term matrix. For this project, TF values are used for the DTM as the texts are sentences rather than paragraphs or other larger bodies of text. TF–IDF values vary by the *number* of documents in the dataset; therefore, the model performance can vary when applied to a dataset with just a few documents. In addition to removing custom stop words and sparse terms, single character letters are also eliminated because they do not add any value to the sentiment significance.

Feature Engineering

N-grams are used as a feature engineering process in this project. Use of n-grams helps to understand the sentiment of a sentence as a whole. As mentioned previously, the objective of this project is to predict sentiment class (positive and negative) from financial texts. Both unigram and bigrams are implemented, and the BOW is created from them. Bigram tokens are helpful for keeping negations intact in the text, which is vital for sentiment prediction. For example, the tokens "not" and "good" or "no" and "longer" can be formed into single tokens, now bigrams, such as "not_good" and "no_longer." These and similar tokens can be useful during ML model training and can improve model performance. Exhibit 41 shows a sample of 100 words from the BOW containing both unigram and bigram tokens after removal of custom stop words, sparse terms, and single characters. Note that the BOW contains such tokens as increas, loss, loss_prior, oper_rose, tax_loss, and sale_increas. Such tokens are informative about the embedded sentiment in the texts and are useful for training an ML model. The corresponding word frequency measures for the document term matrix are computed based on this new BOW.

Exhibit 41 One-Hundred Sample Tokens from Final BOW of Entire Corpus

"last"	"last_quarter"	"quarter"	"quarter_componenta"	"componenta"
"componenta_sale"	"sale"	"sale_doubl"	"doubl"	"doubl_same"
"same"	"same_period"	"period"	"period_earlier"	"earlier"
"earlier_while"	"while"	"while_move"	"move"	"move_zero"
"zero"	"zero_pre"	"pre"	"pre_tax"	"tax"
"tax_pre"	"tax_loss"	"loss"	"third"	"third_quarter"
"quarter_sale"	"sale_increas"	"increas"	"increas_by"	"by"
"by_percentsign"	"percentsign"	"percentsign_oper"	"oper"	"oper_by"
"oper_rose"	"rose"	"rose_correspond"	"correspond"	"correspond_period"
"period_repres"	"repres"	"repres_percentsign"	"percentsign_sale"	"oper_total"
"total"	"total_up"	"up"	"up_repres"	"finnish"
"finnish_talentum"	"talentum"	"talentum_report"	"report"	"report_oper"
"oper_increas"	"increas_sale"	"sale_total"	"cloth"	"cloth_retail"
"retail"	"retail_chain"	"chain"	"chain_sepp"	"sepp"
"sepp_ls"	"ls"	"ls_sale"	"consolid"	"consolid_sale"
"incres_percentsign"	"percentsign_reach"	"reach"	"reach_while"	"while_oper"
"oper_amount"	"amount"	"amount_compar"	"compar"	"compar_loss"
"loss_prior"	"prior"	"prior_period"	"foundri"	"foundri_divis"
"divis"	"divis_report"	"report_sale"	"percentsign_correspond"	"period_sale"
"sale_machin"	"machin"	"machin_shop"	"shop"	"shop_divis"

EXAMPLE 6

Calculating and Interpreting Term Frequency Measures

Data scientists Jack and Jill are using financial text data to develop sentiment indicators for forecasting future stock price movements. They have assembled a BOW from the corpus of text being examined and have pulled the following abbreviated term frequency measures tables.

Term Frequency Measures Table 1

SentenceNo	TotalWordsInSentence	Word	TotalWordCount	WordCountInSentence	SentenceCountWithWord
<int>	<int>	<chr>	<int>	<int>	<int>
624	34	a	873	6	687
701	39	the	2397	6	1453
1826	34	a	873	6	687
1963	39	the	2397	6	1453
128	30	of	1491	5	984
223	37	the	2397	5	1453

Term Frequency Measures Table 2

SentenceNo	TotalWordsInSentence	Word	TotalWordCount	WordCountInSentence	SentenceCountWithWord
<int>	<int>	<chr>	<int>	<int>	<int>
28	7	risen	3	1	3
830	7	diminish	2	1	2
1368	9	great	4	1	4
1848	8	injuri	1	1	1
1912	7	cheaper	1	1	1
1952	6	argument	1	1	1

1 Determine and interpret term frequency (TF) at the collection level and at the sentence level for the word (i.e., token) “a” in sentence 1,826 in term frequency measures Table 1 and then for the token “great” in sentence 1,368 in term frequency measures Table 2.

2 Determine and interpret TF–IDF (term frequency–inverse document frequency) for the word “a” in sentence 1,826 in term frequency measures Table 1 and then for the token “great” in sentence 1,368 in term frequency measures Table 2.

Solution to 1:

TF at the collection level is calculated using Equation 10:

TF (Collection Level) = TotalWordCount/Total number of words in collection.

For token “a” in sentence 1,826 (Table 1), TF (Collection Level) is 873/44,151 = 0.019773 or 1.977%.For token “great” in sentence 1,368 (Table 2), TF (Collection Level) is 4/44,151 = 0.000091 or 0.009%.TF at the collection level is an indicator of the frequency, in percentage terms, that a token is used throughout the whole collection of texts (here, 44,151). It is useful for identifying outlier words: Tokens with highest TF values are mostly stop words that do not contribute to differentiating the sentiment embedded in the text (such as “a”), and tokens with lowest TF values are mostly proper nouns or sparse terms that are also not important to the meaning of the text. Conversely, tokens with intermediate TF values potentially carry important information useful for differentiating the sentiment embedded in the text.TF at the sentence level is calculated using Equation 11:

TF (Sentence Level) = WordCountInSentence/TotalWordsInSentence.

For token “a” in sentence 1,826, TF (Sentence Level) is 6/34 = 0.176471 or 17.647%.

For token “great” in sentence 1,368, TF (Sentence Level) is 1/9 = 0.111111 or 11.111%.

TF at the sentence level is an indicator of the frequency, in percentage terms, that a token is used in a particular sentence (i.e., instance). Therefore, it is useful for understanding the importance of the specific token in a given sentence.

Solution to 2:

To calculate TF–IDF, besides TF at the sentence level, document frequency (DF) and inverse document frequency (IDF) are also required.

DF is the number of documents (i.e., sentences) that contain a given word divided by the total number of sentences in the corpus (here, 2,180). DF is calculated using Equation 12:

DF = SentenceCountWithWord/Total number of sentences.

For token "a" in sentence 1,826, DF is 687/2,180 = 0.315138 or 31.514%.

For token "great" in sentence 1,368, DF is 4/2,180 = 0.001835 or 0.184%.

Document frequency is important since tokens occurring frequently across sentences (such as "a") provide no differentiating information in each sentence. Tokens occurring less frequently across sentences (such as "great"), however, may provide useful differentiating information.

IDF is a relative measure of how important a term is across the entire corpus (i.e., collection of texts/sentences). IDF is calculated using Equation 13:

IDF = log(1/DF).

For token "a" in sentence 1,826, IDF is log(1/0.315138) = 1.154746.

For token "great" in sentence 1,368, IDF is log(1/0.001835) = 6.300786.

Using TF and IDF, TF–IDF can now be calculated using Equation 14:

TF–IDF = TF × IDF.

For token "a" in sentence 1,826, TF–IDF = 0.176471 × 1.154746 = 0.203779, or 20.378%.

For token "great" in sentence 1,368, TF–IDF = 0.111111 × 6.300786 = 0.700087, or 70.009%.

As TF–IDF combines TF at the *sentence level* with IDF across the entire corpus, it provides a complete representation of the value of each word. A high TF–IDF value indicates the word appears many times within a small number of documents, signifying an important yet unique term within a sentence (such as "great"). A low TF–IDF value indicates tokens that appear in most of the sentences and are not discriminative (such as "a"). TF–IDF values are useful in extracting the key terms in a document for use as features for training an ML model.

7.3 Model Training

The sentiment class labels (positive and negative) constitute the target variable (y) for model training. They are relabeled as 1 (for positive) and 0 (for negative) to enable calculating the performance metrics, such as receiver operating characteristic (ROC) curve and area under the curve (AUC) from the trained model results. The master dataset that has been cleansed and preprocessed is partitioned into three separate sets: 1) training set; 2) cross-validation (CV) set; and 3) test set. These are in the ratio of 60:20:20, respectively (following common practice). For splitting, simple random sampling is applied within levels of the target variable to balance the class distributions within the splits. The final DTM is built using the sentences (rows), which are the instances, and resulting tokens (columns), which are the feature variables, from the BOW of the training dataset. The final BOW consists of unigram and bigram tokens from the sentences in the training corpus only. The DTM is then filled in with resultant TF values of the tokens from the training corpus.

Similarly, the DTMs for the CV set and the test set are built using tokens from the final training BOW for tuning, validating, and testing of the model. To be clear, the final BOW from the training corpus is used for building DTMs across all the splits because the model has been trained on that final BOW. Thus, the columns (think, features) of all three DTMs are the same, but the number of rows varies because a different number of sentences are in each split. The DTMs are filled with resultant term frequency values calculated using sentences in the corpuses of the respective splits—sentences from the CV set corpus and sentences from the test set corpus. Exhibit 42 tabulates the summary of dimensions of the data splits and their uses in the model training process. As mentioned, the columns of DTMs for the splits are the same, equal to the number of unique tokens (i.e., features) from the final training corpus BOW, which is 9,188. Note that this number of unique tokens (9,188) differs from that in the master corpus (11,501) based on the sentences that are included in the training corpus after the random sampling.

Exhibit 42 Summary of the Three Data Splits

Corpus	Split %	Number of Sentences	DTM Dimensions	Purpose
Master	100%	2180	2180 × 11501	Used for data exploration
Training	60%	1309	1309 × 9188	Used for ML model training
CV	20%	435	435 × 9188	Used for tuning and validating the trained model
Test	20%	436	436 × 9188	Used for testing the trained, tuned, and validated model

Method Selection

Alternative ML methods, including SVM, decision trees, and logistic regression, were examined because these techniques are all considered potentially suitable for this particular task (i.e., supervised learning), type of data (i.e., text), and size of data (i.e., wider data with many potential variables). The SVM and logistic regression methods appeared to offer better performance than decision trees. For brevity, we discuss logistic regression in the remainder of the chapter. Logistic regression was used to train the model, using the training corpus DTM containing 1,309 sentences. As a reminder, in this project texts are the sentences and the classifications are positive and negative sentiment classes (labeled 1 and 0, respectively). The tokens are feature variables, and the sentiment class is the target variable. Text data typically contain thousands of tokens. These result in sparse DTMs because each column represents a token feature and the values are mostly zeros (i.e., not all the tokens are present in every text). Logistic regression can deal with such sparse training data because the regression coefficients will be close to zero for tokens that are not present in a significant number of sentences. This allows the model to ignore a large number of minimally useful features. Regularization further helps lower the coefficients when the features rarely occur and do not contribute to the model training.

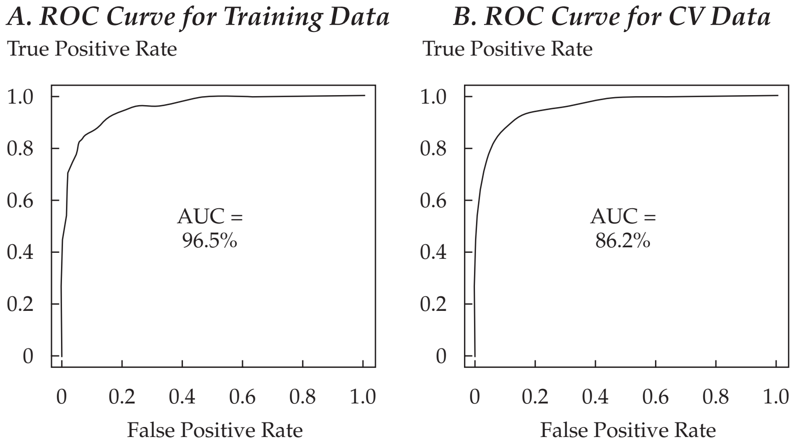
Logistic regression is applied on the final training DTM for model training. As this method uses maximum likelihood estimation, the output of the logistic model is a probability value ranging from 0 to 1. However, because the target variable is binary, coefficients from the logistic regression model are not directly used to predict the value of the target variable. Rather, a mathematical function uses the logistic regression coefficient (β) to calculate probability (p) of sentences having positive sentiment (y =

1).[3] If p for a sentence is 0.90, there is a 90% likelihood that the sentence has positive sentiment. Theoretically, the sentences with p > 0.50 likely have positive sentiment. Because this is not always true in practice, however, it is important to find an ideal threshold value of p. We elaborate on this point in a subsequent example. The threshold value is a cutoff point for p values, and the ideal threshold p value is influenced by the dataset and model training. When the p values (i.e., probability of sentences having positive sentiment) of sentences are above this ideal threshold p value, then the sentences are *highly* likely to have positive sentiment (y = 1). The ideal threshold p value is estimated heuristically using performance metrics and ROC curves, as will be demonstrated shortly.

Performance Evaluation and Tuning

The trained ML model is used to predict the sentiments of the sentences in the training and CV DTMs. Exhibit 43 displays the ROC curves for the training (Panel A) and CV (Panel B) data. Remember that the x-axis is false positive rate, FP/(TN + FP), and the y-axis is true positive rate, TP/(TP + FN). As the model is trained using the training DTM, it clearly performs well on the same training data (so there is no concern about underfitting) but does not perform as well on the CV data. This is apparent as the ROC curves are significantly different between the training and CV datasets. The AUC is 96.5% on training data and 86.2% on CV data. This finding suggests that the model performs comparatively poorly (with a higher rate of error or misclassification) on the CV data when compared to training data. Thus, the implication is that the model is overfitted.

Exhibit 43 ROC Curves of Model Results for Training and CV Data Before Regularization

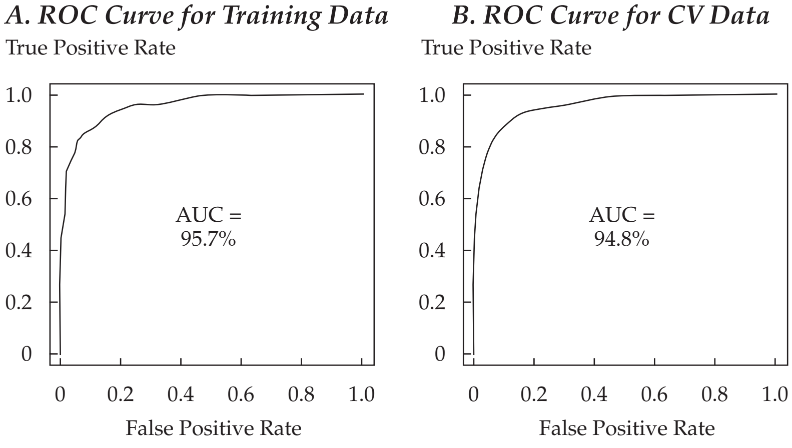


As the model is overfitted, least absolute shrinkage and selection operator (LASSO) regularization is applied to the logistic regression. LASSO regularization penalizes the coefficients of the logistic regression to prevent overfitting of the model. The penalized regression will select the tokens (features) that have statistically significant (i.e., non-zero) coefficients and that contribute to the model fit; LASSO does this while disregarding the other tokens. Exhibit 44 shows the ROC curves for the new model

3 This mathematical function is an exponential function of the form: $P(y = 1) = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$ where the βs are the logistic regression coefficients.

that uses regularized logistic regression. The ROC curves look similar for model performance on both datasets, with an AUC of 95.7% on the training dataset (Panel A) and 94.8% on the CV dataset (Panel B). These findings suggest that the model performs similarly on both training and CV data and thus indicate a good fitting model (one that is not overfitted).

Exhibit 44 ROC Curves of Model Results for Training and CV Data After Regularization



Regularization along with careful feature selection help to prevent overfitting in logistic regression models. Another model was trained using all token features, including stop words, sparse terms, and single characters, with no regularization. That model showed an AUC of 99.1% when applied on the training dataset and an AUC of 89.4% when applied on the CV dataset, suggesting that the model is overfitting. As the AUC values in all of the models discussed are not far from 100%, these models are clearly not underfitting. In sum, the final ML model for this project uses logistic regression with LASSO regularization.

To further evaluate the model, error analysis is conducted by calculating a confusion matrix using the ML model results from the cross-validation dataset. The threshold p value of 0.5 is used as a cutoff point. When target value $p > 0.5$, the prediction is assumed to be $y = 1$ (meaning, positive sentiment). Otherwise, the prediction is assumed to be $y = 0$ (negative sentiment). A confusion matrix, with performance metrics and overall scores for the model results using the CV data, is shown in Exhibit 45.

Exhibit 45 Confusion Matrix of Model Results for CV Data with Threshold p Value = 0.50

Confusion Matrix for CV Data with Threshold = 0.5

		Actual Training Labels	
		Class "1"	Class "0"
Predicted Results	Class "1"	284 (TP)	38 (FP)
	Class "0"	7 (FN)	106 (TN)

Performance Metrics

TP = 284, FP = 38, FN = 7, TN = 106

$P = 284 / (284+38) = 0.88$

$R = 284 / (284+7) = 0.98$

$\text{F1 Score} = (2 \times 0.88 \times 0.98) / (0.88 + 0.98) = 0.93$

$\text{Accuracy} = (284 + 106) / (284 + 38 + 106 + 7) = 0.90$

The model accuracy is 90% with a theoretically suggested (default) threshold p value of 0.5. The CV data are used to tune the threshold value for best model performance. Various p values from 0.01 to 0.99 are systematically evaluated individually, and confusion matrixes and performance metrics are calculated using each of these p values. Based on these metrics, the p value resulting in the highest model accuracy is selected as the ideal threshold p value. However, there are often trade-offs: Minimizing false positives (FPs) comes at a cost of increasing false negatives (FNs), and vice versa. Prioritizing various performance statistics (e.g., precision versus recall) depends on the context and relative consequences of FP and FN on the project applications. In this project, the values of negative sentiment and positive sentiment sentences are assumed to be equal, thus the impacts of FP and FN are also equal. It is common practice to simulate many model results using different threshold p values and to search for maximized accuracy and F1 statistics that minimize these trade-offs. As noted earlier, accuracy and F1 scores are overall performance measures that give equal weight to FP and FN.

Exhibit 46 shows the overall performance measures (i.e., F1 score and accuracy) for various threshold p values. The threshold p value that results in the highest accuracy and F1 score can now be identified. From the charts in Exhibit 45, the ideal threshold p value appears to be around 0.60. To investigate further, a table of performance measures (i.e., precision, recall, F1 score, and accuracy) is generated for a series of threshold p values ranging from 0.45 to 0.75. The table in Exhibit 47 demonstrates that threshold p values between 0.60 and 0.63 result in the highest accuracy and F1 score for the CV dataset. As a result of this analysis, a final threshold p value of 0.60 is selected.

Exhibit 46 Threshold Values Versus Overall Performance Measures

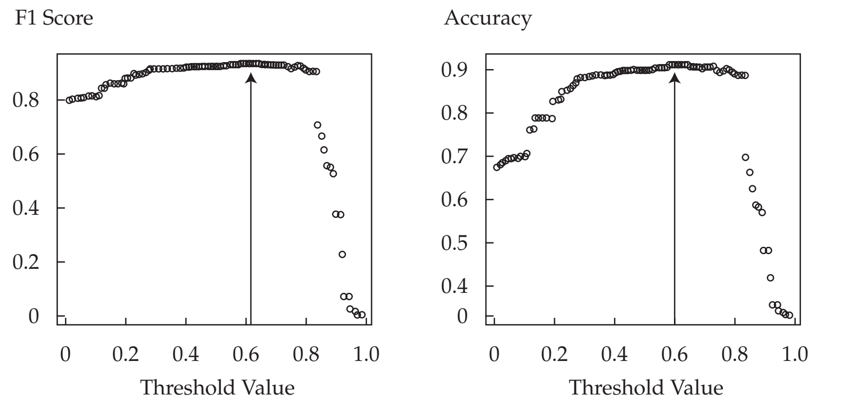


Exhibit 47 Performance Measures of the Model for a Series of Threshold Values

Threshold	Precision	Recall	F1	Accuracy
0.45	0.8750000	0.986254296	0.927302100	0.8965517
0.46	0.8827160	0.982817869	0.930081301	0.9011494
0.47	0.8827160	0.982817869	0.930081301	0.9011494
0.48	0.8819876	0.975945017	0.926590538	0.8965517
0.49	0.8819876	0.975945017	0.926590538	0.8965517
0.50	0.8819876	0.975945017	0.926590538	0.8965517
0.51	0.8819876	0.975945017	0.926590538	0.8965517
0.52	0.8819876	0.975945017	0.926590538	0.8965517
0.53	0.8902821	0.975945017	0.931147541	0.9034483
0.54	0.8930818	0.975945017	0.932676519	0.9057471
0.55	0.8930818	0.975945017	0.932676519	0.9057471
0.56	0.8958991	0.975945017	0.934210526	0.9080460
0.57	0.8958991	0.975945017	0.934210526	0.9080460
0.58	0.8958991	9.975945017	0.934210526	0.9080460
0.59	0.9015873	0.975945017	0.937293729	0.9126437
0.60	0.9044586	0.975945017	0.938842975	0.9149425
0.61	0.9044586	0.975945017	0.938842975	0.9149425
0.62	0.9044586	0.975945017	0.938842975	0.9149425
0.63	0.9041534	0.972508591	0.937086093	0.9126437
0.64	0.9041534	0.972508591	0.937086093	0.9126537
0.65	0.9041534	0.972508591	0.937086093	0.9126437
0.66	0.9035370	0.965635739	0.933554817	0.9080460
0.67	0.9035370	0.965635739	0.933554817	0.9080460
0.68	0.9064516	0.965635739	0.935108153	0.9103448
0.69	0.9064516	0.965635739	0.935108153	0.9103448
0.70	0.9061489	0.962199313	0.933333333	0.9080460
0.71	0.9061489	0.962199313	0.933333333	0.9080460

(continued)

Exhibit 47 (Continued)

Threshold	Precision	Recall	F1	Accuracy
0.72	0.9090909	0.962199313	0.934891486	0.9103448
0.73	0.9090909	0.962199313	0.934891486	0.9103448
0.74	0.9078947	0.948453608	0.927731092	0.9011494
0.75	0.9072848	0.941580756	0.924114671	0.8965517

* The shaded row shows the selected threshold p value (0.60) and the performance metrics for the selected model.

Finally, the confusion matrix using the ideal threshold p value of 0.60 is constructed to observe the performance of the final model. When target value p > 0.60, the prediction is assumed to be y = 1 (indicating positive sentiment); otherwise, the prediction is assumed to be y = 0 (negative sentiment). The confusion matrix for the CV data is shown in Exhibit 48. It is clear that the model performance metrics have improved in the final model compared to the earliest case when the threshold p value was 0.50. Now, accuracy and F1 score have both increased by one percentage point to 91% and 94%, respectively, while precision has increased by two percentage points to 90%.

Exhibit 48 Confusion Matrix of Model Results for CV Data with Threshold p Value = 0.60

Confusion Matrix for CV Data with Threshold = 0.6

		Actual Training Labels	
		Class "1"	Class "0"
Predicted Results	Class "1"	284 (TP)	30 (FP)
	Class "0"	7 (FN)	114 (TN)

Performance Metrics

TP = 284, FP = 30, FN = 7, TN = 114

P = 284 / (284+30) = 0.90

R = 284 / (284+7) = 0.98

F1 Score = (2 × 0.90 × 0.98) / (0.90 + 0.98) = 0.94

Accuracy = (284 + 114) / (284 + 30 + 114 + 7) = 0.91

7.4 Results and Interpretation

The final ML model with the appropriate threshold p value has been validated and is now ready for use. The model can be used to predict the sentiment of new sentences from the test data corpus as well as new sentences from similar financial text data sources, such as news wires, earnings call transcripts, and quarterly financial reports. The final model is a collection of penalized regression coefficients for unigram and bigram tokens from the BOW of the training corpus. To use the model to predict the sentiment of new sentences, tokenization and identical cleansing and preprocessing operations must be performed on the new sentences. All the processes performed on the training data must be performed on the new data to which the model will be applied (as was done for the test dataset). The model will use the trained penalized regression coefficients on the term frequency (TF) values of the tokens in the document term matrix (DTM) of the new sentences and will determine the target value (p). The columns of the DTM of the new sentences are the same as those of the training DTM, but the TF values are calculated based on the test corpus. Using the threshold p value of 0.60, the sentiment class for each sentence in the test corpus will be predicted.

The model is now applied on the test data that contains 436 sentences. Note that the test data were not used to train or validate/tune the model and are new to the model. The test data were preprocessed identically to the training and CV data while a part of the master corpus. The model is then applied to the test DTM, and the results are obtained. Exhibit 49 displays 30 sample results from the test corpus. The results table contains cleansed and preprocessed sentences, actual sentiment, target p values from the model, and predicted sentiment. Note that this sample contains three cases of misclassification: the 10th sentence (text), where p = 0.46; the 26th text, where p = 0.77; and the 30th text, where p = 0.71. Therefore, accuracy of this 30-text sample is 27/30 = 90%.

Exhibit 49 Thirty Sample Results of Test Data

Sentence	Sentiment	p	Predicted Sentiment
exclude non recur item pre tax surg percentsign	1	0.81	1
adp news feb finnish retail kesko oyj hel kesbv said today total sale exclud valu ad tax vat stood at januari down percentsign on yea	0	0.12	0
india trade with russia current stand at four billion dollar grow per cent fiscal	1	0.83	1
refin margin was bbl combar bbl prior	1	0.81	1
scania morgan Stanley lift share target on swedish heavi duti truck bus maker scania ab crown euro crown euro	1	0.83	1
deal is like bring save	1	0.83	1
will also strengthen ruukki offshore busi	1	0.83	1
last week finnish metl technolog group announc plan sell more than percent technolog unit further compani strategy goal becom world largest stainless steel maker	1	0.83	1
nest oil board propos dividend full compar with ago	1	0.81	1
pre tax loss total compar loss first quarter	1	0.46	0
pretax total compar loss fourth quarter	1	0.74	1
re use back into pet bottle has also steadili increas rate use strap tape has pick up again after dip pector said previous	1	0.95	1
satama sale would be higher than befor	1	0.83	1
octob finnish wood product technolog supplier raut oyj hel rutav said today swung first nine month versus loss same period earlier	1	0.79	1
ebit total compar loss correspond period	1	0.74	1
finnish consum packag manufactur huhtamaki oyj said swung euro first nine month loss euro same period	1	0.77	1
finnish dental care group oral hammaslaakarit oyj post total euro first nine month versus loss euro same period	1	0.79	1
finnish silicon water manufactur okmet oyj said swung euro first nine month loss euro earlier	1	0.77	1
adp news feb finnish print circuit board pcb maker aspocomp group oyj hel acg said today swung versus loss	1	0.79	1
mn pretax third quarter	1	0.83	1

(continued)

Exhibit 49 (Continued)

Sentence	Sentiment	p	Predicted Sentiment
oper total compar correspond period	1	0.81	1
raut post euro third quarter compar loss euro correspond period	1	0.74	1
russian export duti will active harvest finland sale russia will increas also	1	0.91	1
compani expect sale signific increas	1	0.91	1
compani amount ee which was percentsign more than	1	0.81	1
third quarter fiscal efor swung loss versus correspond period fiscal	0	0.77	1
acando ab acanb ss fell percent kronor lowest close sinc dec	0	0.20	0
compani oper loss total compar	0	0.27	0
last paseng flew airlin down percent	0	0.12	0
loss after financi item total compar correspond period	0	0.71	1

Exhibit 50 shows the confusion matrix for the test data. Accuracy and F1 score are 90% and 93%, respectively, while precision and recall are 89% and 98%, respectively. Therefore, it is apparent that the model performs similarly on the training, CV, and test datasets. These findings suggest that the model is robust and is not overfitting. They also suggest that the model should generalize well out-of-sample and can thus be used to predict the sentiment classes for new sentences from similar financial text data sources. Of course, these new text data must first be subjected to identical tokenization, cleansing, and preprocessing as done for the training dataset.

Exhibit 50 Confusion Matrix of Model Results for Test Data with Threshold p Value = 0.60

Confusion Matrix for Test Data

		Actual Training Labels	
		Class "1"	Class "0"
Predicted Results	Class "1"	284 (TP)	35 (FP)
	Class "0"	7 (FN)	110 (TN)

Performance Metrics

TP = 284, FP = 35, FN = 7, TN = 110

P = 284 / (284+35) = 0.89

R = 284 / (284+7) = 0.98

F1 Score = (2 × 0.89 × 0.98) / (0.89 + 0.98) = 0.93

Accuracy = (284 + 110) / (284 + 35 + 110 + 7) = 0.90

To recap, this project involves converting unstructured data (i.e., text data from financial data sources) into structured data (i.e., tokens, sentences, and term frequency values) in a document term matrix that is used as input for training, validating, and testing machine learning-based models (here, logistic regression) for predicting classification (here, sentiment classes). Similar models can be built and used in different contexts to understand the sentiment embedded in larger texts. The derived sentiment classification can be useful as a visualization tool to provide insight about the text without reading large documents. These sentiment classifications can also be used as structured input data for larger ML models that have a specific purpose, such as to predict future stock price movements.

EXAMPLE 7

Comparing Performance Metrics for Confusion Matrixes with Different Threshold p Values

In the previous analysis using the cross-validation dataset, performance measures for the sentiment classification ML model were calculated for a wide range (from 0.45 to 0.75) of threshold p values. The threshold value of 0.60 was determined to be the p value that maximizes model accuracy and F1 score; the confusion matrix for this model is shown in Exhibit 48. Use the following confusion matrixes with threshold p values of 0.75 and 0.45, A and B, respectively, to answer the following questions.

Confusion Matrix A

Confusion Matrix for CV, Threshold = 0.75

N = 436		Actual Training Labels	
		Class "1"	Class "0"
Predicted Results	Class "1"	281	28
	Class "0"	17	110

Performance Metrics

TP = 281, FP = 28, FN = 17, TN = 110
Precision = TP/(TP + FP) = 0.91
Recall = TP/(TP + FN) = 0.94
F1 Score = HMean: Prec. & Recall = 0.93
Accuracy = (TP + TN)/N = 0.90

Confusion Matrix B

Confusion Matrix for CV, Threshold = 0.45

N = 436		Actual Training Labels	
		Class "1"	Class "0"
Predicted Results	Class "1"	281	41
	Class "0"	4	110

Performance Metrics

TP = 281, FP = 41, FN = 4, TN = 110
Precision = TP/(TP + FP) = 0.87
Recall = TP/(TP + FN) = 0.99
F1 Score = HMean: Prec. & Recall = 0.93
Accuracy = (TP + TN)/N = 0.90

1 Compare the performance metrics of confusion matrix A (using a threshold p value of 0.75) with the confusion matrix in Exhibit 48 (using a threshold p value of 0.60).

2 Compare the performance metrics of confusion matrix B (using a threshold p value of 0.45) with the confusion matrix in Exhibit 48 (using a threshold p value of 0.60).

3 Contrast the performance metrics of confusion matrixes A and B, and explain the trade-offs implied between them.

Solution to 1:

Since confusion matrix A has fewer true positives (TPs) and fewer true negatives (TNs) than the confusion matrix in Exhibit 48 (281 vs. 284 and 110 vs. 114, respectively), confusion matrix A has lower accuracy and a lower F1 score compared to the one in Exhibit 48 (0.90 vs. 0.91 and 0.93 vs. 0.94, respectively). Also, although confusion matrix A has slightly better precision, 0.91 vs. 0.90, due to a few less false positives (FPs), it has significantly lower recall, 0.94 vs. 0.98, due to having many more false negatives (FNs), 17 vs. 7, than the confusion matrix in Exhibit 48. On balance, the ML model using the threshold p value of 0.60 is the superior model for this sentiment classification problem.

Solution to 2:

Confusion matrix B has the same number of TPs (281) and TNs (110) as confusion matrix A. Therefore, confusion matrix B also has lower accuracy (0.90) and a lower F1 score (0.93) compared to the one in Exhibit 48. Although confusion matrix B has slightly better recall, 0.99 vs. 0.98, due to fewer FNs, it has somewhat lower precision, 0.87 vs. 0.90, due to having many more FPs, 41 vs.

30, than the confusion matrix in Exhibit 48. Again, it is apparent that the ML model using the threshold p value of 0.60 is the better model in this sentiment classification context.

Solution to 3:

The main differences in performance metrics between confusion matrixes A and B are in precision and recall. Confusion matrix A has higher precision, at 0.91 vs. 0.87, but confusion matrix B has higher recall, at 0.99 vs. 0.94. These differences highlight the trade-off between FP (Type I error) and FN (Type II error). Precision is useful when the cost of FP is high, such as when an expensive product that is fine mistakenly fails quality inspection and is scrapped; in this case, FP should be minimized. Recall is useful when the cost of FN is high, such as when an expensive product is defective but mistakenly passes quality inspection and is sent to the customer; in this case, FN should be minimized. In the context of sentiment classification, FP might result in buying a stock for which sentiment is incorrectly classified as positive when it is actually negative. Conversely, FN might result in avoiding (or even shorting) a stock for which the sentiment is incorrectly classified as negative when it is actually positive. The model behind the confusion matrix in Exhibit 48 strikes a balance in the trade-off between precision and recall.

SUMMARY

In this reading, we have discussed the major steps in big data projects involving the development of machine learning (ML) models—namely, those combining textual big data with structured inputs.

- Big data—defined as data with volume, velocity, variety, and potentially lower veracity—has tremendous potential for various fintech applications, including several related to investment management.
- The main steps for traditional ML model building are conceptualization of the problem, data collection, data preparation and wrangling, data exploration, and model training.
- For textual ML model building, the first four steps differ somewhat from those used in the traditional model: Text problem formulation, text curation, text preparation and wrangling, and text exploration are typically necessary.
- For structured data, data preparation and wrangling entail data cleansing and data preprocessing. Data cleansing typically involves resolving incompleteness errors, invalidity errors, inaccuracy errors, inconsistency errors, non-uniformity errors, and duplication errors.
- Preprocessing for structured data typically involves performing the following transformations: extraction, aggregation, filtration, selection, and conversion.
- Preparation and wrangling text (unstructured) data involves a set of text-specific cleansing and preprocessing tasks. Text cleansing typically involves removing the following: html tags, punctuations, most numbers, and white spaces.

- Text preprocessing requires performing normalization that involves the following: lowercasing, removing stop words, stemming, lemmatization, creating bag-of-words (BOW) and n-grams, and organizing the BOW and n-grams into a document term matrix (DTM).
- Data exploration encompasses exploratory data analysis, feature selection, and feature engineering. Whereas histograms, box plots, and scatterplots are common techniques for exploring structured data, word clouds are an effective way to gain a high-level picture of the composition of textual content. These visualization tools help share knowledge among the team (business subject matter experts, quants, technologists, etc.) to help derive optimal solutions.
- Feature selection methods used for text data include term frequency, document frequency, chi-square test, and a mutual information measure. Feature engineering for text data includes converting numbers into tokens, creating n-grams, and using name entity recognition and parts of speech to engineer new feature variables.
- The model training steps (method selection, performance evaluation, and model tuning) often do not differ much for structured versus unstructured data projects.
- Model selection is governed by the following factors: whether the data project involves labeled data (supervised learning) or unlabeled data (unsupervised learning); the type of data (numerical, continuous, or categorical; text data; image data; speech data; etc.); and the size of the dataset.
- Model performance evaluation involves error analysis using confusion matrixes, determining receiver operating characteristics, and calculating root mean square error.
- To carry out an error analysis for each model, a confusion matrix is created; true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) are determined. Then, the following performance metrics are calculated: accuracy, F1 score, precision, and recall. The higher the accuracy and F1 score, the better the model performance.
- To carry out receiver operating characteristic (ROC) analysis, ROC curves and area under the curve (AUC) of various models are calculated and compared. The more convex the ROC curve and the higher the AUC, the better the model performance.
- Model tuning involves managing the trade-off between model bias error, associated with underfitting, and model variance error, associated with overfitting. A fitting curve of in-sample (training sample) error and out-of-sample (cross-validation sample) error on the y-axis versus model complexity on the x-axis is useful for managing the bias vs. variance error trade-off.
- In a real-world big data project involving text data analysis for classifying and predicting sentiment of financial text for particular stocks, the text data are transformed into structured data for populating the DTM, which is then used as the input for the ML algorithm.
- To derive term frequency (TF) at the sentence level and TF–IDF, both of which can be inputs to the DTM, the following frequency measures should be used to create a term frequency measures table: TotalWordsInSentence; TotalWordCount; TermFrequency (Collection Level); WordCountInSentence; SentenceCountWithWord; Document Frequency; and Inverse Document Frequency.

PRACTICE PROBLEMS

The following information relates to Questions 1–15

Aaliyah Schultz is a fixed-income portfolio manager at Aries Investments. Schultz supervises Ameris Steele, a junior analyst.

A few years ago, Schultz developed a proprietary machine learning (ML) model that aims to predict downgrades of publicly-traded firms by bond rating agencies. The model currently relies only on structured financial data collected from different sources. Schultz thinks the model's predictive power may be improved by incorporating sentiment data derived from textual analysis of news articles and Twitter content relating to the subject companies.

Schultz and Steele meet to discuss plans for incorporating the sentiment data into the model. They discuss the differences in the steps between building ML models that use traditional structured data and building ML models that use textual big data. Steele tells Schultz:

Statement 1 The second step in building text-based ML models is text preparation and wrangling, whereas the second step in building ML models using structured data is data collection.

Statement 2 The fourth step in building both types of models encompasses data/text exploration.

Steele expresses concern about using Twitter content in the model, noting that research suggests that as much as 10%–15% of social media content is from fake accounts. Schultz tells Steele that she understands her concern but thinks the potential for model improvement outweighs the concern.

Steele begins building a model that combines the structured financial data and the sentiment data. She starts with cleansing and wrangling the raw structured financial data. Exhibit 1 presents a small sample of the raw dataset before cleansing: Each row represents data for a particular firm.

Exhibit 1 Sample of Raw Structured Data Before Cleansing

ID	Ticker	IPO Date	Industry (NAICS)	EBIT	Interest Expense	Total Debt
1	ABC	4/6/17	44	9.4	0.6	10.1
2	BCD	November 15, 2004	52	5.5	0.4	6.2
3	HIJ	26-Jun-74	54	8.9	1.2	15.8
4	KLM	14-Mar-15	72	5.7	1.5	0.0

After cleansing the data, Steele then preprocesses the dataset. She creates two new variables: an "Age" variable based on the firm's IPO date and an "Interest Coverage Ratio" variable equal to EBIT divided by interest expense. She also deletes the "IPO Date" variable from the dataset. After applying these transformations, Steele scales the financial data using normalization. She notes that over the full sample dataset, the "Interest Expense" variable ranges from a minimum of 0.2 and a maximum of 12.2, with a mean of 1.1 and a standard deviation of 0.4.

Steele and Schultz then discuss how to preprocess the raw text data. Steele tells Schultz that the process can be completed in the following three steps:

Step 1 Cleanse the raw text data.

Step 2 Split the cleansed data into a collection of words for them to be normalized.

Step 3 Normalize the collection of words from Step 2 and create a distinct set of tokens from the normalized words.

With respect to Step 1, Steele tells Schultz:

> "I believe I should remove all html tags, punctuations, numbers, and extra white spaces from the data before normalizing them."

After properly cleansing the raw text data, Steele completes Steps 2 and 3. She then performs exploratory data analysis. To assist in feature selection, she wants to create a visualization that shows the most informative words in the dataset based on their term frequency (TF) values. After creating and analyzing the visualization, Steele is concerned that some tokens are likely to be noise features for ML model training; therefore, she wants to remove them.

Steele and Schultz discuss the importance of feature selection and feature engineering in ML model training. Steele tells Schultz:

> "Appropriate feature selection is a key factor in minimizing model overfitting, whereas feature engineering tends to prevent model underfitting."

Once satisfied with the final set of features, Steele selects and runs a model on the training set that classifies the text as having positive sentiment (Class "1" or negative sentiment (Class "0"). She then evaluates its performance using error analysis. The resulting confusion matrix is presented in Exhibit 2.

Exhibit 2 Confusion Matrix

		Actual Training Results	
		Class "1"	Class "0"
Predicted Results	Class "1"	TP = 182	FP = 52
	Class "0"	FN = 31	TN = 96

1 Which of Steele's statements relating to the steps in building structured data-based and text-based ML models is correct?

A Only Statement 1 is correct.

B Only Statement 2 is correct.

C Statement 1 and Statement 2 are correct.

2 Steele's concern about using Twitter data in the model *best* relates to:

A volume.

B velocity.

C veracity.

3 What type of error appears to be present in the IPO Date column of Exhibit 1?

A invalidity error.

B inconsistency error.

C non-uniformity error.

4 What type of error is most likely present in the last row of data (ID #4) in Exhibit 1?

A Inconsistency error

B Incompleteness error

C Non-uniformity error

5 During the preprocessing of the data in Exhibit 1, what type of data transformation did Steele perform during the data preprocessing step?

A Extraction

B Conversion

C Aggregation

6 Based on Exhibit 1, for the firm with ID #3, Steele should compute the scaled value for the "Interest Expense" variable as:

A 0.008.

B 0.083.

C 0.250.

7 Is Steele's statement regarding Step 1 of the preprocessing of raw text data correct?

A Yes.

B No, because her suggested treatment of punctuation is incorrect.

C No, because her suggested treatment of extra white spaces is incorrect.

8 Steele's Step 2 can be *best* described as:

A tokenization.

B lemmatization.

C standardization.

9 The output created in Steele's Step 3 can be *best* described as a:

A bag-of-words.

B set of n-grams.

C document term matrix.

10 Given her objective, the visualization that Steele should create in the exploratory data analysis step is a:

A scatter plot.

B word cloud.

C document term matrix.

11 To address her concern in her exploratory data analysis, Steele should focus on those tokens that have:

A low chi-square statistics.

B low mutual information (ML) values.

C very low and very high term frequency (TF) values.

12 Is Steele's statement regarding the relationship between feature selection/feature engineering and model fit correct?

A Yes.

B No, because she is incorrect with respect to feature selection.

C No, because she is incorrect with respect to feature engineering.

13 Based on Exhibit 2, the model's precision metric is *closest* to:

A 78%.

B 81%.

C 85%.

14 Based on Exhibit 2, the model's F1 score is *closest* to:

A 77%.

B 81%.

C 85%.

15 Based on Exhibit 2, the model's accuracy metric is *closest* to:

A 77%.

B 81%.

C 85%.

SOLUTIONS

1 B is correct. The five steps in building structured data-based ML models are: 1) conceptualization of the modeling task, 2) data collection, 3) data preparation and wrangling, 4) data exploration, and 5) model training. The five steps in building text-based ML models are: 1) text problem formulation, 2) data (text) curation, 3) text preparation and wrangling, 4) text exploration, and 5) model training. Statement 1 is incorrect: Text preparation and wrangling is the third step in building text ML models and occurs after the second data (text) curation step. Statement 2 is correct: The fourth step in building both types of models encompasses data/text exploration.

2 C is correct. Veracity relates to the credibility and reliability of different data sources. Steele is concerned about the credibility and reliability of Twitter content, noting that research suggests that as much as 10%–15% of social media content is from fake accounts.

3 C is correct. A non-uniformity error occurs when the data are not presented in an identical format. The data in the "IPO Date" column represent the IPO date of each firm. While all rows are populated with valid dates in the IPO Date column, the dates are presented in different formats (e.g., mm/dd/yyyy, dd/mm/yyyy).

4 A is correct. There appears to be an inconsistency error in the last row (ID #4). An inconsistency error occurs when a data point conflicts with corresponding data points or reality. In the last row, the interest expense data item has a value of 1.5, and the total debt item has a value of 0.0. This appears to be an error: Firms that have interest expense are likely to have debt in their capital structure, so either the interest expense is incorrect or the total debt value is incorrect. Steele should investigate this issue by using alternative data sources to confirm the correct values for these variables.

5 A is correct. During the data preprocessing step, Steele created a new "Age" variable based on the firm's IPO date and then deleted the "IPO Date" variable from the dataset. She also created a new "Interest Coverage Ratio" variable equal to EBIT divided by interest expense. Extraction refers to a data transformation where a new variable is extracted from a current variable for ease of analyzing and using for training an ML model, such as creating an age variable from a date variable or a ratio variable. Steele also performed a selection transformation by deleting the IPO Date variable, which refers to deleting the data columns that are not needed for the project.

6 B is correct. Steele uses normalization to scale the financial data. Normalization is the process of rescaling numeric variables in the range of [0, 1]. To normalize variable X, the minimum value ($X_{\min}$) is subtracted from each observation (X_i), and then this value is divided by the difference between the maximum and minimum values of X ($X_{\max} - X_{\min}$):

$$X_{i\ (\text{normalized})} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

The firm with ID #3 has an interest expense of 1.2. So, its normalized value is calculated as:

$$X_{i\ (\text{normalized})} = \frac{1.2 - 0.2}{12.2 - 0.2} = 0.083$$

7 B is correct. Although most punctuations are not necessary for text analysis and should be removed, some punctuations (e.g., percentage signs, currency symbols, and question marks) may be useful for ML model training. Such punctuations should be substituted with annotations (e.g., /percentSign/, /dollarSign/, and /questionMark/) to preserve their grammatical meaning in the text. Such annotations preserve the semantic meaning of important characters in the text for further text processing and analysis stages.

8 A is correct. Tokenization is the process of splitting a given text into separate tokens. This step takes place after cleansing the raw text data (removing html tags, numbers, extra white spaces, etc.). The tokens are then normalized to create the bag-of-words (BOW).

9 A is correct. After the cleansed text is normalized, a bag-of-words is created. A bag-of-words (BOW) is a collection of a distinct set of tokens from all the texts in a sample dataset.

10 B is correct. Steele wants to create a visualization for Schultz that shows the most informative words in the dataset based on their term frequency (TF, the ratio of the number of times a given token occurs in the dataset to the total number of tokens in the dataset) values. A word cloud is a common visualization when working with text data as it can be made to visualize the most informative words and their TF values. The most commonly occurring words in the dataset can be shown by varying font size, and color is used to add more dimensions, such as frequency and length of words.

11 C is correct. Frequency measures can be used for vocabulary pruning to remove noise features by filtering the tokens with very high and low TF values across all the texts. Noise features are both the most frequent and most sparse (or rare) tokens in the dataset. On one end, noise features can be stop words that are typically present frequently in all the texts across the dataset. On the other end, noise features can be sparse terms that are present in only a few text files. Text classification involves dividing text documents into assigned classes. The frequent tokens strain the ML model to choose a decision boundary among the texts as the terms are present across all the texts (an example of underfitting). The rare tokens mislead the ML model into classifying texts containing the rare terms into a specific class (an example of overfitting). Thus, identifying and removing noise features are critical steps for text classification applications.

12 A is correct. A dataset with a small number of features may not carry all the characteristics that explain relationships between the target variable and the features. Conversely, a large number of features can complicate the model and potentially distort patterns in the data due to low degrees of freedom, causing overfitting. Therefore, appropriate feature selection is a key factor in minimizing such model overfitting. Feature engineering tends to prevent underfitting in the training of the model. New features, when engineered properly, can elevate the underlying data points that better explain the interactions of features. Thus, feature engineering can be critical to overcome underfitting.

13 A is correct. Precision, the ratio of correctly predicted positive classes (true positives) to all predicted positive classes, is calculated as:

$$\text{Precision (P)} = \text{TP}/(\text{TP} + \text{FP}) = 182/(182 + 52) = 0.7778\ (78\%).$$

14 B is correct. The model's F1 score, which is the harmonic mean of precision and recall, is calculated as:

$$\text{F1 score} = (2 \times \text{P} \times \text{R})/(\text{P} + \text{R}).$$

$$\text{F1 score} = (2 \times 0.7778 \times 0.8545)/(0.7778 + 0.8545) = 0.8143\ (81\%).$$

15 A is correct. The model's accuracy, which is the percentage of correctly predicted classes out of total predictions, is calculated as:

Accuracy = (TP + TN)/(TP + FP + TN + FN).

Accuracy = (182 + 96)/(182 + 52 + 96 + 31) = 0.7701 (77%).

READING

9

Excerpt from "Probabilistic Approaches: Scenario Analysis, Decision Trees, and Simulations"

by Aswath Damodaran

Aswath Damodaran is at the Stern School of Business at New York University (USA).

LEARNING OUTCOMES

Mastery	The candidate should be able to:
☐	**a.** describe steps in running a simulation;
☐	**b.** explain three ways to define the probability distributions for a simulation's variables;
☐	**c.** describe how to treat correlation across variables in a simulation;
☐	**d.** describe advantages of using simulations in decision making;
☐	**e.** describe some common constraints introduced into simulations;
☐	**f.** describe issues in using simulations in risk assessment;
☐	**g.** compare scenario analysis, decision trees, and simulations.

INTRODUCTION

1

Scenario analysis, which applies probabilities to a small number of possible outcomes, and decision trees, which use tree diagrams of possible outcomes, are techniques used to assess risk. Simulations are also used to assess risk.

Damodaran, Aswath, *Strategic Risk Taking: A Framework for Risk Management*, 1st Edition, © 2008, pp. 164–183. Reprinted by permission of Aswath Damodaran.

2 SIMULATIONS

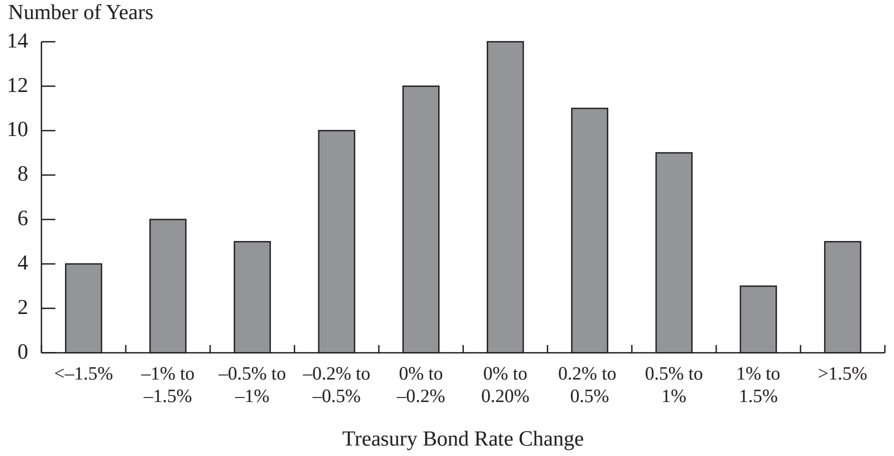
If scenario analysis and decision trees are techniques that help us to assess the effects of discrete risk, simulations provide a way of examining the consequences of continuous risk. To the extent that most risks that we face in the real world can generate hundreds of possible outcomes, a simulation will give us a fuller picture of the risk in an asset or investment.

2.1 Steps in Simulation

Unlike scenario analysis, where we look at the values under discrete scenarios, simulations allow for more flexibility in how we deal with uncertainty. In its classic form, distributions of values are estimated for each parameter in the analysis (growth, market share, operating margin, beta, etc.). In each simulation, we draw one outcome from each distribution to generate a unique set of cash flows and value. Across a large number of simulations, we can derive a distribution for the value of investment or an asset that will reflect the underlying uncertainty we face in estimating the inputs to the valuation. The steps associated with running a simulation are as follows:

1. **Determine "probabilistic" variables**: In any analysis, there are potentially dozens of inputs, some of which are predictable and some of which are not. Unlike scenario analysis and decision trees, where the number of variables that are changed and the potential outcomes have to be few in number, there is no constraint on how many variables can be allowed to vary in a simulation. At least in theory, we can define probability distributions for each and every input in a valuation. The reality, though, is that this will be time consuming and may not provide much of a payoff, especially for inputs that have only marginal impact on value. Consequently, it makes sense to focus attention on a few variables that have a significant impact on value.
2. **Define probability distributions for these variables**: This is a key and the most difficult step in the analysis. Generically, there are three ways in which we can go about defining probability distributions:
   - A *Historical data*: For variables that have a long history and reliable data over that history, it is possible to use the historical data to develop distributions. Assume, for instance, that you are trying to develop a distribution of expected changes in the long-term Treasury bond rate (to use as an input in investment analysis). You could use the histogram in Figure 6.6, based upon the annual changes in Treasury bond rates every year from 1928 to 2005, as the distribution for future changes.

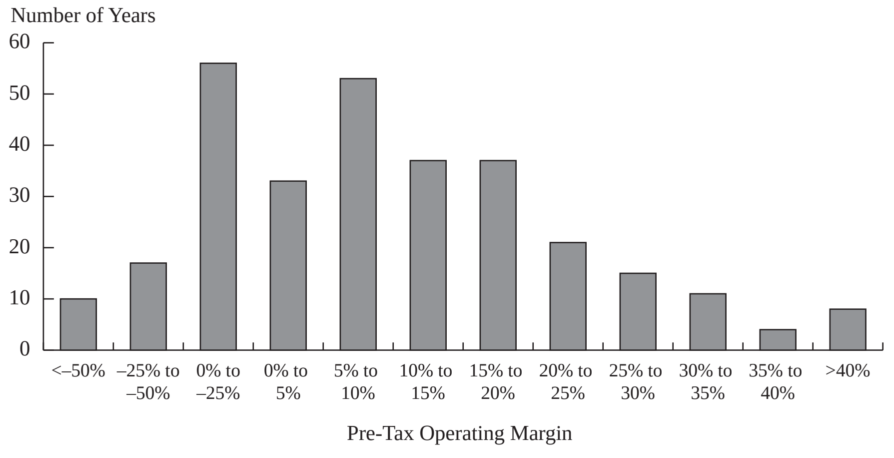
Figure 6.6 Change in T-Bond Rate: 1928–2005



Implicit in this approach is the assumption that there have been no structural shifts in the market that will render the historical data unreliable.

B *Cross sectional data*: In some cases, you may be able to substitute data on differences in a specific variable across existing investments that are similar to the investment being analyzed. Consider two examples. Assume that you are valuing a software firm and are concerned about the uncertainty in operating margins. Figure 6.7 provides a distribution of pre-tax operating margins across software companies in 2006:

Figure 6.7 Pre-tax Operating Margin across Software Companies (US): January 2006



If we use this distribution, we are in effect assuming that the cross sectional variation in the margin is a good indicator of the uncertainty we face in estimating it for the software firm in question. In a second example, assume that you work for Target, the retailer, and that you are trying to estimate the sales per square foot for a new store investment. Target could use the distribution on this variable across existing stores as the basis for its simulation of sales at the new store.

C *Statistical distribution and parameters*: For most variables that you are trying to forecast, the historical and cross sectional data will be insufficient or unreliable. In these cases, you have to pick a statistical distribution that best captures the variability in the input and estimate the parameters for that distribution. Thus, you may conclude that operating margins will be distributed uniformly, with a minimum of 4% and a maximum of 8%, and that revenue growth is normally distributed with an expected value of 8% and a standard deviation of 6%. Many simulation packages available for personal computers now provide a rich array of distributions to choose from, but picking the right distribution and the parameters for the distribution remains difficult for two reasons. The first is that few inputs that we see in practice meet the stringent requirements that statistical distributions demand; revenue growth, for instance, cannot be normally distributed because the lowest value it can take on is –100%. Consequently, we have to settle for statistical distributions that are close enough to the real distribution that the resulting errors will not wreak havoc on our conclusion. The second is that the parameters still need to be estimated, once the distribution is picked. For this, we can draw on historical or cross sectional data; and for the revenue growth input, we can look at revenue growth in prior years or revenue growth rate differences across peer group companies. The caveats about structural shifts that make historical data unreliable and peer group companies not being comparable continue to apply.

The probability distributions used are discrete for some inputs and continuous for others, and are based upon historical data for some and statistical distributions for others.

3 **Check for correlation across variables**: While it is tempting to jump to running simulations right after the distributions have been specified, it is important that we check for correlations across variables. Assume, for instance, that you are developing probability distributions for both interest rates and inflation. While both inputs may be critical in determining value, they are likely to be correlated with each other; high inflation is usually accompanied by high interest rates. When there is strong correlation, positive or negative, across inputs, you have two choices. One is to pick only one of the two inputs to vary; it makes sense to focus on the input that has the bigger impact on value. The other is to build the correlation explicitly into the simulation; this does require more sophisticated simulation packages and adds more detail to the estimation process. As with the distribution, the correlations can be estimated by looking at the past.

4 **Run the simulation**: For the first simulation, you draw one outcome from each distribution and compute the value based upon those outcomes. This process can be repeated as many times as desired, though the marginal contribution of each simulation drops off as the number of simulations increases. The number of simulations you run will be determined by the following:

A *Number of probabilistic inputs*: The larger the number of inputs that have probability distributions attached to them, the greater will be the required number of simulations.

B *Characteristics of probability distributions*: The greater the diversity of distributions in an analysis, the larger will be the number of required simulations. Thus, the number of required simulations will be smaller in a simulation where all of the inputs have normal distributions than in one where some have normal distributions, some are based upon historical data distributions, and some are discrete.

c *Range of outcomes*: The greater the potential range of outcomes on each input, the greater will be the number of simulations.

Most simulation packages allow users to run thousands of simulations, with little or no cost attached to increasing that number. Given that reality, it is better to err on the side of too many simulations rather than too few.

There have generally been two impediments to good simulations. The first is informational: estimating distributions of values for each input into a valuation is difficult to do. In other words, it is far easier to estimate an expected growth rate of 8% in revenues for the next 5 years than it is to specify the distribution of expected growth rates—the type of distribution, parameters of that distribution—for revenues. The second is computational: until the advent of personal computers, simulations tended to be too time and resource intensive for the typical analysis. Both these constraints have eased in recent years and simulations have become more feasible.

2.2 An Example of a Simulation

Running a simulation is simplest for firms that consider the same kind of projects repeatedly. These firms can use their experience from similar projects that are already in operation to estimate expected values for new projects. The Home Depot, for instance, analyzes dozens of new home improvement stores every year. It also has hundreds of stores in operation,[1] at different stages of their life cycles; some of these stores have been in operation for more than 10 years and others have been around only for a couple of years. Thus, when forecasting revenues for a new store, the Home Depot can draw on this rich database to make its estimates more precise. The firm has a reasonable idea of how long it takes a new store to become established and how store revenues change as the store ages and new stores open close by.

There are other cases where experience can prove useful for estimating revenues and expenses on a new investment. An oil company, in assessing whether to put up an oil rig, comes into the decision with a clear sense of what the costs are of putting up a rig, and how long it will take for the rig to be productive. Similarly, a pharmaceutical firm, when introducing a new drug, can bring to its analysis its experience with other drugs in the past, how quickly such drugs are accepted and prescribed by doctors, and how responsive revenues are to pricing policy. We are not suggesting that the experience these firms have had in analyzing similar projects in the past removes uncertainty about the project from the analysis. The Home Depot is still exposed to considerable risk on each new store that it analyzes today, but the experience does make the estimation process easier and the estimation error smaller than it would be for a firm that is assessing a unique project.

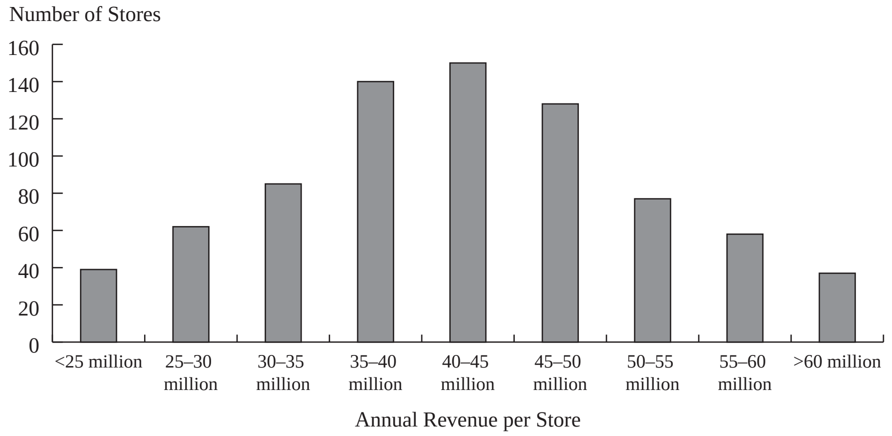
Assume that the Home Depot is analyzing a new home improvement store that will follow its traditional format.[2] There are several estimates the Home Depot needs to make when analyzing a new store. Perhaps the most important is the likely revenues at the store. Given that the Home Depot's store sizes are similar across locations, the firm can get an idea of the expected revenues by looking at revenues at their existing stores. Figure 6.8 summarizes the distribution[3] of annual revenues at existing stores in 2005:

1 At the end of 2005, the Home Depot had 743 Home Depot stores in operation, 707 of which were in the United States.

2 A typical Home Depot store has store space of about 100,000 square feet and carries a wide range of home improvement products, from hardware to flooring.

3 This distribution is a hypothetical one, since the Home Depot does not provide this information to outsiders. It does have the information internally.

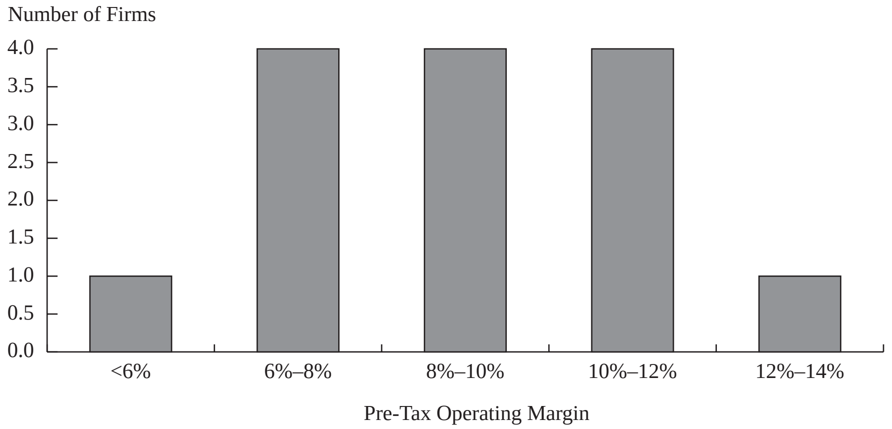
Figure 6.8 Revenues/Store: Home Depot US Stores in 2005



This distribution not only yields an expected revenue per store of about $44 million, but also provides a measure of the uncertainty associated with the estimate, in the form of a standard deviation in revenues per store.

The second key input is the operating margin that Home Depot expects to generate at this store. While the margins are fairly similar across all of its existing stores, there are significant differences in margins across different building supply retailers, reflecting their competitive strengths or weaknesses. Figure 6.9 summarizes differences in pre-tax operating margins across building supply retailers:

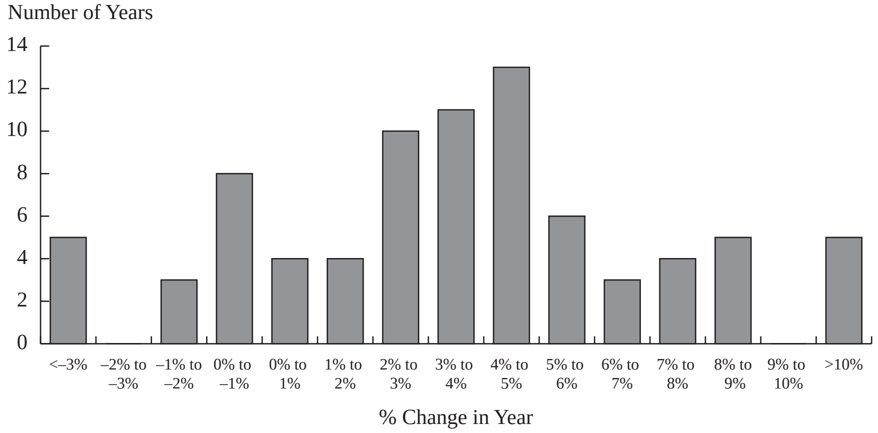
Figure 6.9 Pre-tax Operating Margin at Building Retailers (US) January 2006



Note that this distribution, unlike the revenue distribution, does not have a noticeable peak. In fact, with one outlier in either direction, it is distributed evenly between 6% and 12%.

Finally, the store's future revenues will be tied to an estimate of expected growth, which we will assume will be strongly influenced by overall economic growth in the United States. To get a measure of this growth, we looked at the distribution of real GDP growth from 1925 to 2005 in Figure 6.10:

Figure 6.10 Annual Percent Change in Real GDP for US: 1925–2005



To run a simulation of the Home Depot's store's cash flows and value, we will make the following assumptions:

- **Base revenues**: We will base our estimate of the base year's revenues on Figure 6.8. For computational ease, we will assume that revenue will be normally distributed with an expected value of \$44 million and a standard deviation of \$10 million.
- **Pre-tax operating margin**: Based upon Figure 6.9, the pre-tax operating margin is assumed to be uniformly distributed with a minimum value of 6% and a maximum value of 12%, with an expected value of 9%. Non-operating expenses are anticipated to be \$1.5 million a year.
- **Revenue growth**: We used a slightly modified version of the actual distribution of historical real GDP changes as the distribution of future changes in real GDP.[4] The average real GDP growth over the period was 3%, but there is substantial variation with the worst year delivering a drop in real GDP of more than 8% and the best an increase of more than 8%. The expected annual growth rate in revenues is the sum of the expected inflation rate and the growth rate in real GDP. We will assume that the expected inflation rate is 2%.
- The store is expected to generate cash flows for 10 years and there is no expected salvage value from the store closure.
- The cost of capital for the Home Depot is 10% and the tax rate is 40%.

We can compute the value of this store to the Home Depot, based entirely upon the expected values of each variable:

Expected base-year revenue = \$44 million

Expected base-year after-tax cash flow = (Revenue × Pretax margin – Non-operating expenses)(1 – Tax rate) = (44 × 0.09 – 1.5)(1 – 0.4) = \$1.476 million

Expected growth rate = GDP growth rate + Expected inflation = 3% + 2% = 5%

4 In the modified version, we smoothed out the distribution to fill in the missing intervals and moved the peak of the distribution slightly to the left (to 3-4% from 4-5%) reflecting the larger size of the economy today.

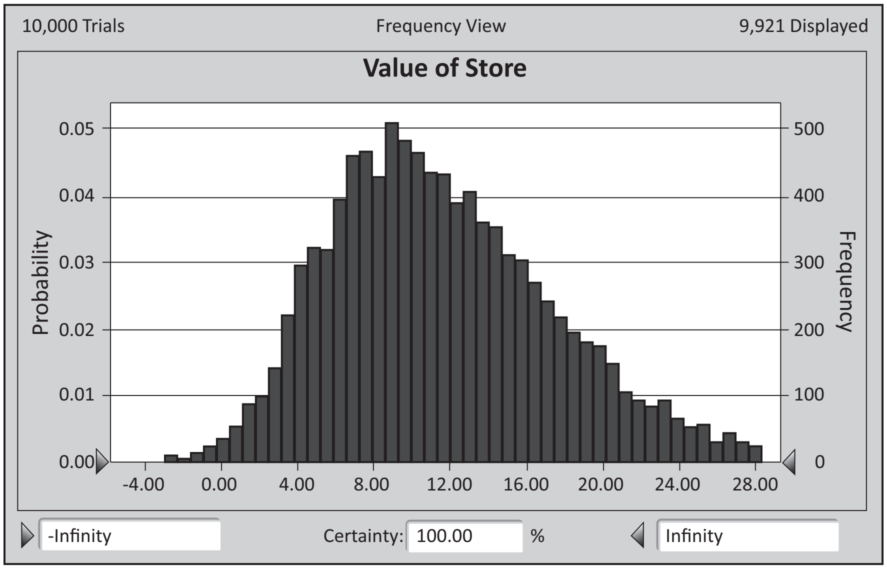
Value[5] of store

$$= \mathrm{CF}(1+g)\frac{\left(1-\dfrac{(1+g)^n}{(1+r)^n}\right)}{r-g} = 1.476(1.05)\frac{\left(1-\dfrac{1.05^{10}}{1.10^{10}}\right)}{0.10-0.05} = \$11.53\text{ million}$$

The risk-adjusted value for this store is \$11.53 million.

We then did a simulation with 10,000 runs, based upon the probability distributions for each of the inputs.[6] The resulting values are graphed in Figure 6.11:

Figure 6.11 Distribution of Estimated Values for HD Store from Simulation



The key statistics on the values obtained across the 10,000 runs are summarized below:

- The average value across the simulations was \$11.67 million, a trifle higher than the risk-adjusted value of \$11.53 million; the median value was \$10.90 million.
- There was substantial variation in values, with the lowest value across all runs of –\$5.05 million and the highest value of \$39.42 million; the standard deviation in values was \$5.96 million.

2.3 Use in Decision Making

A well-done simulation provides us with more than just an expected value for an asset or investment.

1 **Better input estimation**: In an ideal simulation, analysts will examine both the historical and cross sectional data on each input variable before making a judgment on what distribution to use and the parameters of the distribution. In the process, they may be able to avoid the sloppiness that is associated with the use

5 The equation presented here is the equation for the present value of a growing annuity.
6 We used Crystal Ball as the computational program. Crystal Ball is a simulation program produced by Decisioneering, Inc.

of "single best" estimates; many discounted cash flow valuations are based upon expected growth rates that are obtained from services such as Zack's or IBES, which report analysts' consensus estimates.

2 **It yields a distribution for expected value rather than a point estimate**: Consider the valuation example that we completed in the last section. In addition to reporting an expected value of $11.67 million for the store, we also estimated a standard deviation of $5.96 million in that value and a breakdown of the values, by percentile. The distribution reinforces the obvious but important point that valuation models yield estimates of value for risky assets that are imprecise and explains why different analysts valuing the same asset may arrive at different estimates of value.

Note that there are two claims about simulations that we are unwilling to make. The first is that simulations yield better estimates of expected value than conventional risk-adjusted value models. In fact, the expected values from simulations should be fairly close to the expected value that we would obtain using the expected values for each of the inputs (rather than the entire distribution). The second is that simulations, by providing estimates of the expected value and the distribution in that value, lead to better decisions. This may not always be the case since the benefits that decision-makers get by getting a fuller picture of the uncertainty in value in a risky asset may be more than offset by misuse of that risk measure. It is all too common for risk to be double counted in simulations and for decisions to be based upon the wrong type of risk.

2.4 Simulations with Constraints

To use simulations as a tool in risk analysis, we have to introduce a constraint, which, if violated, creates very large costs for the firm and perhaps even causes its demise. We can then evaluate the effectiveness of risk hedging tools by examining the likelihood that the constraint will be violated with each one and weighing that off against the cost of the tool. In this section, we will consider some common constraints that are introduced into simulations.

2.4.1 *Book Value Constraints*

The book value of equity is an accounting construct and, by itself, means little. Firms like Microsoft and Google trade at market values that are several times their book values. At the other extreme, there are firms that trade at half their book value or less. In fact, there are several hundred firms in the United States, some with significant market values that have negative book values for equity. There are two types of restrictions on book value of equity that may call for risk hedging.

1 **Regulatory Capital Restrictions**: Financial service firms such as banks and insurance companies are required to maintain book equity as a fraction of loans or other assets at or above a floor ratio specified by the authorities. Firms that violate these capital constraints can be taken over by the regulatory authorities with the possibility of equity investors losing everything if that occurs. Not surprisingly, financial service firms not only keep a close eye on their book value of equity (and the related ratios) but are also conscious of the possibility that the risk in their investments or positions can manifest itself as a drop in book equity. Value at risk, or VAR, represents efforts by financial service firms to understand the potential risks in their investments and to be ready for the possibility of a catastrophic outcome, though the probability of it occurring might be very small. By simulating the values of their investments under a variety of scenarios, they can identify not only the possibility of falling below the

regulatory ratios but also look for ways of hedging against this event occurring. The payoff to risk hedging then manifests itself as a decline in or even an elimination of the probability that the firm will violate a regulatory constraint.

2 **Negative Book Value for Equity**: As noted, there are hundreds of firms in the United States with negative book values of equity that survive its occurrence and have high market values for equity. There are some countries where a negative book value of equity can create substantial costs for the firms and its investors. For instance, companies with negative book values of equity in parts of Europe are required to raise fresh equity capital to bring their book values above zero. In some countries in Asia, companies that have negative book values of equity are barred from paying dividends. Even in the United States, lenders to firms can have loan covenants that allow them to gain at least partial control of a firm if its book value of equity turns negative. As with regulatory capital restrictions, we can use simulations to assess the probability of a negative book value for equity and to protect against it.

2.4.2 *Earnings and Cash Flow Constraints*

Earnings and cash flow constraints can be either internally or externally imposed. In some firms, managers may decide that the consequences of reporting a loss or not meeting analysis estimates of earnings are so dire, including perhaps the loss of their jobs, that they are willing to expend the resources on risk hedging products to prevent this from happening. The payoff from hedging risk then has nothing to do with firm value maximization and much to do with managerial compensation and incentives. In other firms, the constraints on earnings and cash flows can be externally imposed. For instance, loan covenants can be related to earnings outcomes. Not only can the interest rate on the loan be tied to whether a company makes money or not, but the control of the firm can itself shift to lenders in some cases if the firm loses money. In either case, we can use simulations to both assess the likelihood that these constraints will be violated and to examine the effect of risk hedging products on this likelihood.

2.4.3 *Market Value Constraints*

In discounted cash flow valuation, the value of the firm is computed as a going concern, by discounting expected cash flows at a risk-adjusted discount rate. Deducting debt from this estimate yields equity value. The possibility and potential costs of not being able to meet debt payments is considered only peripherally in the discount rate. In reality, the costs of not meeting contractual obligations can be substantial. In fact, these costs are generally categorized as indirect bankruptcy costs and could include the loss of customers, tighter supplier credit, and higher employee turnover. The perception that a firm is in trouble can lead to further trouble. By allowing us to compare the value of a business to its outstanding claims in all possible scenarios (rather than just the most likely one), simulations allow us to not only quantify the likelihood of distress but also build in the cost of indirect bankruptcy costs into valuation. In effect, we can explicitly model the effect of distress on expected cash flows and discount rates.

2.5 Issues

The use of simulations in investment analysis was first suggested in an article by David Hertz in the Harvard Business Review.[7] He argued that using probability distributions for input variables, rather than single best estimates, would yield more informative

7 Hertz, D. 1964, "Risk Analysis in Capital Investment," *Harvard Business Review*.

output. In the example that he provided in the paper, he used simulations to compare the distributions of returns of two investments: the investment with the higher expected return also had a higher chance of losing money (which was viewed as an indicator of its riskiness). In the aftermath, there were several analysts who jumped on the simulation bandwagon, with mixed results. In recent years, there has been a resurgence in interest in simulation as a tool for risk assessment, especially in the context of using and valuing derivatives. There are several key issues, though, that we have to deal with in the context of using simulations in risk assessment:

1 **Garbage in, garbage out**: For simulations to have value, the distributions chosen for the inputs should be based upon analysis and data, rather than guesswork. It is worth noting that simulations yield great-looking output, even when the inputs are random. Unsuspecting decision makers may therefore be getting meaningless pictures of the risk in an investment. It is also worth noting that simulations require more than a passing knowledge of statistical distributions and their characteristics; analysts who cannot assess the difference between normal and lognormal distributions should not be doing simulations.

2 **Real data may not fit distributions**: The problem with the real world is that the data seldom fits the stringent requirements of statistical distributions. Using probability distributions that bear little resemblance to the true distribution underlying an input variable will yield misleading results.

3 **Non-stationary distributions**: Even when the data fits a statistical distribution or where historical data distributions are available, shifts in the market structure can lead to shifts in the distributions as well. In some cases, this can change the form of the distribution and in other cases, it can change the parameters of the distribution. Thus, the mean and variance estimated from historical data for an input that is normally distributed may change for the next period. What we would really like to use in simulations, but seldom can assess, are forward looking probability distributions.

4 **Changing correlation across inputs**: In the third simulation step, we noted that correlation across input variables can be modeled into simulations. However, this works only if the correlations remain stable and predictable. To the extent that correlations between input variables change over time, it becomes far more difficult to model them.

2.6 Risk-Adjusted Value and Simulations

A common misconception is that decision trees are risk adjusted because they consider the likelihood of adverse events. The same misconception is prevalent in simulations, where the argument is that the cash flows from simulations are somehow risk adjusted because of the use of probability distributions and that the risk-free rate should be used in discounting these cash flows. With one exception, this argument does not make sense. Looking across simulations, the cash flows that we obtain are expected cash flows and are not risk adjusted. Consequently, we should be discounting these cash flows at a risk-adjusted rate.

The exception occurs when you use the standard deviation in values from a simulation as a measure of investment or asset risk and make decisions based upon that measure. In this case, using a risk-adjusted discount rate will result in a double counting of risk. Consider a simple example. Assume that you are trying to choose between two assets, both of which you have valued using simulations and risk-adjusted discount rates. Table 6.3 summarizes your findings:

Table 6.3 Results of Simulation

Asset	Risk-Adjusted Discount Rate	Simulation Expected Value	Simulation Standard Deviation
A	12%	$100	15%
B	15%	$100	21%

Note that you view Asset B to be riskier and have used a higher discount rate to compute value. If you now proceed to reject Asset B, because the standard deviation is higher across the simulated values, you would be penalizing it twice. You can redo the simulation using the risk-free rate as the discount rate for both assets, but a note of caution needs to be introduced. If we then base our choice between these assets on the standard deviation in simulated values, we are assuming that all risk matters in investment choice, rather than only the risk that cannot be diversified away. Put another way, we may end up rejecting an asset because it has a high standard deviation in simulated values, even though adding that asset to a portfolio may result in little additional risk (because much of its risk can be diversified away).

This is not to suggest that simulations are not useful to us in understanding risk. Looking at the variance of the simulated values around the expected value provides a visual reminder that we are estimating value in an uncertain environment. It is also conceivable that we can use it as a decision tool in portfolio management in choosing between two stocks that are equally undervalued but have different value distributions. The stock with the less volatile value distribution may be considered a better investment than another stock with a more volatile distribution.

3 AN OVERALL ASSESSMENT OF PROBABILISTIC RISK ASSESSMENT APPROACHES

Now that we have looked at scenario analysis, decision trees, and simulations, we can consider not only when each one is appropriate but also how these approaches complement or replace risk-adjusted value approaches.

3.1 Comparing the Approaches

Assuming that we decide to use a probabilistic approach to assess risk and could choose between scenario analysis, decision trees, and simulations, which one should we pick? The answer will depend upon how you plan to use the output and what types of risk you are facing:

1 **Selective versus full risk analysis**: In the best-case/worst-case scenario analysis, we look at only three scenarios (the best case, the most likely case, and the worst case) and ignore all other scenarios. Even when we consider multiple scenarios, we will not have a complete assessment of all possible outcomes from risky investments or assets. With decision trees and simulations, we attempt to consider all possible outcomes. In decision trees, we try to accomplish this by converting continuous risk into a manageable set of possible outcomes. With simulations, we use probability distributions to capture all possible outcomes. Put in terms of probability, the sum of the probabilities of the scenarios we examine in scenario analysis can be less than one, whereas the sum of the probabilities of outcomes in decision trees and simulations has to equal one. As

a consequence, we can compute expected values across outcomes in the latter, using the probabilities as weights, and these expected values are comparable to the single estimate risk-adjusted values arrived at using discounted cash flow and relative valuation models.

2 **Type of risk**: As noted above, scenario analysis and decision trees are generally built around discrete outcomes in risky events whereas simulations are better suited for continuous risks. Focusing on just scenario analysis and decision trees, the latter are better suited for sequential risks, since risk is considered in phases, whereas the former is easier to use when risks occur concurrently.

3 **Correlation across risks**: If the various risks that an investment is exposed to are correlated, simulations allow for explicitly modeling these correlations (assuming that you can estimate and forecast them). In scenario analysis, we can deal with correlations subjectively by creating scenarios that allow for them; the high (low) interest rate scenario will also include slower (higher) economic growth. Correlated risks are difficult to model in decision trees.

Table 6.4 summarizes the relationship between risk type and the probabilistic approach used:

Table 6.4 Risk Type and Probabilistic Approaches

Discrete/ Continuous	Correlated/ Independent	Sequential/ Concurrent	Risk Approach
Discrete	Independent	Sequential	Decision tree
Discrete	Correlated	Concurrent	Scenario analysis
Continuous	Either	Either	Simulations

Finally, the quality of the information will be a factor in your choice of approach. Since simulations are heavily dependent upon being able to assess probability distributions and parameters, they work best in cases where there is substantial historical and cross sectional data available that can be used to make these assessments. With decision trees, you need estimates of the probabilities of the outcomes at each chance node, making them best suited for risks that can be assessed either using past data or population characteristics. Thus, it should come as no surprise that when confronted with new and unpredictable risks, analysts continue to fall back on scenario analysis, notwithstanding its subjective ways of dealing with risk.

3.2 Complement or Replacement for Risk-Adjusted Value

Both decision trees and simulations are approaches that can be used as either complements to or substitutes for risk-adjusted value. Scenario analysis, on the other hand, will always be a complement to risk-adjusted value, since it does not look at the full spectrum of possible outcomes.

When any of these approaches are used as complements to risk-adjusted value, caveats continue to apply and bear repeating. All of these approaches use expected rather than risk-adjusted cash flows and *the discount rate that is used should be a risk-adjusted discount rate*: the risk-free rate cannot be used to discount expected cash flows. In all three approaches, though, we still preserve the flexibility to change the risk-adjusted discount rate for different outcomes. Since all of these approaches will also provide a range for estimated value and a measure of variability (in terms of value at the end nodes in a decision tree or as a standard deviation in value in a simulation), it is important that we do not double count for risk. In other words, it is patently unfair to risky investments to discount their cash flows back at a risk-adjusted rate (in simulations and decision trees) and to then reject them because the variability in value is high.

Both simulations and decision trees can be used as alternatives to risk-adjusted valuation, but there are constraints on the process. The first is that the cash flows will be discounted back at a risk-free rate to arrive at value. The second is that we now use the measure of variability in values that we obtain in both these approaches as a measure of risk in the investment. Comparing two assets with the same expected value (obtained with riskless rates as discount rates) from a simulation, we will pick the one with the lower variability in simulated values as the better investment. If we do this, we are assuming that all of the risks that we have built into the simulation are relevant for the investment decision. In effect, we are ignoring the line drawn between risks that could have been diversified away in a portfolio and asset-specific risk on which much of modern finance is built. For an investor considering investing all of his or her wealth in one asset, this should be reasonable. For a portfolio manager comparing two risky stocks that he or she is considering adding to a diversified portfolio or for a publicly traded company evaluating two projects, it can yield misleading results; the rejected stock or project with the higher variance in simulated values may be uncorrelated with the other investments in the portfolio and thus have little marginal risk.

3.3 In Practice

The use of probabilistic approaches has become more common with the surge in data availability and computing power. It is not uncommon now to see a capital budgeting analysis, with twenty to thirty additional scenarios, or a Monte Carlo simulation attached to an equity valuation. In fact, the ease with which simulations can be implemented has allowed its use in a variety of new markets.

- **Deregulated electricity markets**: As electricity markets have been deregulated around the world, companies involved in the business of buying and selling electricity have begun using simulation models to quantify the swings in demand and supply of power, and the resulting price volatility. The results have been used to determine how much should be spent on building new power plants and how best to use the excess capacity in these plants.
- **Commodity companies**: Companies in commodity businesses—oil and precious metals, for instance—have used probabilistic approaches to examine how much they should bid for new sources for these commodities, rather than relying on a single best estimate of the future price. Analysts valuing these companies have also taken to modeling the value of these companies as a function of the price of the underlying commodity.
- **Technology companies**: Shifts in technology can be devastating for businesses that end up on the wrong side of the shift. Simulations and scenario analyses have been used to model the effects on revenues and earnings of the entry and diffusion of new technologies.

Simulations are a key component of Value at Risk and other risk management tools used, especially in firms that have to deal with risk in financial assets.

CONCLUSION

4

Estimating the risk-adjusted value for a risky asset or investment may seem like an exercise in futility. After all, the value is a function of the assumptions that we make about how the risk will unfold in the future. With probabilistic approaches to risk assessment, we estimate not only an expected value but also get a sense of the range of possible outcomes for value, across good and bad scenarios.

- In the most extreme form of scenario analysis, you look at the value in the best case and worst case scenarios and contrast them with the expected value. In its more general form, you estimate the value under a small number of likely scenarios, ranging from optimistic to pessimistic.
- Decision trees are designed for sequential and discrete risks, where the risk in an investment is considered into phases and the risk in each phase is captured in the possible outcomes and the probabilities that they will occur. A decision tree provides a complete assessment of risk and can be used to determine the optimal courses of action at each phase and an expected value for an asset today.
- Simulations provide the most complete assessments of risk since they are based upon probability distributions for each input (rather than a single expected value or just discrete outcomes). The output from a simulation takes the form of an expected value across simulations and a distribution for the simulated values.

With all three approaches, the keys are to avoid double counting risk (by using a risk-adjusted discount rate and considering the variability in estimated value as a risk measure) or making decisions based upon the wrong types of risk.

PRACTICE PROBLEMS

The following information relates to Questions 1–7

Alicia Maxwell, an analyst for a REIT, is evaluating the potential purchase of a hotel property. She plans to use simulation analysis to estimate the distribution of the property's annual operating cash flow for the next five years.

Revenue Construction

Maxwell recognizes that annual gross revenue for the property depends on the nightly room rate and the occupancy rate. She believes that the primary driver for the nightly room rate is the Employment Cost Index (ECI) and that the primary driver for the occupancy rate is the Consumer Sentiment Index (CSI). In the process of simulating revenues, she examines the ECI and the CSI quarterly over the past 20 years and their relation to the nightly room rate and occupancy rates of the REIT's existing properties, respectively. She estimates the following:

Nightly room rate = \$23 + 0.9(ECI_{t-1})

Occupancy rate = 0.25 + 0.7(CSI_{t-1})

Occupancy rates are assumed to be non-negative and cannot exceed 100%.

Maxwell generates 10,000 trials of the ECI and CSI based on the historical mean level of the indexes and their monthly standard deviations. Although the distribution of historical CSI is not symmetric, she assumes that both ECI and CSI are normally distributed. Maxwell is aware that if the inputs are correlated, this may present a problem. She also observes that the CSI and the ECI are correlated with one another and that the relation between the CSI and the occupancy rate is stronger than that between the ECI and nightly room rates. Maxwell estimates the corresponding nightly room rate and the occupancy rate based on these historical relations, multiplies these by the number of hotel nights in a year, and generates 10,000 estimates of annual gross revenue.

Expense Assumptions

Maxwell examines current expenses for the REIT's other hotel properties and selects the distributions for the simulation of operating expenses and management fees. Maxwell estimates operating expenses to be uniformly distributed between 68% and 70% of revenues and that the property management fee for the hotel is uniformly distributed between 5.9% and 6.1% of total annual revenue.

Simulation Results and Analysis

Maxwell has three concerns regarding the results from the simulation trials:

Concern 1: Although the distribution of historical CSI is slightly skewed, Maxwell uses the normal distribution to simulate the monthly CSI.

Concern 2: Property management firms may demand higher property management fees (that is, a higher percentage of revenue) when the CSI is lower.

Concern 3: When comparing the distribution of CSI over the past 20 years with those of the past 30 years, she notices a substantial difference in the mean and standard deviation of the CSI distribution.

In completing her analysis, Maxwell considers her choice of simulation analysis over alternative approaches, such as decision trees and scenario analysis, to be justified. Although scenario analysis and decision trees both consider possible outcomes, neither can be used easily when correlated variables are present, as is the case with CSI and ECI. Further, she notes that, compared with scenario analysis and decision trees, simulation is best suited for continuous risks, whether they be concurrent or sequential.

Based on the results of the simulation analysis, the REIT acquires the hotel. One year later, the REIT is considering the acquisition of another hotel, and Maxwell wants to use the same simulation model. Based on an analysis of the hotel industry, Maxwell notes that recent mergers in the industry have affected competition in the market in which this hotel operates. Consequently, Maxwell needs to update the simulation model.

1 With respect to forecasting annual gross revenue, Maxwell's *best* course of action to deal with the inputs problem should be to:

   A simulate the ECI and the CSI independently.

   B build the correlation explicitly into the simulation.

   C estimate revenues using the ECI only and eliminate the CSI from the simulation.

2 With respect to the simulation, Maxwell should be concerned that:

   A CSI does not follow a symmetric distribution.

   B property management fees are not normally distributed.

   C there is a lack of correlation between the CSI and the management fee percentage.

3 Maxwell's distribution assumption for the property management fee in the simulation is based on:

   A historical data.

   B simulation results.

   C cross-sectional data.

4 Which of Maxwell's assumptions can be *best* described as a simulation constraint? The assumption about the:

   A occupancy rate.

   B operating expenses.

   C property management fee.

5 Which of Maxwell's three concerns can be *best* described as being attributable to non-stationary distributions?

   A Concern 1

   B Concern 2

   C Concern 3

6 Maxwell's justifications for her choice of simulation analysis are correct with respect to:

   A correlated risks.

   B continuous risks.

   C both correlated risks and continuous risks.

7 Considering industry changes over the past year, one update that Maxwell should make to the simulation model is the:

A choice of the distribution of CSI and ECI.

B relation between CSI and occupancy rates.

C removal of constraints on occupancy rates.

SOLUTIONS

1 B is correct. Both the nightly room rate and the occupancy rate are dependent on inputs, ECI and CSI, which are correlated. If there are correlated inputs, there are two solutions to this problem. One is to allow only one of the two inputs to vary, emphasizing the one with the larger impact. The second solution is to the build the correlation explicitly into the simulation. Simulating ECI and CSI independently is not a remedy for the correlated inputs problem. Further, the relation between CSI and the occupancy rate is stronger than that between ECI and nightly room rates, which suggests that CSI should be kept if one of the two inputs is to be removed from the simulation. A is incorrect because the two input variables, ECI and CSI, are correlated, and simulating them independently is not appropriate. The remedies include dropping one of the probabilistic inputs or building the correlation into the simulation explicitly. C is incorrect because the relation between CSI and the occupancy rate is stronger than that between ECI and nightly room rates, which suggests that CSI should be kept if one of the two inputs is to be removed from the simulation.

2 A is correct. Using a normal distribution when the distribution is not normal may lead to misleading results. There is no limitation on the type of probability distribution (for example, normal or uniform); what is important is that the selected distribution reflect the likely distribution of future values. Further, there is no requirement that inputs be correlated with one another. In fact, correlated inputs present issues that must be overcome by either removing one of the probability inputs or explicitly building the correlation into the simulation. B is incorrect because probabilistic variables may follow different distributions, including the normal distribution and the uniform distribution. It is important to pick a statistical distribution that best captures the variability in the input and estimate the parameters for that distribution. C is incorrect because there is no requirement that inputs be correlated with one another. In fact, correlated inputs present issues that must be overcome by either removing one of the probability inputs or explicitly building the correlation into the simulation.

3 C is correct. Maxwell uses the distribution of property management fees for the REIT's other hotel properties to simulate the property management fee. Therefore, the property management fee distribution is based on differences in property management fees across a cross-section of the REIT's existing hotel properties that are similar to the hotel being analyzed. A is incorrect because Maxwell's distribution assumption about the property management fee is not based on historical data for the hotel. The property is new, and therefore Maxwell does not have a history of reliable data for the property management fee to use. B is incorrect because the property management fee distributional assumption is not based on the results of a simulation but rather is based on cross-sectional data. The property management fee is based on the property management fees for the REIT's other hotel properties. Once the distributional assumption is made (that is, the statistical distribution and parameters), Maxwell may then incorporate these into the trials for the management fee.

4 A is correct. Without Maxwell's assumption regarding the constraints that the occupancy rates be between 0% and 100%, the simulation could produce negative occupancy rates or rates above 100%. This assumption serves as a constraint on occupancy rates. Operating expenses are assumed to be a percentage of revenues and are not constrained; what is specified with respect to operating expenses is the distribution. Similarly, for the property management fee, the distribution is specified, but this is not a constraint. B is incorrect because the

assumption of a specific relation between operating expenses and revenues is not a constraint. C is incorrect because the assignment of a distribution, in this case a uniform distribution for the property management fee, is not a constraint. It is merely a step in a simulation in which the statistical distribution and parameters are specified.

5 C is correct. A substantial difference in the mean and standard deviation of CSI over the past 20 years relative to those of the past 30 years suggests a change in (i.e., non-stationary) distribution. Even when the data fits a statistical distribution or when historical data distributions are available, shifts in the market structure may lead to shifts in the distributions as well, as evidenced by the shift between the 20- and 30-year distributions of CSI. In some cases, such shifts can change the form of the distribution and in other cases, they can change the parameters of the distribution. A is incorrect because Concern 1 is a description of real data not fitting distributions. The problem with the real world is that the data seldom fits the stringent requirements of statistical distributions. Using probability distributions that bear little resemblance to the true distribution underlying an input variable will yield misleading results. B is incorrect because Concern 2 is an issue of correlation that is not incorporated into the simulation. Correlation across input variables can be modeled into simulations. This works only if the correlations remain stable and predictable, however. To the extent that correlations between input variables change over time, as is expressed in Concern 2, it becomes far more difficult to model them.

6 C is correct. Correlated risks are difficult to model in decision trees. In addition, adjusting scenario analysis for correlated risks is subjective. Scenario analysis and decision trees are generally built around discrete outcomes in risky events, whereas simulations are better suited for continuous risks. Further, if the various risks to which an investment is exposed are correlated, simulations allow for explicitly modeling these correlations if they can be estimated.

7 B is correct. Even when the data fits a statistical distribution or when historical data distributions are available, shifts in the market structure can lead to shifts in the distributions as well. In some cases, these shifts can change the form of the distribution, and in other cases, they can change the parameters of the distribution. Thus, relations from historical data for an input may change for the next period, affecting the relation between the nightly room rate and CSI, as well as the relation between the occupancy rate and ECI.

The constraints on the occupancy rates are necessary because without them, there may be unrealistic results (such as negative rates or rates exceeding 100%). Further, the choice of the distributions of CSI and ECI is not mentioned to be affected by the changes in the industry over the past year; rather, the relation between CSI and occupancy rates may change.

Economics

STUDY SESSION

Study Session 4	Economics

TOPIC LEVEL LEARNING OUTCOME

The candidate should be able to explain and demonstrate the use of economic concepts and methods in the determination and forecasting of currency exchange rates, the analysis of economic growth, and the analysis of business and financial market regulation.

A country's exchange rates, level of economic activity, and regulatory environment have significant implications for companies operating within its borders. Although predicting exchange rates is extremely difficult, exchange rate equilibrium relationships provide valuable insights for understanding the currency risks inherent in overseas operations and international investments.

ECONOMICS

STUDY SESSION 4

Economics

This study session begins with fundamental foreign exchange concept and theories of exchange rate determination. As a means to understanding exchange rate risk exposures, discussion centers on theoretical long-term equilibrium values. International parity conditions and the carry trade are described. Both direct (capital controls, foreign exchange intervention) and indirect (monetary, fiscal policy) exchange rate influencers are considered. A discussion of long-term growth and its relationship to investment returns follows. The three theories of growth (classical, neoclassical, endogenous) are presented. The session concludes with an overview of regulation, including the types of regulation, roles played by regulation, and considerations to use when evaluating the effects of regulation on an industry.

READING ASSIGNMENTS

Reading 10	Currency Exchange Rates: Understanding Equilibrium Value by Michael R. Rosenberg and William A. Barker, PhD, CFA
Reading 11	Economic Growth and the Investment Decision by Paul R. Kutasovic, PhD, CFA
Reading 12	Economics of Regulation by Chester S. Spatt, PhD

READING

10

Currency Exchange Rates: Understanding Equilibrium Value

by Michael R. Rosenberg and William A. Barker, PhD, CFA

Michael R. Rosenberg (USA). William A. Barker, PhD, CFA (Canada)

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	**a.** calculate and interpret the bid–offer spread on a spot or forward currency quotation and describe the factors that affect the bid–offer spread;
☐	**b.** identify a triangular arbitrage opportunity and calculate its profit, given the bid–offer quotations for three currencies;
☐	**c.** distinguish between spot and forward rates and calculate the forward premium/discount for a given currency;
☐	**d.** calculate the mark-to-market value of a forward contract;
☐	**e.** explain international parity conditions (covered and uncovered interest rate parity, forward rate parity, purchasing power parity, and the international Fisher effect);
☐	**f.** describe relations among the international parity conditions;
☐	**g.** evaluate the use of the current spot rate, the forward rate, purchasing power parity, and uncovered interest parity to forecast future spot exchange rates;
☐	**h.** explain approaches to assessing the long-run fair value of an exchange rate;
☐	**i.** describe the carry trade and its relation to uncovered interest rate parity and calculate the profit from a carry trade;
☐	**j.** explain how flows in the balance of payment accounts affect currency exchange rates;
☐	**k.** explain the potential effects of monetary and fiscal policy on exchange rates;
☐	**l.** describe objectives of central bank or government intervention and capital controls and describe the effectiveness of intervention and capital controls;
☐	**m.** describe warning signs of a currency crisis.

1 INTRODUCTION

Exchange rates are well known to follow a random walk, whereby fluctuations from one day to the next are unpredictable. The business of currency forecasting can be a humbling experience. Alan Greenspan, former chairman of the US Federal Reserve Board, famously noted that "having endeavored to forecast exchange rates for more than half a century, I have understandably developed significant humility about my ability in this area."

Hence, this reading is not about predicting exchange rates, but about the tools the reader can use to better understand long-run equilibrium value. This outlook helps guide the market participant's decisions with respect to risk exposures, as well as whether currency hedges should be implemented and, if so, how they should be managed. After discussing the basics of exchange rate transactions, this reading presents the main theories for currency determination—starting with the international parity conditions—and then describes other important influences, such as current account balances, capital flows, and monetary and fiscal policy.

Although these fundamentals-based models usually perform poorly in predicting future exchange rates in the short run, they are crucial for understanding long-term currency value. Thus, the reading proceeds as follows:

- Section 2 reviews the basic concepts of the foreign exchange market covered in the CFA Level I curriculum reading and expands this previous coverage to incorporate more material on bid–offer spreads.
- In Section 3, we begin to examine determinants of exchange rates, starting with longer-term interrelationships among exchange rates, interest rates, and inflation rates embodied in the international parity conditions. These parity conditions form the key building blocks for many long-run exchange rate models.
- Section 4 examines the FX carry trade, a trading strategy that exploits deviations from uncovered interest rate parity.
- Section 5 examines the relationship between a country's exchange rate and its balance of payments.
- In Section 6, we examine how monetary and fiscal policies can *indirectly* affect exchange rates by influencing the various factors described in our exchange rate model from Section 3.
- Section 7 examines *direct* public sector actions in foreign exchange markets, both through capital controls and by foreign exchange market intervention (buying and selling currencies for policy purposes).
- Section 8 examines historical episodes of currency crisis and some leading indicators that may signal the increased likelihood of a crisis.

A final section summarizes the key points of the reading.

2 FOREIGN EXCHANGE MARKET CONCEPTS

We begin with a brief review of some of the basic conventions of the FX market that were covered in the CFA Level I curriculum. In this section, we cover (1) the basics of exchange rate notation and pricing, (2) arbitrage pricing constraints on spot rate foreign exchange quotes, and (3) forward rates and covered interest rate parity.

An exchange rate is the price of the *base* currency expressed in terms of the *price* currency. For example, a USD/EUR rate of 1.3650 means the euro, the base currency, costs 1.3650 US dollars (an appendix defines the three-letter currency codes used in this reading). The exact notation used to represent exchange rates can vary widely between sources, and occasionally the same exchange rate notation will be used by different sources to mean completely different things. *The reader should be aware that the notation used here may not be the same as that encountered elsewhere.* To avoid confusion, this reading will identify exchange rates using the convention of "P/B," referring to the price of the base currency "B" expressed in terms of the price currency "P."[1]

The spot exchange rate is usually used for settlement on the second business day after the trade date, referred to as $T + 2$ settlement.[2] In foreign exchange markets—as in other financial markets—market participants are presented with a two-sided price in the form of a bid price and an offer price (also called an ask price) quoted by potential counterparties. The bid price is the price, defined in terms of the price currency, at which the counterparty is willing to buy one unit of the base currency. Similarly, the offer price is the price, in terms of the price currency, at which that counterparty is willing to sell one unit of the base currency. For example, given a price request from a client, a dealer might quote a two-sided price on the spot USD/EUR exchange rate of 1.3648/1.3652. This means that the dealer is willing to pay USD 1.3648 to buy one euro and that the dealer is willing to sell one euro for USD 1.3652.

There are two points to bear in mind about bid–offer quotes:

1 *The offer price is always higher than the bid price.* The bid–offer spread—the difference between the offer price and the bid price—is the compensation that counterparties seek for providing foreign exchange to other market participants.

2 *The party in the transaction who requests a two-sided price quote has the option (but not the obligation) to deal at either the bid (to sell the base currency) or the offer (to buy the base currency) quoted by the dealer.* If the party chooses to trade at the quoted prices, the party is said to have either "*hit the bid*" or "*paid the offer*." If the base currency is being sold, the party is said to have hit the bid. If the base currency is being bought, the party is said to have paid the offer.

We will distinguish here between the bid–offer pricing *a client receives from a dealer* and the pricing *a dealer receives from the interbank market*. Dealers buy and sell foreign exchange among themselves in what is called the interbank market.[3] This global network for exchanging currencies among professional market participants allows dealers to adjust their inventories and risk positions, distribute foreign currencies to end users who need them, and transfer foreign exchange rate risk to market participants who are willing to bear it. The interbank market is typically for dealing

1 Notation is generally not standardized in global foreign exchange markets, and there are several common ways of expressing the same currency pair (e.g., JPY/USD, USD:JPY, $/¥). What is common in FX markets, however, is the concept of a "base" and a "price" currency when setting exchange rates. Later in the reading, we will sometimes switch to discussing a "domestic" and a "foreign" currency, quoted as foreign/domestic (*f/d*). *This is only an illustrative device* for more easily explaining various theoretical concepts. The candidate should be aware that currency pairs are not described in terms of "foreign" and "domestic" currencies in professional FX markets. This is because what is the "foreign" and what is the "domestic" currency depend on where one is located, which can lead to confusion. For instance, what is "foreign" and what is "domestic" for a Middle Eastern investor trading CHF against GBP with the New York branch of a European bank, with the trade ultimately booked at the bank's headquarters in Paris?

2 The exception among the major currencies is CAD/USD, for which standard spot settlement is $T + 1$.

3 Although it is known as the interbank market, many non-bank entities now can access this market. These non-bank entities include institutional asset managers, hedge funds, and other large, sophisticated market participants.

sizes of at least 1 million units of the base currency. Of course, the dealing amount can be larger than 1 million units; indeed, interbank market trades generally are measured in terms of multiples of a million units of the base currency.

The bid–offer spread a dealer provides to most clients typically is slightly wider than the bid–offer spread observed in the interbank market. Most currencies, except for the yen, are quoted to four decimal places. The fourth decimal place (0.0001) is referred to as a "pip." The yen is typically quoted to just two decimal places; in yen quotes, the second decimal place (0.01) is referred to as a pip.

For example, if the quote in the interbank USD/EUR spot market is 1.3649/1.3651 (two pips wide), the dealer might quote a client a bid–offer of 1.3648/1.3652 (four pips wide) for a spot USD/EUR transaction. When the dealer buys (sells) the base currency from (to) a client, the dealer is typically expecting to quickly turn around and sell (buy) the base currency in the interbank market. This offsetting transaction allows the dealer to divest the risk exposure assumed by providing a two-sided price to the client and to hopefully make a profit. Continuing our example, suppose the dealer's client hits the dealer's bid and sells EUR to the dealer for USD 1.3648. The dealer is now long EUR (and short USD) and wants to cover this position in the interbank market. To do this, the dealer sells the EUR in the interbank market by hitting the interbank bid. As a result, the dealer *bought* EUR from the client at USD 1.3648 and then *sold* the EUR in the interbank for USD 1.3649. This gives the dealer a profit of USD 0.0001 (one pip) for every EUR transacted. This one pip translates into a profit of USD 100 per EUR million bought from the client. If, instead of hitting his bid, the client paid the offer (1.3652), then the dealer could pay the offer in the interbank market (1.3651), earning a profit of one pip.

The size of the bid–offer spread quoted to dealers' clients in the FX market can vary widely across exchange rates and is not constant over time, even for a single exchange rate. The size of this spread depends primarily on three factors:

- the bid–offer spread in the interbank foreign exchange market for the two currencies involved,
- the size of the transaction, and
- the relationship between the dealer and the client.

We examine each factor in turn.

The size of the bid–offer spread quoted in the interbank market depends on the liquidity in this market. Liquidity is influenced by several factors:

1 *The currency pair involved.* Market participation is greater for some currency pairs than for others. Liquidity in the major currency pairs—for example, USD/EUR, JPY/USD, and USD/GBP—can be quite high. These markets are almost always deep, with multiple bids and offers from market participants around the world. In other currency pairs, particularly some of the more obscure currency cross rates (e.g., MXN/CHF), market participation is much thinner and consequently the bid–offer spread in the interbank market will be wider.

2 *The time of day.* The interbank FX markets are most liquid when the major FX trading centers are open. Business hours in London and New York—the two largest FX trading centers—overlap from approximately 8:00 a.m. to 11:00 a.m. New York time. The interbank FX market for most currency pairs is typically most liquid during these hours. After London closes, liquidity is thinner through the New York afternoon. The Asian session starts when dealers in Tokyo, Singapore, and Hong Kong SAR open for business, typically by 7:00 p.m. New York time. For most currency pairs, however, the Asian session is not as liquid as the London and New York sessions. Although FX markets are open 24 hours a day on business days, between the time New York closes and the time Asia opens, liquidity in interbank markets can be very thin because

Sydney, Australia, tends to be the only active trading center during these hours. For reference, the chart below shows a 24-hour period from midnight (00:00) to midnight (24:00) London time, corresponding standard times in Tokyo and New York, and, shaded in grey, the *approximate* hours of the most liquid trading periods in each market.

Standard Time and Approximate FX Trading Hours in Major Markets: Midnight to Midnight (London Time)

Tokyo	09:00	13:00	17:00	21:00	01:00 Day+1	05:00 Day+1	09:00 Day+1
London	00:00	04:00	08:00	12:00	16:00	20:00	24:00
New York	19:00 Day−1	23:00 Day−1	03:00	07:00	11:00	15:00	19:00

3 *Market volatility.* As in any financial market, when major market participants have greater uncertainty about the factors influencing market pricing, they will attempt to reduce their risk exposures and/or charge a higher price for taking on risk. In the FX market, this response implies wider bid–offer spreads in both the interbank and broader markets. Geopolitical events (e.g., war, civil strife), market crashes, and major data releases (e.g., US non-farm payrolls) are among the factors that influence spreads and liquidity.

The size of the transaction can also affect the bid–offer spread shown by a dealer to clients. Typically, the larger the transaction, the further away from the current spot exchange rate the dealing price will be. Hence, a client who asks a dealer for a two-sided spot CAD/USD price on, for example, USD 50 million will be shown a wider bid–offer spread than a client who asks for a price on USD 1 million. The wider spread reflects the greater difficulty the dealer faces in offsetting the foreign exchange risk of the position in the interbank FX market.[4]

The relationship between the dealer and the client can also affect the size of the bid–offer spread shown by the dealer. For many clients, the spot foreign exchange business is only one business service among many that a dealer provides to that client. For example, the dealer firm might also transact in bond and/or equity securities with the same client. In a competitive business environment, in order to win the client's business for these other services, the dealer might provide a tighter (i.e., smaller) bid–offer spot exchange rate quote. The dealer might also give tighter bid–offer quotes in order to win repeat FX business. A client's credit risk can also be a factor. A client with a poor credit profile may be quoted a wider bid–offer spread than one with good credit. Given the short settlement cycle for spot FX transactions (typically two business days), however, credit risk is not the most important factor in determining the client's bid–offer spread on spot exchange rates.

4 Smaller dealing sizes can also affect the bid–offer quote shown to clients. "Retail" quotes are typically for dealing sizes smaller than 1 million units of the base currency and can range all the way down to foreign exchange transactions conducted by individuals (for example, going to a local bank in order to purchase foreign currency for an overseas holiday). Depending on the dealing venue, the bid–offer spreads for these retail transactions can be very large compared with those in the interbank market. Whereas the bid–offer spread in the interbank market for most liquid currency pairs can be a pip or two, going to the teller window at a local bank branch or using a credit card to conduct FX transactions can result in a retail bid–offer spread of several hundreds of pips.

2.1 Arbitrage Constraints on Spot Exchange Rate Quotes

The bid–offer quotes a dealer shows in the interbank FX market must respect two arbitrage constraints; otherwise the dealer creates riskless arbitrage opportunities for other interbank market participants.[5]

First, the bid shown by a dealer in the interbank market cannot be higher than the current interbank offer, and the offer shown by a dealer cannot be lower than the current interbank bid. If the bid–offer quotes shown by a dealer are inconsistent with the then-current interbank market quotes, other market participants will buy from the cheaper source and sell to the more expensive source. This arbitrage will eventually bring the two prices back into line. For example, suppose that the current spot USD/EUR price in the interbank market is 1.3649/1.3651. If a dealer showed a misaligned price quote of 1.3652/1.3654, then other market participants would pay the offer in the interbank market, *buying* EUR at a price of USD 1.3651, and then *sell* the EUR to the dealer by hitting the dealer's bid at USD 1.3652—thereby making a riskless profit of one pip on the trade. This arbitrage would continue as long as the dealer's bid–offer quote violated the arbitrage constraint.

Second, the cross-rate bids (offers) posted by a dealer must be lower (higher) than the implied cross-rate offers (bids) available in the interbank market. A currency dealer located in a given country typically provides exchange rate quotations between that country's currency and various foreign currencies. If a particular currency pair is not explicitly quoted, it can be inferred from the quotes for each currency in terms of the exchange rate with a third nation's currency. For example, given exchange rate quotes for the currency pairs A/B and C/B, we can back out the implied cross rate of A/C. This implied A/C cross rate must be consistent with the A/B and C/B rates. This again reflects the basic principle of arbitrage: If identical financial products are priced differently, then market participants will buy the cheaper one and sell the more expensive one until the price difference is eliminated. In the context of FX cross rates, there are two ways to trade currency A against currency C: (1) using the cross rate A/C or (2) using the A/B and C/B rates. Because, in the end, both methods involve selling (buying) currency C in order to buy (sell) currency A, the exchange rates for these two approaches must be consistent. If the exchange rates are not consistent, the arbitrageur will buy currency C from a dealer if it is undervalued (relative to the cross rate) and sell currency A. If currency C is overvalued by a dealer (relative to the cross rate), it will be sold and currency A will be bought.

To illustrate this **triangular arbitrage** among three currencies, suppose that the interbank market bid–offer in USD/EUR is 1.3649/1.3651 and that the bid–offer in JPY/USD is 76.64/76.66. We need to use these two interbank bid–offer quotes to calculate the market-implied bid–offer quote on the JPY/EUR cross rate.

Begin by considering the transactions required to *sell* JPY and *buy* EUR, going through the JPY/USD and USD/EUR currency pairs. We can view this process intuitively as follows:

Sell JPY Buy EUR	=	Sell JPY Buy USD	then	Sell USD Buy EUR

5 We will confine our attention to the interbank FX market because arbitrage presumes the ability to deal simultaneously with different market participants and in different markets, the ability to access "wholesale" bid–offer quotes, and the market sophistication to spot arbitrage opportunities. These criteria are typically limited to interbank market participants—that is, the professional FX market.

Note that the “Buy USD” and “Sell USD” in the expressions on the right-hand side of the equal sign will cancel out to give the JPY/EUR cross rate. In equation form, we can represent this relationship as follows:

$$\left(\frac{\text{JPY}}{\text{EUR}}\right) = \left(\frac{\text{JPY}}{\cancel{\text{USD}}}\right)\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)$$

Now, let’s incorporate the bid–offer rates in order to do the JPY/EUR calculation. A rule of thumb is that when we speak of a bid or offer exchange rate, we are referring to the bid or offer for the currency in the denominator (the base currency).

i. The left-hand side of the above equal sign is “Sell JPY, Buy EUR.” In the JPY/EUR price quote, the EUR is in the denominator (it is the base currency). Because we want to buy the currency in the denominator, we need an exchange rate that is an offer rate. Thus, we will be calculating the *offer* rate for JPY/EUR.

ii. The first term on the right-hand side of the equal sign is “Sell JPY, Buy USD.” Because we want to buy the currency in the denominator of the quote, we need an exchange rate that is an offer rate. Thus, we need the *offer* rate for JPY/USD.

iii. The second term on the right-hand side of the equal sign is “Sell USD, Buy EUR.” Because we want to buy the currency in the denominator of the quote, we need an exchange rate that is an offer rate. Thus, we need the *offer* rate for USD/EUR.

Combining all of this together conceptually and putting in the relevant offer rates leads to a JPY/EUR offer rate of

$$\left(\frac{\text{JPY}}{\text{EUR}}\right)_{\text{offer}} = \left(\frac{\text{JPY}}{\cancel{\text{USD}}}\right)_{\text{offer}}\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)_{\text{offer}} = 76.66 \times 1.3651 = 104.65$$

Perhaps not surprisingly, calculating the implied JPY/EUR *bid* rate uses the same process as above, but with “Buy JPY, Sell EUR” for the left-hand side of the equation, which leads to

$$\left(\frac{\text{JPY}}{\text{EUR}}\right)_{\text{bid}} = \left(\frac{\text{JPY}}{\cancel{\text{USD}}}\right)_{\text{bid}}\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)_{\text{bid}} = 76.64 \times 1.3649 = 104.61$$

As one would expect, the implied cross-rate bid (104.61) is less than the offer (104.65).

This simple formula seems relatively straightforward: To get the implied *bid* cross rate, simply multiply the *bid* rates for the other two currencies. However, depending on the quotes provided, it may be necessary to *invert* one of the quotes in order to complete the calculation.

This is best illustrated with an example. Consider the case of calculating the implied GBP/EUR cross rate if you are given USD/GBP and USD/EUR quotes. Simply using the provided quotes will not generate the desired GBP/EUR cross rate:

$$\frac{\text{GBP}}{\text{EUR}} \neq \left(\frac{\text{USD}}{\text{GBP}}\right)\left(\frac{\text{USD}}{\text{EUR}}\right)$$

Instead, because the USD is in the numerator in both currency pairs, we will have to invert one of the pairs to derive the GBP/EUR cross rate.

The following equation represents the cross-rate relationship we are trying to derive:

$$\frac{\text{GBP}}{\text{EUR}} = \left(\frac{\text{GBP}}{\text{USD}}\right)\left(\frac{\text{USD}}{\text{EUR}}\right)$$

But we don’t have the GBP/USD quote. We can, however, invert the USD/GBP quote and use that in our calculation. Let’s assume the bid–offer quote provided is for USD/GBP and is 1.5644/1.5646. With this quote, if we want to *buy* GBP (the currency in the denominator), we will buy GBP at the offer and the relevant quote is 1.5646. We can invert this quote to arrive at the needed GBP/USD quote: 1 ÷ 1.5646 = 0.63914. Note that, in this example, when we buy the GBP, we are also selling the USD. When we invert the provided USD/GBP offer quote, we obtain 0.63914 GBP/USD. This is the price at which we sell the USD—that is, the GBP/USD *bid*.[6]

Similarly, to get a GBP/USD *offer*, we use the inverse of the USD/GBP *bid* of 1.5644: 1 ÷ 1.5644 = 0.63922. (Note that we extended the calculated GBP/USD 0.63914/0.63922 quotes to five decimal places to avoid truncation errors in subsequent calculations.)

We can now finish the calculation of the bid and offer cross rates for GBP/EUR. Using the previously provided 1.3649/1.3651 as the bid–offer in USD/EUR, we calculate the GBP/EUR *bid* rate as follows:

$$\left(\frac{\text{GBP}}{\text{EUR}}\right)_{\text{bid}} = \left(\frac{\text{GBP}}{\cancel{\text{USD}}}\right)_{\text{bid}}\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)_{\text{bid}} = 0.63914 \times 1.3649 = 0.8724$$

Similarly, the implied GBP/EUR *offer* rate is

$$\left(\frac{\text{GBP}}{\text{EUR}}\right)_{\text{offer}} = \left(\frac{\text{GBP}}{\cancel{\text{USD}}}\right)_{\text{offer}}\left(\frac{\cancel{\text{USD}}}{\text{EUR}}\right)_{\text{offer}} = 0.63922 \times 1.3651 = 0.8726$$

Note that the implied *bid* rate is less than the implied *offer* rate, as it must be to prevent arbitrage.

We conclude this section on arbitrage constraints with some simple observations:

- The arbitrage constraint on implied cross rates is similar to that for spot rates (posted bid rates cannot be higher than the market's offer; posted offer rates cannot be lower than the market's bid). The only difference is that this second arbitrage constraint is applied *across* currency pairs instead of involving a *single* currency pair.
- In reality, any violations of these arbitrage constraints will quickly disappear. Both human traders and automatic trading algorithms are constantly on alert for any pricing inefficiencies and will arbitrage them away almost instantly. If Dealer 1 is buying a currency at a price higher than the price at which Dealer 2 is selling it, the arbitrageur will buy the currency from Dealer 2 and resell it to Dealer 1. As a result of buying and selling pressures, Dealer 2 will raise his offer prices and Dealer 1 will reduce her bid prices to the point where arbitrage profits are no longer available.
- Market participants do not need to calculate cross rates *manually* because electronic dealing machines (which are essentially just specialized computers) will automatically calculate cross bid–offer rates given any two underlying bid–offer rates.

EXAMPLE 1

Bid–Offer Rates

The following are spot rate quotes in the interbank market:

6 It may help here to remember our rule of thumb from above: When we speak of a bid or offer exchange rate, we are referring to the bid or offer for the currency in the denominator (the base currency).

USD/EUR	1.4559/1.4561
JPY/USD	81.87/81.89
CAD/USD	0.9544/0.9546
SEK/USD	6.8739/6.8741

1 What is the bid–offer on the SEK/EUR cross rate implied by the interbank market?

A 0.2118/0.2119

B 4.7209/4.7214

C 10.0077/10.0094

2 What is the bid–offer on the JPY/CAD cross rate implied by the interbank market?

A 78.13/78.17

B 85.78/85.78

C 85.76/85.80

3 If a dealer quoted a bid–offer rate of 85.73/85.75 in JPY/CAD, then a triangular arbitrage would involve buying:

A CAD in the interbank market and selling it to the dealer, for a profit of JPY 0.01 per CAD.

B JPY from the dealer and selling it in the interbank market, for a profit of CAD 0.01 per JPY.

C CAD from the dealer and selling it in the interbank market, for a profit of JPY 0.01 per CAD.

4 If a dealer quoted a bid–offer of 85.74/85.81 in JPY/CAD, then you could:

A not make any arbitrage profits.

B make arbitrage profits buying JPY from the dealer and selling it in the interbank market.

C make arbitrage profits buying CAD from the dealer and selling it in the interbank market.

5 A market participant is considering the following transactions:

Transaction 1	Buy CAD 100 million against the USD at 15:30 London time.
Transaction 2	Sell CAD 100 million against the KRW at 21:30 London time.
Transaction 3	Sell CAD 10 million against the USD at 15:30 London time.

Given the proposed transactions, what is the *most likely* ranking of the bid–offer spreads, from tightest to widest, under normal market conditions?

A Transactions 1, 2, 3

B Transactions 2, 1, 3

C Transactions 3, 1, 2

Solution to 1:

C is correct. Using the provided quotes and setting up the equations so that the cancellation of terms results in the SEK/EUR quote,

$$\frac{SEK}{EUR} = \frac{SEK}{USD} \times \frac{USD}{EUR}$$

Hence, to calculate the SEK/EUR bid (offer) rate, we multiply the SEK/USD and USD/EUR bid (offer) rates to get the following:

Bid: 10.0077 = 6.8739 × 1.4559

Offer: 10.0094 = 6.8741 × 1.4561

Solution to 2:

C is correct. Using the intuitive equation-based approach,

$$\frac{\text{JPY}}{\text{CAD}} = \frac{\text{JPY}}{\text{USD}} \times \left(\frac{\text{CAD}}{\text{USD}}\right)^{-1} = \frac{\text{JPY}}{\cancel{\text{USD}}} \times \frac{\cancel{\text{USD}}}{\text{CAD}}$$

This equation shows that we have to invert the CAD/USD quotes to get the USD/CAD bid–offer rates of 1.04756/1.04778. That is, given the CAD/USD quotes of 0.9544/0.9546, we take the inverse of each and interchange bid and offer, so that the USD/CAD quotes are (1/0.9546)/(1/0.9544), or 1.04756/1.04778. Multiplying the JPY/USD and USD/CAD bid–offer rates then leads to the following:

Bid: 85.76 = 81.87 × 1.04756

Offer: 85.80 = 81.89 × 1.04778

Solution to 3:

C is correct. The implied interbank cross rate for JPY/CAD is 85.76/85.80 (the answer to Question 2). Hence, the dealer is offering to sell the CAD (the base currency in the quote) too cheaply, at an offer rate that is below the interbank bid rate (85.75 versus 85.76, respectively). Triangular arbitrage would involve buying CAD from the dealer (paying the dealer's offer) and selling CAD in the interbank market (hitting the interbank bid), for a profit of JPY 0.01 (85.76 – 85.75) per CAD transacted.

Solution to 4:

A is correct. The arbitrage relationship is not violated: The dealer's bid (offer) is not above (below) the interbank market's offer (bid). The implied interbank cross rate for JPY/CAD is 85.76/85.80 (the solution to Question 2).

Solution to 5:

C is correct. The CAD/USD currency pair is most liquid when New York and London are both in their most liquid trading periods at the same time (approximately 8:00 a.m. to 11:00 a.m. New York time, or about 13:00 to 16:00 London time). Transaction 3 is for a smaller amount than Transaction 1. Transaction 2 is for a less liquid currency pair (KRW/CAD is traded less than CAD/USD) and occurs outside of normal dealing hours in all three major centers (London, North America, and Asia); the transaction is also for a large amount.

2.2 Forward Markets

Outright forward contracts (often referred to simply as forwards) are agreements to exchange one currency for another on a future date at an exchange rate agreed upon today. Any exchange rate transaction that has a settlement date longer than $T + 2$ is a forward contract.

Forward exchange rates must satisfy an arbitrage relationship that equates the investment return on two alternative but equivalent investments. To simplify the explanation of this arbitrage relationship, and to focus on the intuition behind forward rate calculations, we will ignore the bid–offer spread on exchange rates and money market instruments. In addition, we will alter our exchange rate notation from price/base currency (P/B) to "foreign/domestic currency" (f/d), making the assumption that the domestic currency for an investor is the base currency in the exchange rate quotation. Using this (f/d) notation will make it easier to illustrate the choice an investor faces between domestic and foreign investments, as well as the arbitrage relationships that equate the returns on these investments when their risk characteristics are equivalent.

Consider an investor with one unit of domestic currency to invest for one year. The investor faces two alternatives:

A One alternative is to invest cash for one year at the domestic risk-free rate (i_d). At the end of the year, the investment would be worth $(1 + i_d)$.

B The other alternative is to convert the domestic currency to foreign currency at the spot rate of $S_{f/d}$ and invest for one year at the foreign risk-free rate (i_f). At the end of the period, the investor would have $S_{f/d}(1 + i_f)$ units of foreign currency. These funds then must be converted back to the investor's domestic currency. If the exchange rate to be used for this end-of-year conversion is set at the start of the period using a one-year forward contract, then the investor will have eliminated the foreign exchange risk associated with converting at an unknown future spot rate. If we let $F_{f/d}$ denote the forward rate, the investor would obtain $(1/F_{f/d})$ units of the domestic currency for each unit of foreign currency sold forward. Hence, in domestic currency, at the end of the year, the investment would be worth $S_{f/d}(1 + i_f)(1/F_{f/d})$.

The two investment alternatives above (A and B) are risk free and therefore must offer the same return. If they did not offer the same return, investors could earn a riskless arbitrage profit by borrowing in one currency, lending in the other, and using the spot and forward exchange markets to eliminate currency risk. Equating the returns on these two investment alternatives—that is, putting investments A and B on opposite sides of the equal sign—leads to the following relationship:

$$\left(1 + i_d\right) = S_{f/d}\left(1 + i_f\right)\left(\frac{1}{F_{f/d}}\right)$$

To see the intuition behind forward rate calculations, note that the right-hand side of the expression (for investment B) also shows the chronological order of this investment: Convert from domestic to foreign currency at the spot rate ($S_{f/d}$); invest this foreign currency amount at the foreign risk-free interest rate $(1 + i_f)$; and then at maturity, convert the foreign currency investment proceeds back into the domestic currency using the forward rate $(1/F_{f/d})$.

For simplicity, we assumed a one-year horizon in the preceding example. However, the argument holds for any investment horizon. The risk-free assets used in this arbitrage relationship are typically bank deposits quoted using the London Interbank Offered Rate (Libor) for each currency involved. The day count convention for almost all Libor deposits is Actual/360.[7] Incorporating this day count convention into our arbitrage formula leads to

$$\left(1 + i_d\left[\frac{\text{Actual}}{360}\right]\right) = S_{f/d}\left(1 + i_f\left[\frac{\text{Actual}}{360}\right]\right)\left(\frac{1}{F_{f/d}}\right)$$

7 This notation means that interest is calculated as if there were 360 days in a year. However, the actual number of days the funds are on deposit is used to calculate the interest payable. The main exception to the Actual/360 day count convention is the GBP, for which the convention is Actual/365. For the purposes of this reading, we will use Actual/360 consistently in order to avoid complication. In practice, however, one should confirm and apply the correct day count convention for each rate. Applying incorrect day counts could give the illusion of an arbitrage opportunity where none actually exists.

This equation can be rearranged to isolate the forward rate:

$$F_{f/d} = S_{f/d}\left(\frac{1 + i_f\left[\frac{\text{Actual}}{360}\right]}{1 + i_d\left[\frac{\text{Actual}}{360}\right]}\right) \tag{1}$$

Equation 1 describes **covered interest rate parity**. Our previous work shows that covered interest rate parity is based on an arbitrage relationship among risk-free interest rates and spot and forward exchange rates. Because of this arbitrage relationship between investment alternatives, Equation 1 can also be described as saying that the covered (i.e., currency-hedged) interest rate differential between the two markets is zero.

The covered interest rate parity equation can also be rearranged to give an expression for the forward premium or discount:

$$F_{f/d} - S_{f/d} = S_{f/d}\left(\frac{\left[\frac{\text{Actual}}{360}\right]}{1 + i_d\left[\frac{\text{Actual}}{360}\right]}\right)\left(i_f - i_d\right)$$

The domestic currency will trade at a forward premium ($F_{f/d} > S_{f/d}$) if, and only if, the foreign risk-free interest rate exceeds the domestic risk-free interest rate ($i_f > i_d$). Equivalently, in this case, the foreign currency will trade at a lower rate in the forward contract (relative to the spot rate) and we would say that the foreign currency trades at a forward discount. In other words, if it is possible to earn more interest in the foreign market than in the domestic market, then the forward discount for the foreign currency will offset the higher foreign interest rate. Otherwise, covered interest rate parity would not hold and arbitrage opportunities would exist.

When the foreign currency is at a higher rate in the forward contract, relative to the spot rate, we say that the foreign currency trades at a forward premium. In the case of a forward premium for the foreign currency, the foreign risk-free interest rate will be less than the domestic risk-free interest rate. Additionally, as can be seen in the equation above, the premium or discount is proportional to the spot exchange rate ($S_{f/d}$), proportional to the interest rate differential ($i_f - i_d$) between the markets, and approximately proportional to the time to maturity (Actual/360).

Although we have illustrated the **covered interest rate parity** equation (Equation 1) in terms of foreign and domestic currencies (using the notation *f/d*), this equation can also be expressed in our more standard exchange rate quoting convention of price and base currencies (P/B):

$$F_{P/B} = S_{P/B}\left(\frac{1 + i_P\left[\frac{\text{Actual}}{360}\right]}{1 + i_B\left[\frac{\text{Actual}}{360}\right]}\right)$$

When dealing in professional FX markets, it may be more useful to think of the covered interest rate parity equation and the calculation of forward rates in this P/B notation rather than in foreign/domestic (*f/d*) notation. Domestic and foreign are relative concepts that depend on where one is located, and because of the potential for confusion, these terms are not used for currency quotes in professional FX markets.

EXAMPLE 2

Calculating the Forward Premium (Discount)

The following table shows the mid-market rate (i.e., the average of the bid and offer) for the current CAD/AUD spot exchange rate as well as for AUD and CAD 270-day Libor (annualized):

Spot (CAD/AUD)	1.0145
270-day Libor (AUD)	4.87%
270-day Libor (CAD)	1.41%

The forward premium (discount) for a 270-day forward contract for CAD/AUD would be *closest* to:

A –0.0346.

B –0.0254.

C +0.0261.

Solution:

B is correct. The equation to calculate the forward premium (discount) is as follows:

$$F_{P/B} - S_{P/B} = S_{P/B}\left(\frac{\left[\frac{\text{Actual}}{360}\right]}{1 + i_B\left[\frac{\text{Actual}}{360}\right]}\right)(i_P - i_B)$$

Because AUD is the base currency in the CAD/AUD quote, putting in the information from the table gives us

$$F_{P/B} - S_{P/B} = 1.0145\left(\frac{\left[\frac{270}{360}\right]}{1 + 0.0487\left[\frac{270}{360}\right]}\right)(0.0141 - 0.0487) = -0.0254$$

In professional FX markets, forward exchange rates are typically quoted in terms of points—the difference between the forward exchange rate quote and the spot exchange rate quote, scaled so that the points can be directly related to the last decimal place in the spot quote. Thus, the forward rate quote is typically shown as the bid–offer on the spot rate and the number of forward points at each maturity,[8] as shown in Exhibit 1.

Exhibit 1 Sample Spot and Forward Quotes (Bid–Offer)

Maturity	Spot Rate
Spot (USD/EUR)	1.3549/1.3651

(continued)

8 "Maturity" is defined in terms of the time between spot settlement—usually $T + 2$—and the settlement of the forward contract.

Exhibit 1 (Continued)

	Forward Points
One month	–5.6/–5.1
Three months	–15.9/–15.3
Six months	–37.0/–36.3
Twelve months	–94.3/–91.8

Note the following:

- As always, the offer in the bid–offer quote is larger than the bid. In this example, the forward points are negative (i.e., the forward rate for the EUR is at a discount to the spot rate) but the bid is a smaller number (–5.6 versus –5.1 at the one-month maturity).
- The absolute number of forward points is a function of the term of the forward contract: A longer contract term results in a larger number of points.
- Because this is an OTC market, a client is not restricted to dealing *only* at the dates/maturities shown. Dealers typically quote standard forward dates, but forward deals can be arranged for any forward date the client requires. The forward points for these non-standard (referred to as "broken") forward dates will typically be interpolated on the basis of the points shown for the standard settlement dates.
- The quoted points are already scaled to each maturity—they are not annualized—so there is no need to adjust them.

To convert any of these quoted forward points into a forward rate, divide the number of points by 10,000 (to scale it down to the same four decimal places in the USD/EUR spot quote) and then add the result to the spot exchange rate quote.[9] Be careful, however, about which side of the market (bid or offer) is being quoted. For example, suppose a market participant is *selling* the EUR forward against the USD and is given a USD/EUR quote. The EUR is the base currency; thus, the market participant must use the *bid* rates (i.e., hit the bid). Using the data in Exhibit 1, the three-month forward *bid* rate in this case would be based on the spot bid and the forward points bid and hence would be

$$1.3549 + \left(\frac{-15.9}{10{,}000}\right) = 1.35331$$

The market participant would be selling EUR three months forward at a price of USD 1.35331 per EUR.

2.3 The Mark-to-Market Value of a Forward Contract

Next, we consider the mark-to-market value of forward contracts. As with other financial instruments, the mark-to-market value of forward contracts reflects the profit (or loss) that would be realized from closing out the position at current market prices. To close out a forward position, we must offset it with an equal and opposite forward position using the spot exchange rate and forward points available in the market when the offsetting position is created. When a forward contract is initiated,

9 Because the JPY/USD exchange rate is quoted to only two decimal places, forward points for the dollar–yen currency pair are divided by 100.

the mark-to-market value of the contract is zero, and no cash changes hands. From that moment onward, however, the mark-to-market value of the forward contract will change as the spot exchange rate changes and as interest rates change in either of the two currencies.

Let's look at an example. Suppose that a market participant bought GBP 10 million for delivery against the AUD in six months at an "all-in" forward rate of 1.6100 AUD/GBP. (The all-in forward rate is simply the sum of the spot rate and the scaled forward points.) Three months later, the market participant wants to close out this forward contract. This would require selling GBP 10 million three months forward using the AUD/GBP spot exchange rate and forward points in effect at that time.[10] Assume the bid–offer quotes for spot and forward points three months prior to the settlement date are as follows:

Spot rate (AUD/GBP)	1.6210/1.6215
Three-month points	130/140

To sell GBP (the base currency in the AUD/GBP quote), we will be calculating the *bid* side of the market. Hence, the appropriate all-in three-month forward rate to use is

$$1.6210 + 130/10{,}000 = 1.6340.$$

This means that the market participant originally bought GBP 10 million at an AUD/GBP rate of 1.6100 and subsequently sold that amount at a rate of 1.6340. These GBP amounts will net to zero at the settlement date (GBP 10 million both bought and sold), but the AUD amounts will not, because the forward rate has changed. The AUD cash flow at the settlement date will be

$$(1.6340 - 1.6100) \times 10{,}000{,}000 = +\text{AUD } 240{,}000.$$

This is a cash *inflow* because the market participant was long the GBP with the original forward position and the GBP subsequently appreciated (the AUD/GBP rate increased).

This cash flow will be paid at the settlement day, which is still three months away. To calculate the mark-to-market value of the dealer's position, we must discount this cash flow to the present. The present value of this amount is found by discounting the settlement day cash flow by the three-month discount rate. Because this amount is in AUD, we use the three-month AUD discount rate. Let's use Libor and suppose that three-month AUD Libor is 4.80% (annualized). The present value of this future AUD cash flow is then

$$\frac{\text{AUD } 240{,}000}{1 + 0.048\left[\frac{90}{360}\right]} = \text{AUD } 237{,}154$$

This result is the mark-to-market value of the original long GBP 10 million six-month forward when it is closed out three months prior to settlement.

To summarize, the process for marking to market a forward position is relatively straightforward:

1 Create an offsetting forward position that is equal to the original forward position. (In the example above, the market participant was long GBP 10 million forward, so the offsetting forward contract would be to sell GBP 10 million.)

[10] Note that the offsetting forward contract is defined in terms of the original position taken: The original position in this example was long GBP 10 million, so the offsetting contract is short GBP 10 million. There is an ambiguity here, however: To be *long* GBP 10 million at 1.6100 AUD/GBP is equivalent to being *short* AUD 16,100,000 (10,000,000 × 1.6100) at the same forward rate. To avoid this ambiguity, for the purposes of this reading we will state what the relevant forward position is for mark-to-market purposes. The net gain or loss from the transaction will be reflected in the alternate currency.

2 Determine the appropriate all-in forward rate for this new, offsetting forward position. If the base currency of the exchange rate quote is being sold (bought), then use the bid (offer) side of the market.

3 Calculate the cash flow at the settlement day. This amount will be based on the original contract size times the difference between the original forward rate and that calculated in step 2. If the currency the market participant was originally long (short) subsequently appreciated (depreciated), then there will be a cash *inflow* (*outflow*). (In the above example, the market participant was long the GBP, which subsequently appreciated, leading to a cash inflow at the settlement day.)

4 Calculate the present value of this cash flow at the future settlement date. The currency of the cash flow and the discount rate must match. (In the example above, the cash flow at the settlement date was in AUD, so an AUD Libor was used to calculate the present value.)

The factors that affect the bid–offer spread for forward points are the same as those we discussed for spot bid–offer rates: the interbank market liquidity of the underlying currency pair, the size of the transaction, and the relationship between the client and the dealer. For forward bid–offer spreads, we can also add a fourth factor: the term of the forward contract. Generally, the longer the term of the forward contract, the wider the bid–offer spread. This relationship holds because as the term of the contract increases,

- liquidity in the forward market tends to decline;
- the exposure to counterparty credit risk increases; and
- the interest rate risk of the contract increases (forward rates are based on interest rate differentials, and a longer duration means greater price sensitivity to movements in interest rates).

EXAMPLE 3

Forward Rates and the Mark-to-Market Value of Forward Positions

A dealer is contemplating trade opportunities in the CHF/GBP currency pair. The following are the current spot rates and forward points being quoted for the CHF/GBP currency pair:

Spot rate (CHF/GBP)	1.4939/1.4941
One month	–8.3/–7.9
Two months	–17.4/–16.8
Three months	–25.4/–24.6
Four months	–35.4/–34.2
Five months	–45.9/–44.1
Six months	–56.5/–54.0

1 The current all-in bid rate for delivery of GBP against the CHF in three months is *closest* to:

A 1.49136.

B 1.49150.

C 1.49164.

2 The all-in rate that the dealer will be quoted today by another dealer to sell the CHF six months forward against the GBP is *closest* to:

A 1.48825.

B 1.48835.

C 1.48870.

Some time ago, Laurier Bay Capital, an investment fund based in Los Angeles, hedged a long exposure to the New Zealand dollar by selling NZD 10 million forward against the USD; the all-in forward price was 0.7900 (USD/NZD). Three months prior to the settlement date, Laurier Bay wants to mark this forward position to market. The bid–offer for the USD/NZD spot rate, the three-month forward points, and the three-month Libors (annualized) are as follows:

Spot rate (USD/NZD)	0.7825/0.7830
Three-month points	−12.1/−10.0
Three-month Libor (NZD)	3.31%
Three-month Libor (USD)	0.31%

3 The mark-to-market value for Laurier Bay's forward position is *closest* to:

A −USD 87,100.

B +USD 77,437.

C +USD 79,938.

Now, suppose that instead of having a long exposure to the NZD, Laurier Bay Capital had a long forward exposure to the USD, which it hedged by selling USD 10 million forward against the NZD at an all-in forward rate of 0.7900 (USD/NZD). Three months prior to settlement date, it wants to close out this short USD forward position.

4 Using the above table, the mark-to-market value for Laurier Bay's short USD forward position is *closest* to:

A −NZD 141,117.

B −NZD 139,959.

C −NZD 87,100.

Solution to 1:

A is correct. The current all-in three-month bid rate for GBP (the base currency) is equal to 1.4939 + (−25.4/10,000) = 1.49136.

Solution to 2:

C is correct. The dealer will sell CHF against the GBP, which is equivalent to buying GBP (the base currency) against the CHF. Hence, the *offer* side of the market will be used for forward points. The all-in forward price will be 1.4941 + (−54.0/10,000) = 1.48870.

Solution to 3:

C is correct. Laurier Bay sold NZD 10 million forward to the settlement date at an all-in forward rate of 0.7900 (USD/NZD). To mark this position to market, the fund would need an offsetting forward transaction involving buying NZD 10 million three months forward to the settlement date. The NZD amounts on the settlement date net to zero. For the offsetting forward contract, because the NZD is the base currency in the USD/NZD quote, buying NZD forward means paying the offer for both the spot rate and the forward points. This scenario leads to an all-in three-month forward rate of 0.7830 – 0.0010 = 0.7820. On the settlement day, Laurier Bay will receive USD 7,900,000 (NZD 10,000,000 × 0.7900 USD/NZD) from the original forward contract and pay out USD 7,820,000 (NZD 10,000,000 × 0.7820 USD/NZD) based on the offsetting forward contract. The result is a net cash flow on the settlement day of 10,000,000 × (0.7900 – 0.7820) = +USD 80,000.

This is a cash inflow because Laurier Bay sold the NZD forward and the NZD depreciated against the USD. This USD cash inflow will occur in three months. To calculate the mark-to-market value of the original forward position, we need to calculate the present value of this USD cash inflow using the three-month USD discount rate (we use USD Libor for this purpose):

$$\frac{\text{USD } 80{,}000}{1+0.0031\left[\frac{90}{360}\right]} = +\text{USD } 79{,}938$$

Solution to 4:

B is correct. Laurier Bay initially sold USD 10 million forward, and it will have to buy USD 10 million forward to the same settlement date (i.e., in three months' time) in order to close out the initial position. Buying USD using the USD/NZD currency pair is the same as selling the NZD. Because the NZD is the base currency in the USD/NZD quote, selling the NZD means calculating the *bid* rate:

$$0.7825 + (-12.1/10{,}000) = 0.78129$$

At settlement, the USD amounts will net to zero (USD 10 million both bought and sold). The NZD amounts will not net to zero, however, because the all-in forward rate changed between the time Laurier Bay initiated the original position and the time it closed out this position. At initiation, Laurier Bay contracted to sell USD 10 million and receive NZD 12,658,228 (i.e., 10,000,000/0.7900) on the settlement date. To close out the original forward contract, Laurier Bay entered into an offsetting forward contract to receive USD 10 million and pay out NZD 12,799,345 (i.e., 10,000,000/0.78129) at settlement. The difference between the NZD amounts that Laurier Bay will receive and pay out on the settlement date equals

NZD 12,658,228 – NZD 12,799,345 = –NZD 141,117.

This is a cash *outflow* for Laurier Bay because the fund was *short* the USD in the original forward position and the USD subsequently *appreciated* (i.e., the NZD subsequently depreciated, because the all-in forward rate in USD/NZD dropped from 0.7900 to 0.78129). This NZD cash outflow occurs in three months' time, and we must calculate its present value using the three-month NZD Libor:

$$\frac{-\text{NZD } 141{,}117}{1+0.0331\left(\frac{90}{360}\right)} = -\text{NZD } 139{,}959$$

3 A LONG-TERM FRAMEWORK FOR EXCHANGE RATES

Having reviewed the basic tools of the FX market, we now turn our focus to how they are used in practice. At the heart of the trading decision in FX markets lies a view on equilibrium market prices. An understanding of equilibrium pricing will assist the investor in framing decisions regarding risk exposures and how they should be managed.

In this and the following sections, we lay out a framework for developing a view on equilibrium exchange rates. We begin by examining international parity conditions, which describe the inter-relationships that jointly determine *long-run* movements in exchange rates, interest rates, and inflation. These parity conditions are the basic building blocks for describing long-term equilibrium levels for exchange rates. In subsequent sections, we expand beyond the parity conditions by discussing other factors that influence a currency's value.

Always keep in mind that exchange rate movements reflect complex interactions among multiple forces. In trying to untangle this complex web of interactions, we must clearly delineate the following concepts:

1 Long run versus short run: Many of the factors that determine exchange rate movements exert subtle but persistent influences over long periods of time. Although a poor guide for short-term prediction, longer-term equilibrium values act as an anchor for exchange rate movements.

2 Expected versus unexpected changes: In reasonably efficient markets, prices will adjust to reflect market participants' expectations of future developments. When a key factor—say, inflation—is trending gradually in a particular direction, market pricing will eventually come to reflect expectations that this trend will continue. In contrast, large, unexpected movements in a variable (for example, a central bank intervening in the foreign exchange market) can lead to immediate, discrete price adjustments. This concept is closely related to risk. For example, a moderate but steady rate of inflation will not have the same effect on market participants as an inflation rate that is very unpredictable. The latter clearly describes a riskier financial environment. Market pricing will reflect risk premia—that is, the compensation that traders and investors demand for being exposed to unpredictable outcomes. Whereas expectations of long-run equilibrium values tend to evolve slowly, risk premia—which are closely related to confidence and reputation—can change quickly in response to unexpected developments.

3 Relative movements: An exchange rate represents the relative price of one currency in terms of another. Hence, for exchange rate determination, the level or variability of key factors in any particular country is typically much less important than the *differences* in these factors across countries. For example, knowing that inflation is increasing in Country A may not give much insight into the direction of the A/B exchange rate without also knowing what is happening with the inflation rate in Country B.

As a final word of caution—and this cannot be emphasized enough: *There is no simple formula, model, or approach that will allow market participants to precisely forecast exchange rates.* Models that work well in one period may fail in others. Models that work for one set of exchange rates may fail to work for others.

Nonetheless, market participants must have a market view to guide their decisions, even if this view requires significant revision as new information becomes available. The following sections provide a framework for understanding FX markets, a guide for thinking through the complex forces driving exchange rates. As with all theory, however, it does not eliminate the need for insightful analysis of actual economic and market conditions.

3.1 International Parity Conditions

International parity conditions form the building blocks of most models of exchange rate determination. The key international parity conditions are as follows:

1 covered interest rate parity;

2 uncovered interest rate parity;
3 forward rate parity;
4 purchasing power parity; and
5 the international Fisher effect.

Parity conditions show how expected inflation differentials, interest rate differentials, forward exchange rates, current spot exchange rates, and expected future spot exchange rates would be linked in an ideal world. These conditions typically make simplifying assumptions, such as zero transaction costs, perfect information that is available to all market participants, risk neutrality, and freely adjustable market prices.

Although empirical studies find that the parity conditions rarely hold in the short term, they do help form a broadly based, long-term view of exchange rates and accompanying risk exposures. The exception to the rule that parity conditions do not hold in the short term is covered interest rate parity, which is the only parity condition that is enforced by arbitrage. We examine this parity condition first.

3.1.1 *Covered Interest Rate Parity*

We have already discussed covered interest rate parity in our examination of forward exchange rates. Under this parity condition, *an investment in a foreign money market instrument that is completely hedged against exchange rate risk should yield exactly the same return as an otherwise identical domestic money market investment*. Given the spot exchange rate and the domestic and foreign yields, the forward exchange rate must equal the rate that gives these two alternative investment strategies—invest either in a domestic money market instrument or in a fully currency-hedged foreign money market instrument—exactly the same holding period return. If one strategy gave a higher holding period return than the other, then an investor could short-sell the lower-yielding approach and invest the proceeds in the higher-yielding approach, earning riskless arbitrage profits in the process. In real-world financial markets, such a disparity will be quickly arbitraged away so that no further arbitrage profits are available. Covered interest rate parity is thus said to be a no-arbitrage condition.

For covered interest rate parity to hold exactly, it must be assumed that there are zero transaction costs and that the underlying domestic and foreign money market instruments being compared are identical in terms of liquidity, maturity, and default risk. Where capital is permitted to flow freely, spot and forward exchange markets are liquid, and financial market conditions are relatively stress free, covered interest rate differentials are generally found to be close to zero and covered interest rate parity holds.

3.1.2 *Uncovered Interest Rate Parity*

According to the **uncovered interest rate parity** condition, the *expected* return on an uncovered (i.e., unhedged) foreign currency investment should equal the return on a comparable domestic currency investment. Uncovered interest rate parity states that *the change in spot rate over the investment horizon should, on average, equal the differential in interest rates between the two countries. That is, the expected appreciation/depreciation of the exchange rate will just offset the yield differential.*

To explain the intuition behind this concept, let's switch, as we did with the examples for covered interest rate parity, from the standard price/base currency notation (P/B) to foreign/domestic currency notation (*f/d*) in order to emphasize the choice between foreign and domestic investments. As before, we also will assume that for the investor, the base currency is the domestic currency. (In *covered* interest rate parity, we assumed the investor transacted at a forward rate that was locked in at strategy initiation. In *uncovered* interest rate parity, the investor is assumed to transact at a future spot rate that is unknown at the time the strategy is initiated and the investor's currency position in the future is not hedged—that is, uncovered.)

For our example, assume that this investor has a choice between a one-year domestic money market instrument and an unhedged one-year foreign-currency-denominated money market investment. Under the assumption of uncovered interest rate parity, the investor will compare the *known* return on the domestic investment with the *expected* all-in return on the unhedged foreign-currency-denominated investment (which includes the foreign yield as well as any movements in the exchange rate, in $S_{f/d}$ terms). The choice between these two investments will depend on which market offers the higher expected return on an unhedged basis.

For example, assume that the return on the one-year foreign money market instrument is 10% while the return on the one-year domestic money market instrument is 4%. From the investor's perspective, the 4% expected return on the one-year domestic investment in domestic currency terms is known with complete certainty. This is not the case for the uncovered investment in the foreign currency money market instrument. In domestic currency terms, the investment return on an uncovered (or unhedged) foreign-currency-denominated investment is equal to

$$\left(1 + i_f\right)\left(1 - \%\Delta S_{f/d}\right) - 1$$

Intuitively, the formula says that the investor's return on a foreign investment is a function of both the foreign interest rate and the change in the spot rate, whereby a depreciation in the foreign currency reduces the investor's return. The percentage change in $S_{f/d}$ enters with a minus sign because an *increase* in $S_{f/d}$ means the foreign currency *declines* in value, thereby reducing the all-in return from the domestic currency perspective of our investor. This all-in return depends on *future* movements in the $S_{f/d}$ rate, which cannot be known until the end of the period. This return can be approximated using the following equation:[11]

$$\cong i_f - \%\Delta S_{f/d}$$

Using the previous example, consider three cases:

1 The $S_{f/d}$ rate is expected to remain unchanged.
2 The domestic currency is expected to appreciate by 10%.
3 The domestic currency is expected to appreciate by 6%.

In the first case, the investor would prefer the foreign-currency-denominated money market investment because it offers a 10% (= 10% – 0%) expected return, while the comparable domestic investment offers only 4%. In the second case, the investor would prefer the domestic investment because the expected return on the foreign-currency-denominated investment is 0% (= 10% – 10%). In the third case, uncovered interest rate parity holds because both investments offer a 4% (for the foreign investment, 10% – 6%) expected return. In this case, the risk-neutral investor is assumed to be indifferent between the alternatives.

Note that in the third case, in which uncovered interest rate parity holds, while the *expected* return over the one-year investment horizon is the same for both instruments, that expected return is *just a point on the distribution* of possible total return outcomes. The all-in return on the foreign money market instrument is uncertain because the *future* $S_{f/d}$ rate is uncertain. Hence, when we say that the investor would

11 This approximate formula holds because the product ($i \times \%\Delta S$) is small compared to the interest rate (i) and the percentage change in the exchange rate ($\%\Delta S$). For simplicity of exposition, we will use the ≅ symbol when we introduce an approximation but will subsequently treat the relationship as an equality (=) unless the distinction is important for the issue being discussed.

be indifferent between owning domestic and foreign investments because they both offer the same *expected* return (4%), we are assuming that the investor is *risk neutral*.[12] Thus, uncovered interest rate parity assumes that there are enough risk-neutral investors to force equality of expected returns.

Using our example's foreign/domestic (*f/d*) notation, uncovered interest rate parity says the expected change in the spot exchange rate over the investment horizon should be reflected in the interest rate differential:[13]

$$\%\Delta S^{e}_{f/d} = i_f - i_d \quad (2)$$

where ΔS^e indicates the change in the spot rate expected for *future* periods.

In our example, if the yield spread between the foreign and domestic investments is 6% ($i_f - i_d$ = 6%), then this spread implicitly reflects the expectation that the domestic currency will strengthen versus the foreign currency by 6%.

Uncovered interest rate parity assumes that the country with the *higher* interest rate or money market yield will see its currency *depreciate*. The depreciation of the currency offsets the initial higher yield so that the (expected) all-in return on the two investment choices is the same. Hence, if the uncovered interest rate parity condition held consistently in the real world, it would rule out the possibility of earning excess returns from going long a high-yield currency and going short a low-yield currency: The depreciation of the high-yield currency would exactly offset the yield advantage that the high-yield currency offers. Taking this scenario to its logical conclusion, if uncovered interest rate parity held at all times, investors would have no incentive to shift capital from one currency to another because expected returns on otherwise identical money market investments would be equal across markets and risk-neutral investors would be indifferent among them.

Most studies find that over short- and medium-term periods, the rate of depreciation of the high-yield currency is less than what would be implied by uncovered interest rate parity. In many cases, high-yield currencies have been found to *strengthen*, not weaken. There is, however, evidence that uncovered interest rate parity works better over very long-term horizons.

Such findings have significant implications for foreign exchange investment strategies. If high-yield currencies do not depreciate in line with the path predicted by the uncovered interest rate parity condition, then high-yield currencies should exhibit a tendency to outperform low-yield currencies over time. If so, investors could adopt strategies that overweight high-yield currencies at the expense of low-yield currencies and generate attractive returns in the process. Such approaches are known as FX carry trade strategies. We will discuss them in greater depth in Section 4.

3.1.3 *Forward Rate Parity*

Forward rate parity states that the forward exchange rate will be an unbiased predictor of the future spot exchange rate. It does not state that the forward rate will be a perfect forecast, just an unbiased one; the forward rate may overestimate or underestimate the future spot rate from time to time, but on average, it will equal the future spot rate. Forward rate parity builds upon two other parity conditions, covered interest rate parity and uncovered interest rate parity.

---

12 Risk-neutral investors base their decisions solely on the expected return and are indifferent to the investments' risk.

13 The mathematically inclined reader may note that this equation cannot hold simultaneously for $S_{f/d}$ and $S_{d/f}$ (= $1/S_{f/d}$) because their percentage changes are not of exactly equal magnitude. This reflects our earlier approximation. Using the exact return on the unhedged foreign instrument would alleviate this issue but would produce a less intuitive equation.

The covered interest rate parity condition describes the relationship among the spot exchange rate, the forward exchange rate, and interest rates. Let's keep using the foreign/domestic exchange rate notation (*f/d*) to simplify the explanation. The arbitrage condition that underlies covered interest rate parity (illustrated in Section 2) can be rearranged to give an expression for the forward premium or discount:

$$F_{f/d} - S_{f/d} = S_{f/d}\left(\frac{\left[\frac{\text{Actual}}{360}\right]}{1 + i_d\left[\frac{\text{Actual}}{360}\right]}\right)\left(i_f - i_d\right)$$

The domestic currency will trade at a forward premium ($F_{f/d} > S_{f/d}$) if, and only if, the foreign risk-free interest rate exceeds the domestic risk-free interest rate ($i_f > i_d$).

For the sake of simplicity, we assume that the investment horizon is one year, so that

$$F_{f/d} - S_{f/d} = S_{f/d}\left(\frac{i_f - i_d}{1 + i_d}\right)$$

Because the $1 + i_d$ denominator will be close to 1, we can approximate the above equation as follows:

$$F_{f/d} - S_{f/d} \cong S_{f/d}\left(i_f - i_d\right)$$

This covered interest rate parity equation can be rearranged to show the forward discount or premium as a percentage of the spot rate:

$$\frac{F_{f/d} - S_{f/d}}{S_{f/d}} \cong i_f - i_d$$

We have also shown that if uncovered interest rate parity holds, then the expected change in the spot rate is equal to the interest rate differential:

$$\%\Delta S^e_{f/d} = i_f - i_d$$

We can link the covered interest rate parity and uncovered interest rate parity equations as follows:

$$\frac{F_{f/d} - S_{f/d}}{S_{f/d}} = \%\Delta S^e_{f/d} = i_f - i_d$$

Thus, the forward premium (discount) on a currency, expressed in percentage terms, equals the expected percentage appreciation (depreciation) of the domestic currency (assuming that the uncovered interest rate parity condition holds).

In theory, then, *the forward exchange rate will be an unbiased forecast of the future spot exchange rate if both covered and uncovered interest rate parity hold:*

$$F_{f/d} = S^e_{f/d}$$

This condition is often referred to as **forward rate parity**.

We know covered interest rate parity must hold because it is enforced by arbitrage. *The question of whether forward rate parity holds is thus dependent upon whether uncovered interest rate parity holds.*

How might uncovered interest rate parity be enforced? It is not enforced by arbitrage because there is no combination of trades that will lock in a (riskless) profit. It could, however, hold if speculators willing to take risk enter the market. If the forward rate is above (below) speculators' expectations of the future spot rate, then risk-neutral speculators will buy the domestic currency in the spot (forward) market and simultaneously sell it in the forward (spot) market. These transactions would push the forward premium into alignment with the consensus expectation of the future spot rate. If the speculators' expectations are correct, they will make a profit.

Note, however, that spot exchange rates are volatile and determined by a complex web of influences: Interest rate differentials are only one among many factors. So, speculators can also lose. Because speculators are rarely, if ever, truly risk neutral, and without an arbitrage relationship to enforce it, uncovered interest rate parity is often violated. *As a result, we can conclude that forward exchange rates are typically poor predictors of future spot exchange rates in the short run*. Over the longer term, uncovered interest rate parity and forward rate parity have more empirical support.

EXAMPLE 4

Covered and Uncovered Interest Rate Parity: Predictors of Future Spot Rates

An Australia-based fixed-income asset manager is deciding how to allocate money between Australia and Japan. Note that the base currency in the exchange rate quote (AUD) is the domestic currency for the asset manager.

JPY/AUD spot rate (mid-market)	79.25
One-year forward points (mid-market)	–301.9
One-year Australian deposit rate	5.00%
One-year Japanese deposit rate	1.00%

1 Based on uncovered interest rate parity, over the next year, the expected change in the JPY/AUD rate is *closest* to a(n):

A decrease of 10%.

B decrease of 4%.

C increase of 4%.

2 The *best* explanation of why this prediction may not be very accurate is that:

A covered interest rate parity does hold in this case.

B the forward points indicate that a riskless arbitrage opportunity exists.

C there is no arbitrage condition that forces uncovered interest rate parity to hold.

3 Using the forward points to forecast the future JPY/AUD spot rate one year ahead assumes that:

A investors are risk neutral.

B spot rates follow a random walk.

C it is not necessary for uncovered interest rate parity to hold.

4 Forecasting that the JPY/AUD spot rate one year from now will equal 79.25 assumes that:

A investors are risk neutral.

B spot rates follow a random walk.

C it is necessary for uncovered interest rate parity to hold.

5 If the asset manager completely hedged the currency risk associated with a one-year Japanese deposit using a forward rate contract, the one-year all-in holding return, in AUD, would be *closest* to:

A 0%.

B 1%.

C 5%.

The fixed-income manager collects the following information and uses it, along with the international parity conditions, to estimate investment returns and future exchange rate movements.

Today's One-Year Libor		Currency Pair	Spot Rate Today
JPY	0.10%	JPY/USD	81.30
USD	0.10%	USD/GBP	1.5950
GBP	3.00%	JPY/GBP	129.67

6 If covered interest rate parity holds, the all-in one-year investment return to a Japanese investor whose currency exposure to the GBP is fully hedged is *closest* to:

A 0.10%.

B 0.17%.

C 3.00%.

7 If uncovered interest rate parity holds, today's expected value for the JPY/GBP currency pair one year from now would be *closest* to:

A 126.02.

B 129.67.

C 130.05.

8 If uncovered interest rate parity holds, between today and one year from now, the expected movement in the JPY/USD currency pair is *closest* to:

A –1.60%.

B +0.00%.

C +1.63%.

Solution to 1:

B is correct. The expected depreciation of the Australian dollar (decline in the JPY/AUD rate) is equal to the interest rate differential between Australia and Japan (5% – 1%).

Solution to 2:

C is correct. There is no arbitrage condition that forces uncovered interest rate parity to hold. In contrast, arbitrage virtually always ensures that covered interest rate parity holds. This is the case for our table, where the –302 point discount is calculated from the covered interest rate parity equation.

Solution to 3:

A is correct. Using forward rates (i.e., adding the forward points to the spot rate) to forecast future spot rates assumes that uncovered interest rate parity and forward rate parity hold. Uncovered interest rate parity assumes that investors are risk neutral. If these conditions hold, then movements in the spot exchange rate, although they *approximate* a random walk, will not actually be a random walk because current interest spreads will determine expected exchange rate movements.

Solution to 4:

B is correct. Assuming that the current spot exchange rate is the best predictor of future spot rates assumes that exchange rate movements follow a random walk. If uncovered interest rate parity holds, the current exchange rate will not be the best predictor unless the interest rate differential happens to be zero. Risk neutrality is needed to enforce uncovered interest rate parity, but it will not make the current spot exchange rate the best predictor of future spot rates.

Solution to 5:

C is correct. A fully hedged JPY investment would provide the same return as the AUD investment: 5%. This represents covered interest rate parity, an arbitrage condition.

Solution to 6:

A is correct. If covered interest rate parity holds (and it very likely does, because this is a pure arbitrage relationship), then the all-in investment return to a Japanese investor in a one-year, fully hedged GBP Libor position would be identical to a one-year JPY Libor position: 0.10%. No calculations are necessary.

Solution to 7:

A is correct. If uncovered interest rate parity holds, then forward rate parity will hold and the expected spot rate one year forward is equal to the one-year forward exchange rate. This forward rate is calculated in the usual manner, given the spot exchange rates and Libors:

$$S^e = F = 129.67\left(\frac{1.001}{1.03}\right) = 126.02$$

Solution to 8:

B is correct. Given uncovered interest rate parity, the expected change in a spot exchange rate is equal to the interest rate differential. At the one-year term, there is no difference between USD Libor and JPY Libor.

3.1.4 *Purchasing Power Parity*

So far, we have looked at the relationship between exchange rates and interest rate differentials. Now, we turn to examining the relationship between exchange rates and inflation differentials. The basis for this relationship is known as **purchasing power parity (PPP)**.

Various versions of PPP exist. The foundation for all of the versions is the law of one price. According to the **law of one price**, identical goods should trade at the same price across countries when valued in terms of a common currency. To simplify the explanation, as we did with our examples for covered and uncovered interest rate parity, let's continue to use the foreign/domestic currency quote convention (*f/d*) and the case where the base currency in the P/B notation is the domestic currency for the investor in the *f/d* notation.

The law of one price asserts that the foreign price of good x, P_f^x, should equal the exchange rate–adjusted price of the identical good in the domestic country, P_d^x:

$$P_f^x = S_{f/d} \times P_d^x$$

For example, for a EUR-based consumer, if the price of good x in the euro area is EUR 100 and the nominal exchange rate stands at 1.40 USD/EUR, then the price of good x in the United States should equal USD 140.

The **absolute version of PPP** simply extends the law of one price to the broad range of goods and services that are consumed in different countries. Expanding our example above to include all goods and services, not just good *x*, the broad price level of the foreign country (P_f) should equal the currency-adjusted broad price level in the domestic country (P_d):

$$P_f = (S_{f/d})(P_d)$$

This equation implicitly assumes that all domestic and foreign goods are tradable and that the domestic and foreign price indexes include the same bundle of goods and services with the same exact weights in each country. Rearranging this equation and solving for the nominal exchange rate ($S_{f/d}$), the absolute version of PPP states that the nominal exchange rate will be determined by the ratio of the foreign and domestic broad price indexes:

$$S_{f/d} = P_f/P_d$$

The absolute version of PPP asserts that the equilibrium exchange rate between two countries is determined entirely by the ratio of their national price levels. However, it is highly unlikely that this relationship actually holds in the real world. The absolute version of PPP assumes that goods arbitrage will equate the prices of all goods and service across countries, but if transaction costs are significant and/or not all goods and services are tradable, then goods arbitrage will be incomplete. Hence, sizable and persistent departures from absolute PPP are likely.

However, if it is assumed that transaction costs and other trade impediments are constant over time, it might be possible to show that *changes* in exchange rates and *changes* in national price levels are related, even if the relationship between exchange rate *levels* and national price *levels* does not hold. According to the **relative version of PPP**, the percentage change in the spot exchange rate ($\%\Delta S_{f/d}$) will be completely determined by the difference between the foreign and domestic inflation rates ($\pi_f - \pi_d$):[14]

$$\%\Delta S_{f/d} \cong \pi_f - \pi_d \qquad \textbf{(3)}$$

Intuitively, the relative version of PPP implies that the exchange rate changes to offset changes in competiveness arising from inflation differentials. For example, if the foreign inflation rate is assumed to be 9% while the domestic inflation rate is assumed to be 5%, then the $S_{f/d}$ exchange rate must rise by 4% ($\%\Delta S_{f/d}$ = 9% – 5% = 4%) in order to maintain the relative competitiveness of the two regions: The currency of the high-inflation country should depreciate relative to the currency of the low-inflation country. If the $S_{f/d}$ exchange rate remained unchanged, the higher foreign inflation rate would erode the competitiveness of foreign companies relative to domestic companies.

Where the relative version of PPP focuses on *actual* changes in exchange rates being driven by *actual* differences in national inflation rates, the ***ex ante*** **version of PPP** asserts that the *expected* changes in the spot exchange rate are entirely driven by *expected* differences in national inflation rates. *Ex ante* PPP tells us that countries that are expected to run *persistently* high inflation rates should expect to see their currencies depreciate over time, while countries that are expected to run relatively low inflation rates on a sustainable basis should expect to see their currencies appreciate over time. *Ex ante* PPP can be expressed as

$$\%\Delta S^e_{f/d} = \pi^e_f - \pi^e_d \qquad \textbf{(4)}$$

14 We will occasionally need to convert from a relationship expressed in levels of the relevant variables to a relationship among rates of change. If $X = (Y \times Z)$, then

$(1 + \%\Delta X) = (1 + \%\Delta Y)(1 + \%\Delta Z)$ and $\%\Delta X \approx \%\Delta Y + \%\Delta Z$ because $(\%\Delta Y \times \%\Delta Z)$ is "small."

Similarly, it can be shown that if $X = (Y/Z)$, then

$(1 + \%\Delta X) = (1 + \%\Delta Y)/(1 + \%\Delta Z)$ and $\%\Delta X \approx \%\Delta Y - \%\Delta Z$.

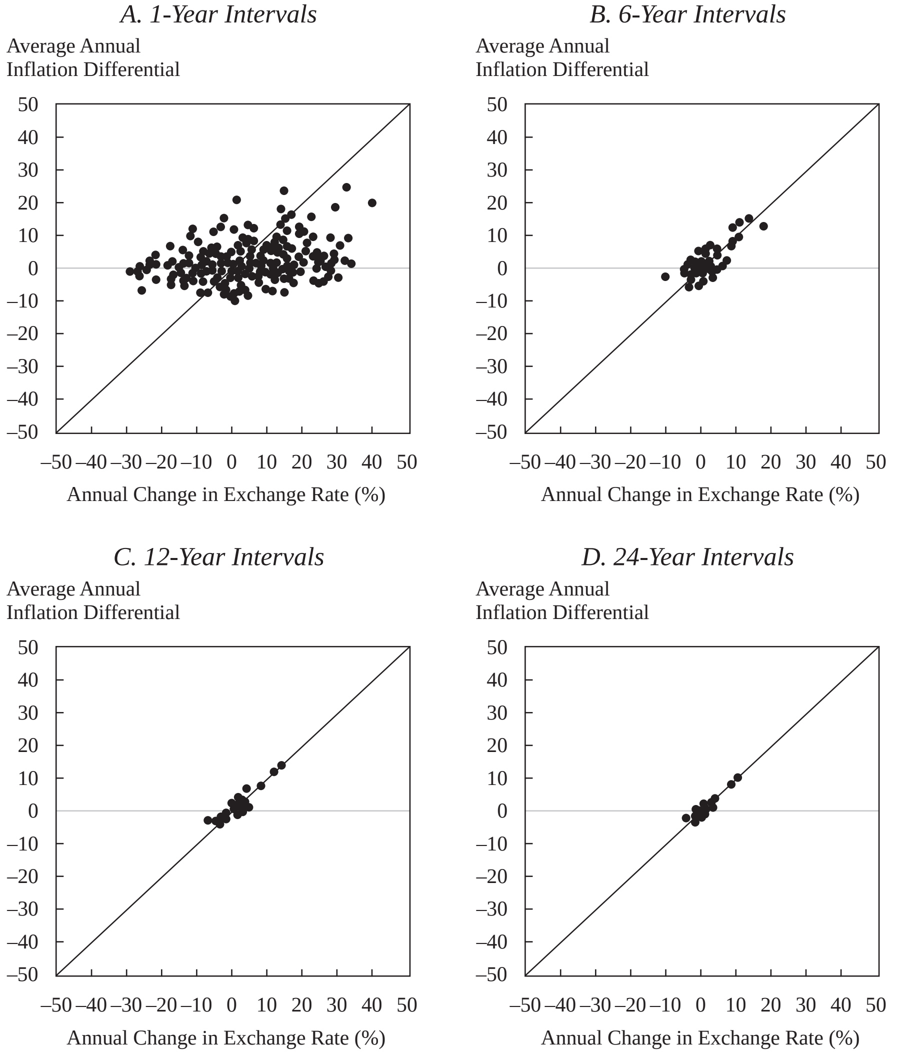
Applying this conversion to the equation for absolute PPP gives Equation 3.

where it is understood that the use of expectations (the superscript *e*) indicates that we are now focused on *future* periods. That is, $\%\Delta S^e_{f/d}$ represents the expected percentage change in the spot exchange rate, while π^e_d and π^e_f represent the expected domestic and foreign inflation rates over the same period.

Studies find that while *over shorter horizons nominal exchange rate movements may appear random, over longer time horizons nominal exchange rates tend to gravitate toward their long-run PPP equilibrium values.*

Exhibit 2 illustrates the success, or lack thereof, of the relative version of PPP at different time horizons: 1 year, 6 years, 12 years, and 24 years. Each chart plots the inflation differential (vertical axis) against the percentage change in the exchange rate (horizontal axis). If PPP holds, the points should fall along the upward-sloping diagonal line. As indicated, there is no clear relationship between changes in exchange rates and inflation differentials over a one-year time horizon. If the horizon is lengthened to six years and beyond, however, a strong positive relationship becomes apparent. Hence, *PPP appears to be a valid framework for assessing long-run fair value in the FX markets, even though the path to PPP equilibrium is excruciatingly slow*. On average, it takes roughly three to five years to narrow a given PPP deviation by roughly 50%.

Exhibit 2 Effect of Relative Inflation Rates on Exchange Rates over Different Time Horizons



Source: Isard, Faruqee, Kincaid, and Fetherston (2001).

3.1.5 *The Fisher Effect and Real Interest Rate Parity*

So far, we have examined the relationships between exchange rates and interest rate differentials and between exchange rates and inflation differentials. Now, we will begin to bring these concepts together by examining how exchange rates, interest rates, and inflation rates interact.

According to what economists call the Fisher effect, one can break down the nominal interest rate (i) in a given country into two parts: (1) the real interest rate (r) in that particular country and (2) the expected inflation rate (π^e) in that country:

$$i = r + \pi^e$$

To relate this concept to exchange rates, we can write the Fisher equation for both the domestic country and a foreign country. If the Fisher effect holds, the nominal interest rates in both countries will equal the sum of their respective real interest rates and expected inflation rates:

$$i_d = r_d + \pi_d^e$$

$$i_f = r_f + \pi_f^e$$

Let's take a closer look at the macroeconomic forces that drive the trend in nominal yield spreads. Subtracting the top equation from the bottom equation shows that the nominal yield spread between the foreign and domestic countries ($i_f - i_d$) equals the sum of two parts: (1) the foreign–domestic real yield spread ($r_f - r_d$) and (2) the foreign–domestic expected inflation differential ($\pi_f^e - \pi_d^e$):

$$i_f - i_d = \left(r_f - r_d\right) + \left(\pi_f^e - \pi_d^e\right)$$

We can rearrange this equation to solve for the *real* interest rate differential instead of the *nominal* interest rate differential:

$$\left(r_f - r_d\right) = \left(i_f - i_d\right) - \left(\pi_f^e - \pi_d^e\right)$$

To tie this material to our previous work on exchange rates, recall our expression for uncovered interest rate parity:

$$\%\Delta S_{f/d}^e = i_f - i_d$$

The nominal interest rate spread ($i_f - i_d$) equals the expected change in the exchange rate ($\%\Delta S_{f/d}^e$).

Recall also the expression for *ex ante* PPP:

$$\%\Delta S_{f/d}^e = \pi_f^e - \pi_d^e$$

The difference in expected inflation rates equals the expected change in the exchange rate. Combining these two expressions, we derive the following:

$$i_f - i_d = \pi_f^e - \pi_d^e$$

The nominal interest rate spread is equal to the difference in expected inflation rates. We can therefore conclude that if uncovered interest rate parity and *ex ante* PPP hold,

$$(r_f - r_d) = 0.$$

The real yield spread between the domestic and foreign countries ($r_f - r_d$) will be zero, and the level of real interest rates in the domestic country will be identical to the level of real interest rates in the foreign country.

The proposition that real interest rates will converge to the same level across different markets is known as the **real interest rate parity** condition.

Finally, if real interest rates are equal across markets, then it also follows that the foreign–domestic nominal yield spread is determined solely by the foreign–domestic expected inflation differential:

$$i_f - i_d = \pi_f^e - \pi_d^e$$

This is known as the **international Fisher effect**.[15]

15 The reader should be aware that some authors refer to uncovered interest rate parity as the international Fisher effect. We reserve this term for the relationship between nominal interest rate differentials and expected inflation differentials because the original (domestic) Fisher effect is a relationship between interest rates and expected inflation.

The international Fisher effect and, by extension, real interest rate parity assume that currency risk is the same throughout the world. However, not all currencies carry the same risk. For example, an emerging country may have a high level of indebtedness, which could result in an elevated level of currency risk (i.e., likelihood of currency depreciation). In this case, because the emerging market currency has higher risk, subtracting the expected inflation rate from the nominal interest rate will result in a calculated real interest rate that is higher than in other countries.[16] The emerging country's investors will require a risk premium for holding the currency, which will be reflected in nominal and real interest rates that are higher than would be expected under the international Fisher effect and real interest rate parity conditions.

EXAMPLE 5

PPP and the International Fisher Effect

An Australia-based fixed-income investment manager is deciding how to allocate her portfolio between Australia and Japan. (As before, the AUD is the domestic currency.) Australia's one-year deposit rate is 5%, considerably higher than Japan's at 1%, but the Australian dollar is estimated to be roughly 10% overvalued relative to the Japanese yen based on purchasing power parity. Before making her asset allocation, the investment manager considers the implications of interest rate differentials and PPP imbalances.

1 All else equal, which of the following events would restore the Australian dollar to its PPP value?

- **A** The Japanese inflation rate increases by 4%.
- **B** The Australian inflation rate decreases by 10%.
- **C** The JPY/AUD exchange rate declines by 10%.

2 If real interest rates in Japan and Australia were equal, then under the international Fisher effect, the inflation rate differential between Japan and Australia would be *closest* to:

- **A** 0%.
- **B** 4%.
- **C** 10%.

3 According to the theory and empirical evidence of purchasing power parity, which of the following would *not* be true if PPP holds in the long run?

- **A** An exchange rate's equilibrium path should be determined by the long-term trend in domestic price levels relative to foreign price levels.
- **B** Deviations from PPP might occur over short- and medium-term periods, but fundamental forces should eventually work to push exchange rates toward their long-term PPP path.
- **C** High-inflation countries should tend to see their currencies appreciate over time.

4 Which of the following would *best* explain the failure of the absolute version of PPP to hold?

- **A** Inflation rates vary across countries.

16 Note that economists typically separate the nominal interest rate into the real interest rate, an inflation premium, and a risk premium.

B Real interest rates are converging across countries.

C Trade barriers exist, and different product mixes are consumed across countries.

Solution to 1:

C is correct. If the Australian dollar is overvalued by 10% on a PPP basis, with all else held equal, a depreciation of the JPY/AUD rate by 10% would move the Australian dollar back to equilibrium.

Solution to 2:

B is correct. If the real interest rates were equal, then the difference in nominal yields would be explained by the difference in inflation rates (5% – 1%).

Solution to 3:

C is correct. According to PPP, high-inflation countries should see their currencies depreciate (at least, over the longer term) in order to re-equilibrate real purchasing power between countries.

Solution to 4:

C is correct. The absolute version of PPP assumes that all goods and services are tradable and that the domestic and foreign price indexes include the same bundle of goods and services with the same exact weights in each country.

3.1.6 *International Parity Conditions: Tying All the Pieces Together*

As noted above, the various parity relationships usually do not hold over short time horizons. However, studies show that over longer time periods, there is a discernible interaction among nominal interest rates, exchange rates, and inflation rates across countries, such that the international parity conditions serve as an anchor for longer-term exchange rate movements. We now summarize the key international parity conditions and describe how they are all linked.

1. According to covered interest rate parity, arbitrage ensures that nominal interest rate spreads equal the percentage forward premium (or discount).
2. According to uncovered interest rate parity, the expected percentage change of the spot exchange rate should, on average, be reflected in the nominal interest rate spread.
3. If both covered and uncovered interest rate parity hold—that is, the nominal yield spread equals both the forward premium (or discount) and the expected percentage change in the spot exchange rate—then the forward exchange rate will be an unbiased predictor of the future spot exchange rate.
4. According to the *ex ante* PPP approach to exchange rate determination, the expected change in the spot exchange rate should equal the expected difference between domestic and foreign inflation rates.
5. Assuming the Fisher effect holds in all markets—that is, the nominal interest rate in each market equals the real interest rate plus the expected inflation rate—and also assuming that real interest rates are broadly the same across

all markets (real interest rate parity), then the nominal yield spread between domestic and foreign markets will equal the domestic–foreign expected inflation differential, which is the international Fisher effect.

6 If *ex ante* PPP and the International Fisher effect hold, then expected inflation differentials should equal both the expected change in the exchange rate and the nominal interest rate differential. This relationship implies that the expected change in the exchange rate equals the nominal interest rate differential, which is uncovered interest rate parity.

In sum, if all the key international parity conditions held at all times, then the expected percentage change in the *spot* exchange rate would equal

- the forward premium or discount (expressed in percentage terms);
- the nominal yield spread between countries; and
- the difference between expected national inflation rates.

In other words, *if all these parity conditions held, it would be impossible for a global investor to earn consistent profits on currency movements.* If forward exchange rates accurately predicted the future path of spot exchange rates, there would be no way to make money in forward exchange speculation. If high-yield currencies fell in value versus low-yield currencies exactly in line with the path implied by nominal interest rate spreads, all markets would offer the same currency-adjusted total returns over time. Investors would have no incentive to shift funds from one market to another based solely on currency considerations.

EXAMPLE 6

The Relationships among the International Parity Conditions

1 Which of the following is a no-arbitrage condition?

A Real interest rate parity

B Covered interest rate parity

C Uncovered interest rate parity

2 Forward rates are unbiased predictors of future spot rates if two parity conditions hold. Which of the following is *not* one of these conditions?

A Real interest rate parity

B Covered interest rate parity

C Uncovered interest rate parity

3 The international Fisher effect requires all but which of the following to hold?

A *Ex ante* PPP

B Absolute PPP

C Real interest rate parity

4 The forward premium/discount is determined by nominal interest rate differentials because of:

A the Fisher effect.

B covered interest parity.

C real interest rate parity.

5 If all of the key international parity conditions held at all times, then the expected percentage change in the spot exchange rate would equal all *except* which of the following?

A The real yield spread

B The nominal yield spread

C The expected inflation spread

Solution to 1:

B is correct. Covered interest rate parity is enforced by equating the investment return on two riskless investments (domestic and currency-hedged foreign).

Solution to 2:

A is correct. Both covered and uncovered interest rate parity must hold for the forward rate to be an unbiased predictor of the future spot rate. Real interest rate parity is not required.

Solution to 3:

B is correct. The international Fisher effect is based on real interest rate parity and *ex ante* PPP (not absolute PPP).

Solution to 4:

B is correct. The forward premium/discount is determined by covered interest rate arbitrage.

Solution to 5:

A is correct. If all the international parity conditions held, the real yield spread would equal zero, regardless of expected changes in the spot exchange rate.

4 THE CARRY TRADE

According to uncovered interest rate parity, high-yield currencies are expected to depreciate in value, while low-yield currencies are expected to appreciate in value. If uncovered interest rate parity held at all times, investors would not be able to profit from a strategy that undertook long positions in high-yield currencies and short positions in low-yield currencies. The change in spot rates over the tenor of the forward contracts would cancel out the interest rate differentials locked in at the inception of the position.

Uncovered interest rate parity is one of the most widely tested propositions in international finance. The evidence suggests that uncovered interest rate parity does *not* hold over short and medium time periods. Studies have generally found that *high-yield currencies, on average, have not depreciated, and low-yield currencies have not appreciated, to the levels predicted by interest rate differentials.*

These findings underscore the potential profitability of a trading strategy known as the **FX carry trade**, which involves taking long positions in high-yield currencies and short positions in low-yield currencies. The latter are often referred to as "funding currencies." As a simplified example of the carry trade, assume a trader can borrow Canadian dollars at 1% and earn 9% on an investment in Brazilian reals for one year. To execute the trade to earn 8% from the interest rate differential, the trader will do the following:

1 Borrow Canadian dollars at $t = 0$

2 Sell the dollars and buy Brazilian reals at the spot rate at $t = 0$
3 Invest in a real-denominated investment at $t = 0$
4 Liquidate the Brazilian investment at $t = 1$
5 Sell the reals and buy dollars at the spot rate at $t = 1$
6 Pay back the dollar loan

If the real appreciates, the trader's profits will be greater than 8% because the stronger real will buy more dollars in one year. If the real depreciates, the trader's profits will be less than 8% because the weaker real will buy fewer dollars in the future. If the real falls in value by more than 8%, the trader will experience losses.[17]

Historical evidence shows that carry trade strategies have generated positive returns over extended periods.[18] One argument for the persistence of the carry trade is that the yields in higher interest rate countries reflect a risk premium due to a more unstable economy, while low-yield currencies represent less risky markets. Although small increases in financial market and/or FX volatility are unlikely to materially affect carry strategy profits, elevated levels of volatility and/or perceived risk in the financial markets can quickly turn these profits into substantial losses. That is, during turbulent periods, the returns on long high-yield currency positions will tend to decline dramatically, while the losses on short low-yield currency positions will tend to rise dramatically.

To understand why, we need to understand the nature of the risk and reward in the carry trade. The reward is the gradual accrual of the interest rate differential—income that is unrelated to exchange rate volatility. The risk arises from the potential for sudden adverse exchange rate movements that result in instantaneous capital losses. During periods of low turbulence, investors may feel relatively confident that exchange rate movements will not jeopardize the gradual accrual of the interest rate differential. Because low-volatility regimes have tended to be the norm and often last for extended periods, investors can become complacent, taking on larger carry trade positions in a search for yield but increasing their risk exposures. When volatility in the currency markets spikes, however, the risk of an adverse exchange rate movement rises sharply relative to the gradual flow of income. As the trade moves toward unprofitability, investors may rush to unwind the carry trade, selling high-yielding currencies and re-purchasing low-yielding currencies. These carry trades are often large-scale trades initiated by trading firms and other opportunistic investors, such as hedge funds. Traders often have stop-loss orders in place that are triggered when price declines reach a certain level. When they all attempt to unwind the trades at once, the selling pressure adds to the losses on the long position currency and the buying pressure on the short position currency drives that currency higher, exacerbating the loss. The "flight to quality" during turbulent times and the leverage inherent in the carry trade further compound the losses. The upshot is that *during periods of low volatility, carry trades tend to generate positive returns, but they are prone to significant crash risk in turbulent times.*

---

17 The carry trader's return consists of the intermarket yield spread, the currency appreciation/depreciation, and the foreign investment appreciation/depreciation. Typically, a carry trade is executed using an investment in highly rated government debt so as to mitigate credit risk. In this simplified example, we use an additive approach to determine the trader's returns (i.e., we ignore the currency gain or loss on the 8% interest rate differential).

18 For example, Dimson, Marsh, McGinnie, Staunton, and Wilmot (2012) found that for the period 1900–2011, carry returns averaged 1.1% annually. For their most recent sub-period, 1972–2011, returns were higher at 2.3% annually. Interestingly, returns were sometimes higher when the strategy was based on real interest rate differentials. For the entire period and their most recent sub-period, annual returns were 2.3% and 1.8%, respectively. See the end of this section for a discussion of the use of PPP in a carry trade strategy.

The tendency for carry trades to experience periodic crashes results in a non-normal distribution of returns for both developed and emerging market (EM) carry trades. Relative to a normal distribution, the distributions tend to be more peaked, with fatter tails and negative skewness. The more peaked distribution around the mean implies that carry trades have typically generated a larger number of trades with small gains/losses than would occur with the normal distribution. Although carry trades have generated positive returns on average in the past, the negative skew and fat tails indicate that carry trades have tended to have more frequent and larger losses than would have been experienced had the return distribution been normal.

EXAMPLE 7

Carry Trade Strategies

A currency fund manager is considering allocating a portion of her FX portfolio to carry trade strategies. The fund's investment committee asks the manager a number of questions about why she has chosen to become involved in FX carry trades and how she will manage the risk of potentially large downside moves associated with the unwinding of carry trades. Which of the following would be her *best* responses to the investment committee's questions?

1 Carry trades can be profitable when:
   - A covered interest rate parity does not hold.
   - B uncovered interest rate parity does not hold.
   - C the international Fisher effect does not hold.

2 Over time, the return distribution of the fund's FX carry trades is *most* likely to resemble a:
   - A normal distribution with fat tails.
   - B distribution with fat tails and a negative skew.
   - C distribution with thin tails and a positive skew.

3 The volatility of the fund's returns relative to its equity base is *best* explained by:
   - A leverage.
   - B low deposit rates in the funding currency.
   - C the yield spread between the high- and low-yielding currencies.

4 A Tokyo-based asset manager enters into a carry trade position based on borrowing in yen and investing in one-year Australian Libor.

Today's One-Year Libor		Currency Pair	Spot Rate Today	Spot Rate One Year Later
JPY	0.10%	JPY/USD	81.30	80.00
AUD	4.50%	USD/AUD	1.0750	1.0803

After one year, the all-in return to this trade, measured in JPY terms, would be *closest* to:

   - A +1.84%.
   - B +3.23%.
   - C +5.02%.

Solution to 1:

B is correct. The carry trade is based on the supposition that uncovered interest rate parity does not hold.

Solution to 2:

B is correct. The "crash risk" of carry trades implies a fat-tailed distribution skewed toward a higher probability of large losses (compared with a normal distribution).

Solution to 3:

A is correct. Carry trades are leveraged trades (borrow in the funding currency, invest in the high-yield currency), and leverage increases the volatility in the investor's return on equity.

Solution to 4:

B is correct. To calculate the all-in return for a Japanese investor in a one-year AUD Libor deposit, we must first calculate the current and one-year-later JPY/AUD cross rates. Because USD 1.0000 buys JPY 81.30 today and AUD 1.0000 buys USD 1.0750 today, today's JPY/AUD cross rate is the product of these two numbers: 81.30 × 1.0750 = 87.40 (rounded to two decimal places). Similarly, one year later, the observed cross rate is 80.00 × 1.0803 = 86.42 (rounded to two decimal places).

Accordingly, measured in yen, the investment return for the unhedged Australian Libor deposit is *closest* to:

$$\frac{1}{87.40}(1 + 4.50\%)86.42 - 1 = 0.0333$$

Against this 3.33% *gross* return, however, the manager must charge the borrowing costs to fund the carry trade investment (one-year JPY Libor was 0.10%). Hence, the *net* return on the carry trade is closest to 3.33% – 0.10% = 3.23%.

We can also calculate the profit using a transactional approach. Assuming an initial position of, for example, 100 yen (JPY 100), the investor will obtain JPY 100 × 1/JPY 87.40 = AUD 1.1442. After one year, the investment will be worth AUD 1.1442 × 1.045 = AUD 1.1957. Converting back to yen in one year results in AUD 1.1957 × JPY 86.42/AUD = JPY 103.33. Paying off the yen loan results in a profit of JPY 103.33 – (JPY 100 × 1.001) = JPY 3.23. This is the 3.23% profit calculated previously.

5 THE IMPACT OF BALANCE OF PAYMENTS FLOWS

As noted earlier, the parity conditions may be appropriate for assessing fair value for currencies over long horizons, but they are of little use as a real-time gauge of value. There have been many attempts to find a better framework for determining a currency's short- or long-run equilibrium value. In this section, we examine the influence of trade and capital flows.

A country's balance of payments consists of its current account as well as its capital and financial account.[19] Loosely speaking, the current account reflects flows in the real economy, which refers to that part of the economy engaged in the actual production of goods and services (as opposed to the financial sector). The capital account reflects financial flows. Decisions about trade flows (the current account) and investment/financing flows (the capital account) are typically made by different entities with different perspectives and motivations. Their decisions are brought into alignment by changes in market prices and/or quantities. One of the key prices—perhaps *the* key price—in this process is the exchange rate.

Countries that import more than they export will have a negative current account balance and are said to have current account deficits. Those with more exports than imports will have a current account surplus. A country's current account balance must be matched by an equal and opposite balance in the capital account. Thus, countries with current account deficits must attract funds from abroad in order to pay for the imports (i.e., they must have a capital account surplus).

When discussing the effect of the balance of payments components on a country's exchange rate, one must distinguish between short- and intermediate-term influences on the one hand and longer-term influences on the other. Over the long term, countries that run persistent current account deficits (net borrowers) often see their currencies depreciate because they finance their acquisition of imports through the continued use of debt. Similarly, countries that run persistent current account surpluses (net lenders) often see their currencies appreciate over time.

However, investment/financing decisions are usually the dominant factor in determining exchange rate movements, at least in the short to intermediate term. There are four main reasons for this:

- Prices of real goods and services tend to adjust much more slowly than exchange rates and other asset prices.
- Production of real goods and services takes time, and demand decisions are subject to substantial inertia. In contrast, liquid financial markets allow virtually instantaneous redirection of financial flows.
- Current spending/production decisions reflect only purchases/sales of current production, while investment/financing decisions reflect not only the financing of current expenditures but also the reallocation of existing portfolios.
- *Expected* exchange rate movements can induce very large short-term capital flows. This tends to make the *actual* exchange rate very sensitive to the currency views held by owners/managers of liquid assets.

In this section, we first examine the impact of current account imbalances on exchange rates. Then, we take a closer look at capital flows.

5.1 Current Account Imbalances and the Determination of Exchange Rates

Current account trends influence the path of exchange rates over time through several mechanisms:

- The flow supply/demand channel

19 The official balance of payments accounts make a distinction between the "capital account" and the "financial account" based on the nature of the assets involved. For simplicity, we will use the term "capital account" here to reflect all investment/financing flows.

- The portfolio balance channel
- The debt sustainability channel

We briefly discuss each of these mechanisms next.

5.1.1 *The Flow Supply/Demand Channel*

The flow supply/demand channel is based on a fairly simple model that focuses on the fact that purchases and sales of internationally traded goods and services require the exchange of domestic and foreign currencies in order to arrange payment for those goods and services. For example, if a country sold more goods and services than it purchased (i.e., the country was running a current account surplus), then the demand for its currency should rise, and vice versa. Such shifts in currency demand should exert upward pressure on the value of the surplus nation's currency and downward pressure on the value of the deficit nation's currency.

Hence, countries with persistent current account surpluses should see their currencies appreciate over time, and countries with persistent current account deficits should see their currencies depreciate over time. A logical question, then, would be whether such trends can go on indefinitely. At some point, domestic currency strength should contribute to deterioration in the trade competitiveness of the surplus nation, while domestic currency weakness should contribute to an improvement in the trade competitiveness of the deficit nation. Thus, the exchange rate responses to these surpluses and deficits should eventually help eliminate—in the medium to long run—the source of the initial imbalances.

The amount by which exchange rates must adjust to restore current accounts to balanced positions depends on a number of factors:

- The initial gap between imports and exports
- The response of import and export prices to changes in the exchange rate
- The response of import and export demand to changes in import and export prices

If a country imports significantly more than it exports, export growth would need to far outstrip import growth in percentage terms in order to narrow the current account deficit. A large initial deficit may require a substantial depreciation of the currency to bring about a meaningful correction of the trade imbalance.

A depreciation of a deficit country's currency should result in an increase in import prices in domestic currency terms and a decrease in export prices in foreign currency terms. However, empirical studies often find limited pass-through effects of exchange rate changes on traded goods prices. For example, many studies find that for every 1% decline in a currency's value, import prices rise by only 0.5%—and in some cases by even less—because foreign producers tend to lower their profit margins in an effort to preserve market share. In light of the limited pass-through of exchange rate changes into traded goods prices, the exchange rate adjustment required to narrow a trade imbalance may be far larger than would otherwise be the case.

Many studies find that the response of import and export demand to changes in traded goods prices is often quite sluggish, and as a result, relatively long lags, lasting several years, can occur between (1) the onset of exchange rate changes, (2) the ultimate adjustment in traded goods prices, and (3) the eventual impact of those price changes on import demand, export demand, and the underlying current account imbalance.

5.1.2 *The Portfolio Balance Channel*

The second mechanism through which current account trends influence exchange rates is the so-called portfolio balance channel. Current account imbalances shift financial wealth from deficit nations to surplus nations. Countries with trade deficits will finance their trade with increased borrowing. This behavior may lead to shifts in global asset preferences, which in turn could influence the path of exchange rates. For example, nations running large current account surpluses versus the United States might find that their holdings of US dollar–denominated assets exceed the amount they desire to hold in a portfolio context. Actions they might take to reduce their dollar holdings to desired levels could then have a profound, negative impact on the dollar's value.

5.1.3 *The Debt Sustainability Channel*

The third mechanism through which current account imbalances can affect exchange rates is the so-called debt sustainability channel. According to this mechanism, there should be some upper limit on the ability of countries to run persistently large current account deficits. If a country runs a large and persistent current account deficit over time, eventually it will experience an untenable rise in debt owed to foreign investors. If such investors believe that the deficit country's external debt is rising to unsustainable levels, they are likely to reason that a major depreciation of the deficit country's currency will be required at some point to ensure that the current account deficit narrows significantly and that the external debt stabilizes at a level deemed sustainable.

The existence of persistent current account imbalances will tend to alter the market's notion of what exchange rate level represents the true, long-run equilibrium value. For deficit nations, ever-rising net external debt levels as a percentage of GDP should give rise to steady (but not necessarily smooth) downward revisions in market expectations of the currency's long-run equilibrium value. For surplus countries, ever-rising net external asset levels as a percentage of GDP should give rise to steady upward revisions of the currency's long-run equilibrium value. Hence, one would expect currency values to move broadly in line with trends in debt and/or asset accumulation.

Persistent Current Account Deficits: The US Current Account and the US Dollar

The historical record indicates that the trend in the US current account has been an important determinant of the long-term swings in the US dollar's value but also that there can be rather long lags between the onset of a deterioration in the current account balance and an eventual decline in the dollar's value. For example, the US current account balance deteriorated sharply in the first half of the 1980s, yet the dollar soared over that period. The reason for the dollar's strength over that period was that high US real interest rates attracted large inflows of capital from abroad, which pushed the dollar higher despite the large US external imbalance. Eventually, however, concerns regarding the sustainability of the ever-widening US current account deficit triggered a major dollar decline in the second half of the 1980s.

History repeated itself in the second half of the 1990s, with the US current account balance once again deteriorating while the dollar soared over the same period. This time, the dollar's strength was driven by strong foreign direct investment, as well as both debt- and equity-related flows into the United States. Beginning in 2001, however, the ever-widening US current account deficit, coupled with a decline in US interest rates, made it more difficult for the United States to attract the foreign private capital needed to finance its current account deficit. The dollar eventually succumbed to the weight of ever-larger trade and current account deficits and began a multi-year slide, starting in 2002–2003. Interestingly, the US dollar has undergone three major downward cycles since the advent of floating exchange rates: 1977–1978, 1985–1987, and 2002–2008. In each of those downward cycles, the dollar's slide was driven in large part by concerns over outsized US current account deficits coupled with relatively low nominal and/or real short-term US interest rates, which made it difficult to attract sufficient foreign capital to the United States to finance those deficits.

Exchange Rate Adjustment in Surplus Nations: Japan and China

Japan and, more recently, China represent examples of countries with large current account surpluses and illustrate the pressure that those surpluses can bring to bear on currencies. In the case of Japan, its rising current account surplus has exerted persistent upward pressure on the yen's value versus the dollar over time. Part of this upward pressure simply reflected the increase in demand for yen to pay for Japan's merchandise exports. But some of the upward pressure on the yen might also have stemmed from rising commercial tensions between the United States and Japan.

Protectionist sentiment in the United States rose steadily with the rising bilateral trade deficit that the United States ran with Japan in the postwar period. US policymakers contended that the yen was undervalued and needed to appreciate. With the increasing trade imbalance between the two countries contributing to more heated protectionist rhetoric, Japan felt compelled to tolerate steady upward pressure on the yen. As a result, the yen's value versus the dollar has tended to move in sync with the trend in Japan's current account surplus.

5.2 Capital Flows and the Determination of Exchange Rates

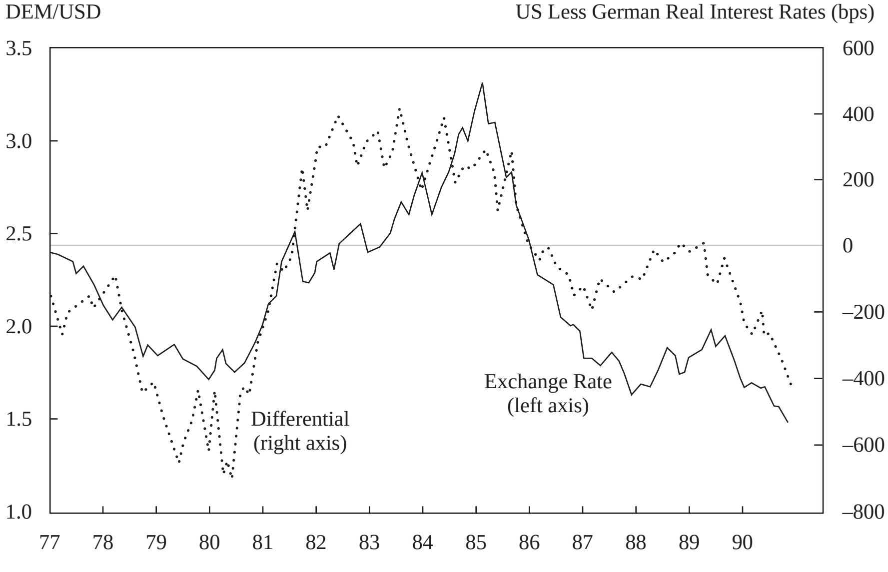
Greater financial integration of the world's capital markets and greater freedom of capital to flow across national borders have increased the importance of global financial flows in determining exchange rates, interest rates, and broad asset price trends. One can cite many examples in which global financial flows either caused or contributed to extremes in exchange rates, interest rates, or asset prices. Two specific examples are given below:

- Yen assets underperformed US dollar assets over much of the 1995–2007 period, a 12-year span of ultra-low Japanese short-term interest rates that gave rise to what became known as the "yen carry trade" as both Japanese and global fund managers borrowed in yen and invested the proceeds in higher-yielding assets in other markets. Such actions helped push the value of the yen significantly lower over time. Periodically, however, such positions became overextended and vulnerable to sudden reversals. In the fall of 1998, a major unwinding of the yen carry trade led to the collapse of several major hedge funds.
- In the first half of the 1980s, a major widening in yield spreads favoring the United States attracted significant amounts of foreign capital to that country, fueling a rise in the dollar. Again in the second half of the 1990s, increasing capital flows to the United States fueled by global demand for US financial assets—specifically US equities during the tech boom—spurred another dramatic rise in the US dollar.

Exhibits 3 and 4 illustrate the pivotal role that real interest rate differentials played in driving the US dollar's value during these periods. As shown in Exhibit 3, the decline of the dollar in the late 1970s, the dramatic rise in its value in the first half of the 1980s, and its subsequent decline in the second half of the 1980s can be explained, to a large extent, by changes in US–foreign real yield spreads. (In Exhibits 3 and 4, DEM/USD indicates the number of German Deutsche Marks per US dollar.) Exhibit 4 shows that the dollar's decline in the first half of the 1990s coincided with a significant narrowing in US–foreign real yield spreads, while the dollar's subsequent rise in the second half of the 1990s coincided with a significant widening in US–foreign real yield spreads.

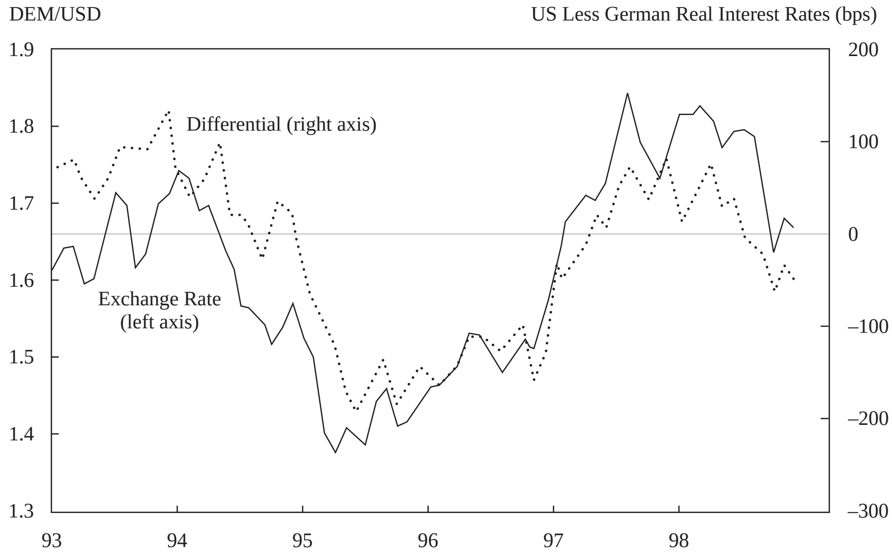
In numerous cases, global capital flows have helped fuel boom-like conditions in emerging market economies for a while before, suddenly and often without adequate warning, those flows reversed. The reversals often caused a major economic downturn, sovereign default, a serious banking crisis, and/or significant currency depreciation. Excessive emerging market capital inflows often plant the seeds of a crisis by contributing to (1) an unwarranted appreciation of the emerging market currency, (2) a huge buildup in external indebtedness, (3) an asset bubble, (4) a consumption binge that contributes to explosive growth in domestic credit and/or the current account deficit, or (5) an overinvestment in risky projects and questionable activities. Governments in emerging markets often resist currency appreciation from excessive capital inflows by using capital controls or selling their currency in the FX market. As an example of the former, in 2016 the Brazilian government was using a tax on foreign exchange transactions to control capital flows and raise government revenue. Note, however, that in general, government control of the exchange rate will not be completely effective because even if a government prohibits investment capital flows, some capital flows will be needed for international trade. In addition, the existence or emergence of black markets for the country's currency will inhibit the ability of the government to fully control the exchange rates for its own currency.

Exhibit 3 Deutsche Mark/US Dollar Exchange Rate and US/German Real Interest Rate Differentials (10-Year Bond Yields Less CPI), 1977–1990



Source: Datastream.

Exhibit 4 Deutsche Mark/US Dollar Exchange Rate and US/German Real Interest Rate Differentials (10-Year Bond Yields Less CPI), 1993–1998



Source: Datastream.

Sometimes, capital flows due to interest rate spreads have little impact on the trend in exchange rates. Consider the case of the Turkish lira. The lira attracted a lot of interest on the part of global fund managers over the 2002–2010 period, in large part because of its attractive yields. Turkish–US short-term yield spreads averaged over 1,000 bps during much of this period. As capital flowed into Turkey, the Turkish authorities intervened in the foreign exchange market in an attempt to keep the lira from appreciating. As a result, international investors were not able to reap the anticipated currency gains over this period. While the return from the movement in the spot exchange rate was fairly small, a long Turkish lira/short US dollar carry trade position generated significant long-run returns, mostly from the accumulated yield spread.

One-sided capital flows can persist for long periods. Consider the case of a high-yield, inflation-prone emerging market country that wants to promote price stability and long-term sustainable growth. To achieve price stability, policymakers in the high-yield economy will initiate a tightening in monetary policy by gradually raising the level of domestic interest rates relative to yield levels in the rest of the world. If the tightening in domestic monetary policy is sustained, inflation expectations for the high-yield economy relative to other economies should gradually decline. The combination of sustained wide nominal yield spreads and a steady narrowing in relative inflation expectations should exert upward pressure on the high-yield currency's value, resulting in carry trade profits over long periods.

Policymakers in high-yield markets can also pursue policies which attract foreign investment; such policies might include tighter fiscal policies, liberalization of financial markets, fewer capital flow restrictions, privatization, and/or a better business environment. Such policies should encourage investors to gradually require a lower risk premium to hold the high-yield currency's assets and revise upward their assessment of the long-run equilibrium value of that country's currency.

The historical evidence suggests that the impact of nominal interest rate spreads on the exchange rate tends to be gradual. Monetary policymakers tend to adjust their official lending rates slowly over time—in part because of the uncertainty that policymakers face and in part because the authorities do not want to disrupt the financial markets. This very gradual change in rates implies a very gradual narrowing of the spread between high-yield and low-yield countries. Similarly, the downward trends in inflation expectations and risk premia in the higher-yield market also tend to unfold gradually. It often takes several years to determine whether structural economic changes will take root and boost the long-run competitiveness of the higher-yield country. Because these fundamental drivers tend to reinforce each other over time, there may be persistence in capital flows and carry trade returns.

5.2.1 *Equity Market Trends and Exchange Rates*

Increasing equity prices can also attract foreign capital. Although exchange rates and equity market returns sometimes exhibit positive correlation, the relationship between equity market performance and exchange rates is not stable. The long-run correlation between the US equity market and the dollar, for example, is very close to zero, but over short to medium periods, correlations tend to swing from being highly positive to being highly negative, depending on market conditions. For instance, between 1990 and 1995, the US dollar fell while the US equity market was strong and the Japanese yen soared while Japanese stocks were weak. In contrast, between 1995 and early 2000, the US dollar soared in tandem with a rising US equity market while the yen weakened in tandem with a decline in the Japanese equity market. *Such instability in the correlation between exchange rates and equity markets makes it difficult to form judgments on possible future currency moves based solely on expected equity market performance.*

Since the global financial crisis, there has been a decidedly negative correlation between the US dollar and the US equity market. Market observers attribute this behavior of the US dollar to its role as a safe haven asset. When investors' appetite for risk is high—that is, when the market is in "risk-on" mode—investor demand for risky assets, such as equities, tends to rise, which drives up their prices. At the same time, investor demand for safe haven assets, such as the dollar, tends to decline, which drives their values lower. The opposite has occurred when the market has been in "risk-off" mode.

EXAMPLE 8

Capital Flows and Exchange Rates

Monique Kwan, a currency strategist at a major foreign exchange dealer, is responsible for formulating trading strategies for the currencies of both developed market (DM) and emerging market (EM) countries. She examines two countries—one DM and one EM—and notes that the DM country has what is considered a low-yield safe haven currency while the EM country has a high-yield currency whose value is more exposed to fluctuations in the global economic growth rate. Kwan is trying to form an opinion about movements in the exchange rate for the EM currency.

1 All else equal, the exchange rate for the EM currency will *most likely* depreciate if the:

A long-run equilibrium value of the high-yield currency is revised upward.

B nominal yield spread between the EM and DM countries increases over time.

C expected inflation differential between the EM and DM countries is revised upward.

2 An increase in safe haven demand would *most likely*:

A increase the risk premium demanded by international investors to hold assets denominated in the EM currency.

B raise the return earned on carry trade strategies.

C exert upward pressure on the value of the EM currency.

Kwan notes that the DM country is running a persistent current account deficit with the EM country. To isolate the influence of this chronic imbalance on exchange rates, she focuses only on the bilateral relationship between the EM and DM countries and makes the simplifying assumption that the external accounts of these two countries are otherwise balanced (i.e., there are no other current account deficits).

3 Over time, and all else equal, the persistent current account deficit with the EM country would *most likely* lead to:

A a large buildup of the EM country's assets held by the DM country.

B an increase in the trade competitiveness of the EM country.

C an upward revision in the long-run equilibrium EM currency value.

Kwan notes that because of the high yield on the EM country's bonds, international investors have recently been reallocating their portfolios more heavily toward this country's assets. As a result of these capital inflows, the EM country has been experiencing boom-like conditions.

4 Given the current boom-like conditions in the EM economy, in the *near term*, these capital inflows are *most likely* to lead to:

A a decrease in inflation expectations in the EM.

B an increase in the risk premium for the EM.

C an increase in the EM currency value.

5 If these capital inflows led to an unwanted appreciation in the real value of its currency, the EM country's government would *most likely*:

A impose capital controls.

B decrease taxes on consumption and investment.

C buy its currency in the foreign exchange market.

6 If government actions were ineffective and the EM country's bubble eventually burst, this would *most likely* be reflected in an increase in:

A the risk premium for the EM.

B the EM currency value.

C the long-run equilibrium EM currency value.

Finally, Kwan turns to examining the link between the value of the DM country's currency and movements in the DM country's main stock market index. One of her research associates tells her that, in general, the correlation between equity market returns and changes in exchange rates has been found to be highly positive over time.

7 The statement made by the research associate is:

A correct.

B incorrect, because the correlation is highly negative over time.

C incorrect, because the correlation is not stable and tends to converge toward zero in the long run.

Solution to 1:

C is correct. All else equal, an increase in the expected inflation differential should lead to depreciation of the EM currency.

Solution to 2:

A is correct. During times of intense risk aversion, investors will crowd into the safe haven currency. This tendency implies an increased risk premium demanded by investors to hold the EM currency.

Solution to 3:

C is correct. Over time, the DM country will see its level of external debt rise as a result of the chronic current account imbalance. Eventually, this trend should lead to a downward revision of the DM currency's long-run equilibrium level (via the debt sustainability channel). This is equivalent to an *increase* in the EM currency's long-run exchange rate. A is incorrect because the DM country's current account deficit is likely to lead to a buildup in DM country assets held by the EM country. B is incorrect because, at some point, the currency strength should contribute to deterioration in the trade competitiveness of the country with the trade surplus (the EM country).

Solution to 4:

C is correct. Given the current investor enthusiasm for the EM country's assets and the boom-like conditions in the country, it is most likely that in the near term, the EM currency will appreciate. At the same time, expected inflation in the EM country is also likely increasing and—given the enthusiasm for EM assets—the risk premium is likely decreasing.

Solution to 5:

A is correct. To reduce unwanted appreciation of its currency, the EM country would be most likely to impose capital controls to counteract the surging capital inflows. Because these inflows are often associated with overinvestment and consumption, the EM government would not be likely to encourage these activities through lower taxes. Nor would the EM country be likely to encourage further currency appreciation by intervening in the market to *buy* its own currency.

Solution to 6:

A is correct. Episodes of surging capital flows into EM countries have often ended badly (with a rapid reversal of these inflows as the bubble bursts). This is most likely to be reflected in an increase in the EM risk premium. It is much less likely that a bursting bubble would be reflected in an increase in either the EM currency value or its long-term equilibrium value.

Solution to 7:

C is correct. Correlations between equity returns and exchange rates are unstable in the short term and tend toward zero in the long run.

6 MONETARY AND FISCAL POLICIES

As the foregoing discussion indicates, government policies can have a significant impact on exchange rate movements. We now examine the channels through which government monetary and fiscal policies are transmitted.

6.1 The Mundell–Fleming Model

The Mundell–Fleming model describes how changes in monetary and fiscal policy within a country affect interest rates and economic activity, which in turn leads to changes in capital flows and trade and ultimately to changes in the exchange rate. The model focuses only on aggregate demand and assumes there is sufficient slack in the economy to allow increases in output without price level increases.

In this model, expansionary monetary policy affects growth, in part, by reducing interest rates and thereby increasing investment and consumption spending. Given flexible exchange rates and expansionary monetary policy, downward pressure on domestic interest rates will induce capital to flow to higher-yielding markets, putting downward pressure on the domestic currency. The more responsive capital flows are to interest rate differentials, the greater the depreciation of the currency.

Expansionary fiscal policy—either directly through increased spending or indirectly via lower taxes—typically exerts upward pressure on interest rates because larger budget deficits must be financed. With flexible exchange rates and mobile capital, the rising domestic interest rates will attract capital from lower-yielding markets, putting upward pressure on the domestic currency. If capital flows are highly sensitive to interest rate differentials, then the domestic currency will tend to appreciate substantially. If, however, capital flows are immobile and very insensitive to interest rate differentials, the policy-induced increase in aggregate demand will increase imports and worsen the trade balance, creating downward pressure on the currency with no offsetting capital inflows to provide support for the currency.[20]

The specific mix of monetary and fiscal policies in a country can have a profound effect on its exchange rate. Consider first the case of high capital mobility. With floating exchange rates and high capital mobility, a domestic currency will appreciate given a restrictive domestic monetary policy and/or an expansionary fiscal policy. Similarly, a domestic currency will depreciate given an expansionary domestic monetary policy and/or a restrictive fiscal policy. In Exhibit 5, we show that the combination of a restrictive monetary policy and an expansionary fiscal policy is extremely bullish for a currency when capital mobility is high; likewise, the combination of an expansionary monetary policy and a restrictive fiscal policy is bearish for a currency. The effect on the currency of monetary and fiscal policies that are both expansionary or both restrictive is indeterminate under conditions of high capital mobility.

20 The effect of fiscal policy on the euro is difficult to ascertain because, although the European Union operates under a common monetary policy, the EU countries' fiscal policies are not centrally controlled.

Exhibit 5 Monetary–Fiscal Policy Mix and the Determination of Exchange Rates under Conditions of High Capital Mobility

	Expansionary Monetary Policy	Restrictive Monetary Policy
Expansionary Fiscal Policy	Indeterminate	Domestic currency appreciates
Restrictive Fiscal Policy	Domestic currency depreciates	Indeterminate

Source: Rosenberg (1996), p. 132.

When capital mobility is low, the effects of monetary and fiscal policy on exchange rates will operate primarily through trade flows rather than capital flows. The combination of expansionary monetary *and* fiscal policy will be bearish for a currency. Earlier we said that expansionary fiscal policy will increase imports and hence the trade deficit, creating downward pressure on the currency. Layering on an expansive monetary policy will further boost spending and imports, worsening the trade balance and exacerbating the downward pressure on the currency.

The combination of restrictive monetary *and* fiscal policy will be bullish for a currency. This policy mix will tend to reduce imports, leading to an improvement in the trade balance.

The impact of expansionary monetary and restrictive fiscal policies (or restrictive monetary and expansionary fiscal policies) on aggregate demand and the trade balance, and hence on the exchange rate, is indeterminate under conditions of low capital mobility. Exhibit 6 summarizes these results.

Exhibit 6 Monetary–Fiscal Policy Mix and the Determination of Exchange Rates under Conditions of Low Capital Mobility

	Expansionary Monetary Policy	Restrictive Monetary Policy
Expansionary Fiscal Policy	Domestic currency depreciates	Indeterminate
Restrictive Fiscal Policy	Indeterminate	Domestic currency appreciates

Source: Adapted from Rosenberg (1996), p. 133.

Exhibit 5 is more relevant for the G–10 countries because capital mobility tends to be high in developed economies. Exhibit 6 is more relevant for emerging market economies that restrict capital movement.

A classic case in which a dramatic shift in the policy mix caused dramatic changes in exchange rates was that of Germany in 1990–1992. During that period, the German government pursued a highly expansionary fiscal policy to help facilitate German unification. At the same time, the Bundesbank pursued an extraordinarily restrictive monetary policy to combat the inflationary pressures associated with unification. The expansive fiscal/restrictive monetary policy mix drove German interest rates sharply higher, eventually causing the German currency to appreciate.

6.2 Monetary Models of Exchange Rate Determination

In the Mundell–Fleming model, monetary policy is transmitted to the exchange rate through its impact on interest rates and output. Changes in the price level and/or the inflation rate play no role. Monetary models of exchange rate determination generally take the opposite perspective: Output is fixed and monetary policy affects exchange rates primarily through the price level and the rate of inflation. In this section, we briefly describe two variations of the monetary approach to exchange rate determination.

The monetary approach asserts that an X percent rise in the domestic money supply will produce an X percent rise in the domestic price level. Assuming that purchasing power parity holds—that is, that changes in exchange rates reflect changes in relative inflation rates—a money supply–induced increase (decrease) in domestic prices relative to foreign prices should lead to a proportional decrease (increase) in the domestic currency's value.

One of the major shortcomings of the pure monetary approach is the assumption that purchasing power parity holds in both the short and long runs. Because purchasing power parity rarely holds in either the short or medium run, the pure monetary model may not provide a realistic explanation of the impact of monetary forces on the exchange rate.

To rectify that problem, Dornbusch (1976) constructed a modified monetary model that assumes prices have limited flexibility in the short run but are fully flexible in the long run. The long-run flexibility of the price level ensures that any increase in the domestic money supply will give rise to a proportional increase in domestic prices and thus contribute to a depreciation of the domestic currency in the long run, which is consistent with the pure monetary model. If the domestic price level is assumed to be inflexible in the short run, however, the model implies that the exchange rate is likely to overshoot its long-run PPP path in the short run. With inflexible domestic prices in the short run, any increase in the nominal money supply results in a decline in the domestic interest rate. Assuming that capital is highly mobile, the decline in domestic interest rates will precipitate a capital outflow, which in the short run will cause the domestic currency to depreciate below its new long-run equilibrium level. In the long run, once domestic nominal interest rates rise, the currency will appreciate and move into line with the path predicted by the conventional monetary approach.

Monetary Policy and Exchange Rates: The Historical Evidence

Historically, changes in monetary policy have had a profound impact on exchange rates. In the case of the US dollar, the Federal Reserve's policy of quantitative easing after the global financial crisis resulted in dollar depreciation from mid-2009 to 2011. The subsequent ending of quantitative easing in 2014, along with the anticipation that the United States would raise interest rates before many other countries, played a key role in driving the dollar higher.

Beginning in 2013, Abenomics—fiscal stimulus, monetary easing, and structural reforms—and the use of quantitative easing in Japan led to a steady decline in interest rates and eventually to negative interest rates in 2016. From 2013 to 2015, the value of the yen changed from roughly JPY 90/USD to JPY 120/USD. Likewise, the use of quantitative easing by the European Central Bank in 2015 led to declines in the value of the euro.

Excessively expansionary monetary policies by central banks in emerging markets have often planted the seeds of speculative attacks on their currencies. In the early 1980s, exchange rate crises in Argentina, Brazil, Chile, and Mexico were all preceded by sharp accelerations in domestic credit expansions. In 2012, Venezuela began a period of triple-digit inflation, followed by a massive currency depreciation and an economic crisis.

EXAMPLE 9

Monetary Policy and Exchange Rates

Monique Kwan, the currency strategist at a major foreign exchange dealer, is preparing a report on the outlook for several currencies that she follows. She begins by considering the outlook for the currency of a developed market country with high capital mobility across its borders and a flexible exchange rate. This DM country also has low levels of public and private debt.

Given these conditions, Kwan tries to assess the impact of each of the following policy changes.

1 For the DM currency, increasing the degree of monetary easing (reducing interest rates and increasing money supply) will *most likely*:

- **A** cause the currency to appreciate.
- **B** cause the currency to depreciate.
- **C** have an indeterminate effect on the currency.

2 The pursuit of an expansionary domestic fiscal policy by the DM country will, in the short run, *most likely:*

A cause the domestic currency's value to appreciate.

B cause the domestic currency's value to depreciate.

C have an indeterminate effect on the domestic currency's value.

Next, Kwan turns her attention to an emerging market country that has low levels of public and private debt. Currently, the EM country has a fixed exchange rate but no controls over international capital mobility. However, the country is considering replacing its fixed exchange rate policy with a policy based on capital controls. These proposed controls are meant to reduce international capital mobility by limiting short-term investment flows ("hot money") in and out of its domestic capital markets.

3 To maintain the exchange rate peg while increasing the degree of monetary easing, the EM country will *most likely* have to:

A tighten fiscal policy.

B decrease interest rates.

C buy its own currency in the FX market.

4 After the EM country replaces its currency peg with capital controls, would its exchange rate be unaffected by a tightening in monetary policy?

A Yes.

B No, the domestic currency would appreciate.

C No, the domestic currency would depreciate.

5 After the EM country replaces its currency peg with capital controls, the simultaneous pursuit of a tight monetary policy and a highly expansionary fiscal policy by the EM country will *most likely*:

A cause the currency to appreciate.

B cause the currency to depreciate.

C have an indeterminate effect on the currency.

Solution to 1:

B is correct. A decrease in the policy rate would most likely cause capital to reallocate to higher-yielding markets. This would lead to currency depreciation.

Solution to 2:

A is correct. An expansionary fiscal policy will lead to higher levels of government debt and interest rates, which will attract international capital flows. (In the long run, however, an excessive buildup in debt may eventually cause downward pressure on the domestic currency.)

Solution to 3:

C is correct. The looser monetary policy will lead to exchange rate depreciation. To counter this effect and maintain the currency peg, the central bank will have to intervene in the FX market, buying the country's own currency. A is incorrect because tighter fiscal policy is associated with lower interest rates and is therefore likely to increase rather than mitigate the downward pressure on the domestic currency. Similarly, B is incorrect because a move to lower interest rates would exacerbate the downward pressure on the currency and hence the pressure on the peg.

Solution to 4:

B is correct. In general, capital controls will not completely eliminate capital flows but will limit their magnitude and responsiveness to investment incentives such as interest rate differentials. At a minimum, flows directly related to financing international trade will typically be allowed. The exchange rate will still respond to monetary policy. With limited capital mobility, however, monetary policy's main influence is likely to come through the impact on aggregate demand and the trade balance. A tighter domestic monetary policy will most likely lead to higher interest rates and less domestic demand, including less demand for imported goods. With fewer imports and with exports held constant, there will be modest upward pressure on the currency.

Solution to 5:

C is correct because (1) capital mobility is low, so the induced increase in interest rates is likely to exert only weak upward pressure on the currency; (2) the combined impact on aggregate demand is indeterminate; and (3) if aggregate demand increases, the downward pressure on the currency due to a worsening trade balance may or may not fully offset the upward pressure exerted by capital flows.

6.3 The Portfolio Balance Approach

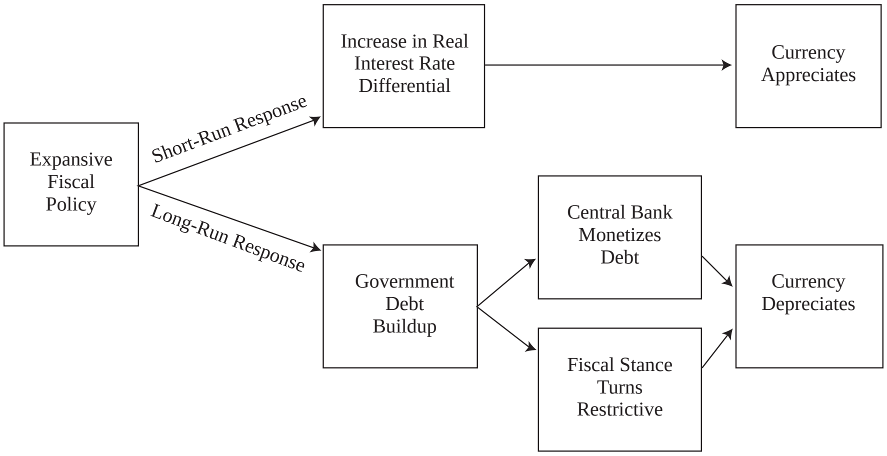
In this section, we re-examine the role fiscal policy plays in determining exchange rates. The Mundell–Fleming model is essentially a short-run model of exchange rate determination. It makes no allowance for the long-term effects of budgetary imbalances that typically arise from sustained fiscal policy actions. The portfolio balance approach to exchange rate determination remedies this limitation. In our previous discussion of the portfolio balance channel, we stated that the currencies of countries with trade deficits will decline over time. We expand that discussion here to more closely examine how exchange rates change over the long term.

In the **portfolio balance approach**, global investors are assumed to hold a diversified portfolio of domestic and foreign assets, including bonds. The desired allocation is assumed to vary in response to changes in expected return and risk considerations. In this framework, a growing government budget deficit leads to a steady increase in the supply of domestic bonds outstanding. These bonds will be willingly held only if investors are compensated in the form of a higher expected return. Such a return could come from (1) higher interest rates and/or a higher risk premium, (2) immediate *depreciation of the currency to a level sufficient to generate anticipation of gains from subsequent currency appreciation*, or (3) some combination of these two factors. The currency adjustments required in the second mechanism are the core of the portfolio balance approach.

One of the major insights one should draw from the portfolio balance model is that *in the long run, governments that run large budget deficits on a sustained basis could eventually see their currencies decline in value.*

The Mundell–Fleming and portfolio balance models can be combined into a single integrated framework in which expansionary fiscal policy under conditions of high capital mobility may be positive for a currency in the short run but negative in the long run. Exhibit 7 illustrates this concept. A domestic currency may rise in value when the expansionary fiscal policy is first put into place. As deficits mount over time and the government's debt obligations rise, however, market participants will begin to wonder how that debt will be financed. If the volume of debt rises to levels that are believed to be unsustainable, market participants may believe that the central bank will eventually be pressured to "monetize" the debt—that is, to buy the government's debt with newly created money. Such a scenario would clearly lead to a rapid reversal of the initial currency appreciation. Alternatively, the market may believe that the government will eventually have to shift toward significant restraint to implement a more restrictive, sustainable fiscal policy over the longer term.

Exhibit 7 The Short- and Long-Run Response of Exchange Rates to Changes in Fiscal Policy



Source: Rosenberg (2003).

EXAMPLE 10

Fiscal Policy and Exchange Rates

Monique Kwan is continuing her analysis of the foreign exchange rate outlook for selected countries. She examines a DM country that has a high degree of capital mobility and a floating-rate currency regime. Kwan notices that although the current outstanding volume of government debt is low, as a percentage of GDP, it is rising sharply as a result of expansionary fiscal policy. Moreover, projections for the government debt-to-GDP ratio point to further increases well into the future.

Kwan uses the Mundell–Fleming and portfolio balance models to form an opinion about both the short-run and long-run implications for the DM country's exchange rate.

1 Over the short run, Kwan is *most likely* to expect:

A appreciation of the DM's currency.

B an increase in the DM's asset prices.

C a decrease in the DM's risk premium.

2 Over the medium term, as the DM country's government debt becomes harder to finance, Kwan would be *most likely* to expect that:

A fiscal policy will turn more accommodative.

B the mark-to-market value of the debt will increase.

C monetary policy will become more accommodative.

3 Assuming that the DM country's government debt becomes harder to finance and there is no change in monetary policy, Kwan is *most likely* to expect that over the longer term, there will be a fiscal policy response that will lead to:

A currency appreciation as yields rise.

B currency depreciation as yields decline.

C an indeterminate impact on the currency, depending on which effect prevails.

Solution to 1:

A is correct. The DM country currently has a low debt load (as a percentage of GDP), and in the short run, its expansionary fiscal policy will lead to higher interest rates and higher real rates relative to other countries. This path should lead to currency appreciation. The higher domestic interest rates will (all else equal) depress local asset prices (so B is incorrect), and the rising debt load is likely to increase rather than decrease the risk premium (so C is incorrect).

Solution to 2:

C is correct. As government debt becomes harder to finance, the government will be tempted to monetize the debt through an accommodative monetary policy. A is incorrect because an inability to finance the debt will make it hard for fiscal policy to become more accommodative. B is incorrect because as investors demand a higher risk premium (a higher return) for holding the DM country's debt, the mark-to-market value of the debt will decline (i.e., bond prices will decrease and bond yields will increase).

Solution to 3:

B is correct. As the DM country's debt ratio deteriorates, foreign investors will demand a higher rate of return to compensate them for the increased risk. Assuming that the central bank will not accommodate (monetize) the rising government debt, the most likely fiscal response is an eventual move toward fiscal consolidation—reducing the public deficit and debt levels that were causing the debt metrics to deteriorate. This policy adjustment would involve issuing fewer government bonds. All else equal, bond yields would decrease, leading to a weaker domestic currency over the longer term.

A is incorrect because currency appreciation is not likely to accompany rising yields when the government is having difficulty financing its deficit. There would be a rising risk premium (a deteriorating investor appetite) for holding DM assets, and hence a currency appreciation would be unlikely despite high DM yields. To avoid paying these high yields on its debt, the DM government would eventually have to take measures to reduce its deficit spending. This approach would eventually help reduce investor risk aversion and DM yields. C is incorrect because given the deterioration in the DM's debt metrics, a depreciation of its exchange rate is likely to be an important part of the restoration of financial market equilibrium.

EXCHANGE RATE MANAGEMENT: INTERVENTION AND CONTROLS

7

Capital flows can be both a blessing and a curse. Capital inflows can be a blessing when they increase domestic investment, thereby increasing a country's economic growth and asset values. Currency appreciation often follows, which increases returns to global investors. Capital inflows can be a curse, however, if they fuel boom-like conditions, asset price bubbles, and overvaluation of a country's currency. If capital inflows then reverse, the result may be a major economic downturn, a significant decline in asset prices, and a large depreciation of the currency. Capital inflows often are driven by a combination of "pull" and "push" factors. Pull factors represent a favorable set of developments that encourage foreign capital inflows. These factors may stem from both the public and private sectors. Examples of better economic management by a government include:

- a decrease in inflation and inflation volatility;
- more-flexible exchange rate regimes;
- improved fiscal positions;
- privatization of state-owned entities;
- liberalization of financial markets; and
- lifting of foreign exchange regulations and controls.

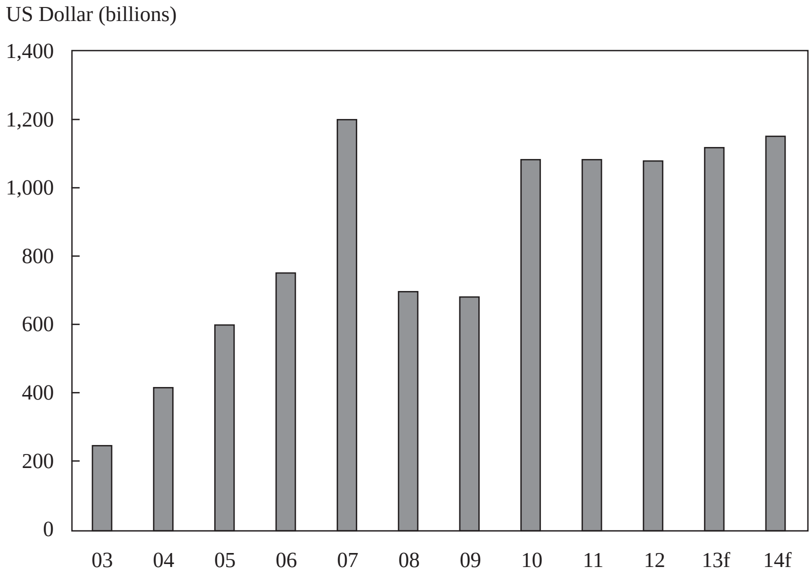
Ideally, these changes will facilitate strong economic growth in the private sector, which will attract further foreign investment. A healthy export sector will generate improvement in the current account balance and an increase in FX reserves, which can be used by the government as a buffer against future speculative attacks. The returns from the currency and assets should increase, increasing the foreign investor's return.

Push factors driving foreign capital inflows are not determined by the domestic policies but arise from the primary sources of internationally mobile capital, notably the investor base in industrial countries. For example, the pursuit of low interest rate policies in industrial countries since the 2008 financial crisis has encouraged global investors to seek higher returns abroad.

Another important push factor is the long-run trend in asset allocation by industrial country investors. For example, many fund managers have traditionally had underweight exposures to emerging market assets, but with the weight of emerging market equities in broad global equity market indexes on the rise (the EM share of world GDP is now around 40%, up from 17% in the 1960s), capital flows to EM countries, in the form of increased allocations to EM equities, are likely to rise.

As indicated in Exhibit 8, net private capital inflows to emerging markets rose steadily between 2003 and 2007, posting nearly a six-fold increase over the period. Both push and pull factors contributed to that surge in capital flows. Net private capital flows to emerging markets tumbled in 2008 and 2009 as heightened risk aversion during the global financial crisis prompted investors to unwind some of their EM exposures in favor of US assets. In 2010, capital flows to emerging markets rose as many EM economies weathered the global financial crisis better than many industrial economies. In addition, the pursuit of ultra-low interest rate policies in the United States, the euro area, and Japan encouraged global investors to invest in higher-yielding EM assets.

Exhibit 8 Net Private Capital Flows to Emerging Markets



Sources: Adapted from Suttle, Koepke, Morkunaite, and Tiftik (2011) and Suttle, Huefner, and Koepke (2013).

However beneficial foreign capital is, policymakers must guard against excessive capital inflows that could quickly be reversed. Capital flow surges planted the seeds of three major currency crises in the 1990s—the Exchange Rate Mechanism (ERM) crisis in the fall of 1992, the Mexican peso crisis in late 1994, and the Asian currency and financial crisis in 1997–1998. Each crisis episode was preceded by a surge in capital inflows and a buildup of huge, highly leveraged speculative positions by local as well as international investors in currencies that eventually came under heavy speculative attack. In the run-up to the ERM crisis, investors—believing that European yield convergence would occur as European monetary union approached—took on highly leveraged long positions in the higher-yielding European currencies financed by short positions in the lower-yielding European currencies. Likewise, in the run-up to the Mexican peso crisis, investors and banks were highly leveraged and made extensive use of derivative products in taking on speculative long Mexican peso/short US dollar positions. And in the run-up to the Asian financial crisis, Asian companies and banks were highly leveraged as they took on a huge volume of short-term dollar- and yen-denominated debt to fund local activities. In each case, the sudden unwinding of those leveraged long speculative positions triggered the attacks on the currencies.

Governments resist excessive inflows and currency bubbles by using capital controls and direct intervention (selling their currency) in the foreign exchange market. Capital controls can take many forms. In the Asian financial crisis, many countries, such as Malaysia, prevented their banks from offering currency transactions in which their currency was sold. As mentioned earlier, Brazil has used a tax to limit currency transactions. In 2006, Thailand required a one-year, non-interest-bearing deposit of 30% of an investment's value to reduce new foreign inflows, which had been appreciating the Thai baht. Vietnam has limited the foreign ownership of local financial institutions. In 2015, Ukraine was removed from the MSCI Frontier Markets equity index after its central bank prevented foreign investors from repatriating funds from the sale of Ukrainian stocks. By 2016, Venezuela had instituted capital controls in the form of four different exchange rates, whereby the rate for selling Venezuelan bolivars for US dollars depended on what the dollars were used for. As a result, many Venezuelans used the black market to obtain dollars.

At one time, capital controls were frowned on as a policy tool for curbing undesired surges in capital inflows. It was generally felt that such controls tended to generate distortions in global trade and finance and that, in all likelihood, market participants would eventually find ways to circumvent the controls. Furthermore, many thought that capital controls imposed by one country could deflect capital flows to other countries, which could complicate monetary and exchange rate policies in those economies. Despite such concerns, the IMF has said that the benefits associated with capital controls may exceed the associated costs. Given the painful lessons that EM policymakers have learned from previous episodes of capital flow surges, some believe that under certain circumstances, capital controls may be needed to prevent exchange rates from overshooting, asset bubbles from forming, and future financial conditions from deteriorating.

Although a case can be made for government intervention and capital controls to limit the potential damage associated with unrestricted inflows of overseas capital, the key issue for policymakers is whether intervention and capital controls will actually work in terms of (1) preventing currencies from appreciating too strongly, (2) reducing the aggregate volume of capital inflows, and (3) enabling monetary authorities to pursue independent monetary policies without having to worry about whether changes in policy rates might attract too much capital from overseas. As an example of the last issue, if a central bank increases interest rates to slow inflation, then capital controls might prevent foreign capital inflows from subsequently depressing interest rates.

Evidence on the effectiveness of direct government intervention suggests that, in the case of industrial countries, the volume of intervention is often quite small relative to the average daily turnover of G–10 currencies in the foreign exchange market. Hence, most studies conclude that the effect of intervention in developed market economies is limited. For most developed market countries, the ratio of official FX reserves held by the respective central banks to the average daily turnover of foreign exchange trading in that currency is negligible. Most industrial countries hold insufficient reserves to significantly affect the supply of and demand for their currency.[21]

The evidence on the effectiveness of government intervention in emerging market currencies is more mixed. Intervention appears to contribute to lower EM exchange rate volatility, but no statistically significant relationship has emerged between the level of EM exchange rates and intervention. Some studies find, however, that EM policymakers might have greater success in controlling exchange rates than their industrial country counterparts because the ratio of EM central bank FX reserve holdings to average daily FX turnover in their domestic currencies is actually quite sizable. With considerably greater firepower in their reserve arsenals, emerging market central banks appear to be in a stronger position than their developed market counterparts to influence the level and path of their exchange rates. What's more, with emerging market central banks' FX reserve holdings expanding at a near-record clip in the past decade, the effectiveness of intervention may be greater now than in the past.

21 If a central bank is intervening in an effort to weaken, rather than strengthen, its own currency, it could (at least in principle) create and sell an unlimited amount of its currency and accumulate a correspondingly large quantity of FX reserves. However, persistent intervention in the FX market can undermine the efficacy of domestic monetary policy.

8 WARNING SIGNS OF A CURRENCY CRISIS

If capital inflows come to a sudden stop, the result may be a financial crisis, in which the economy contracts, asset values plummet, and the currency sharply depreciates. History is filled with examples of currencies that have come under heavy selling pressure within short windows of time. For example, between August 2008 and February 2009, 23 currencies dropped by 25% or more against the US dollar. These included the developed market currencies of Australia, Sweden, and the United Kingdom, which dropped by 35% or more, and the emerging market currencies of Brazil, Russia, and South Korea, which fell by more than 50%.[22]

Currency crises often occur suddenly, with many investors caught by surprise. Once a wave of selling begins, investors and borrowers must immediately reposition their portfolios to avoid excessive capital losses. For example, assume a carry trader had gone long Brazilian real and borrowed US dollars. Upon an initial depreciation of the real, the trader would be inclined to exit the trade by selling real and buying dollars. Or consider a Brazilian public or private borrower that had financed in US dollars. The borrower would also be selling real to buy dollars in order to cover future repayment of the dollar debt. Either of these actions will intensify selling pressure on the depreciated currency. It is this massive liquidation of vulnerable positions, often reinforced by speculative offshore selling, that is largely responsible for the excessive exchange rate movements that occur during currency crises.

Because most crisis episodes have not been adequately anticipated, a great deal of effort has been spent developing early warning systems. One of the problems in developing an early warning system is that views on the underlying causes of currency crises differ greatly. One school of thought contends that currency crises tend to be precipitated by deteriorating economic fundamentals, while a second school contends that currency crises can occur out of the blue, with little evidence of deteriorating fundamentals preceding them.

If, according to the first school of thought, deteriorating economic fundamentals often precede crises, and if those economic fundamentals tend to deteriorate steadily and predictably, then it should be possible to construct an early warning system to anticipate when a currency might be vulnerable.

The second school of thought argues that, although evidence of deteriorating economic fundamentals might explain a relatively large number of currency collapses, there might be cases in which economies with relatively sound fundamentals have their currencies come under attack. Clearly, these currency crises would be more difficult to predict. Events that are largely unrelated to domestic economic fundamentals include sudden adverse shifts in market sentiment that become self-fulfilling prophecies and contagion from crises in other markets. A crisis may spread to a country when, for example, the country devalues its currency to keep its exports competitive with those of another country that devalued.

Recognizing that no single model can correctly anticipate the onset of all crisis episodes, an early warning system might nevertheless be useful in assisting investors in structuring and/or hedging their global portfolios. An ideal early warning system would need to incorporate a number of important features. First, it should have a strong record of predicting actual crises but also should not issue false alarms. Second, it should include macroeconomic indicators whose data are available on a timely basis. If data arrive with a long lag, a crisis could be under way before the early warning system starts flashing red. Third, because currency crises tend to be triggered in

22 Glick and Hutchison (2011).

countries with a number of economic problems, not just one, an ideal early warning system should be broad based, incorporating a wide range of symptoms that crisis-prone currencies might exhibit.

Many studies have been conducted to develop an early warning system for currency crises, typically by constructing a model in which a number of variables constitute the early warning system. Various definitions of currency crises have been used. Although the variables and methodologies differ from one study to the next, the following conditions were identified in one or more studies.[23]

1 Prior to a currency crisis, the capital markets have been liberalized to allow the free flow of capital.
2 There are large inflows of foreign capital (relative to GDP) in the period leading up to a crisis, with short-term funding denominated in a foreign currency being particularly problematic.
3 Currency crises are often preceded by (and often coincide with) banking crises.
4 Countries with fixed or partially fixed exchange rates are more susceptible to currency crises than countries with floating exchange rates.
5 Foreign exchange reserves tend to decline precipitously as a crisis approaches.
6 In the period leading up to a crisis, the currency has risen substantially relative to its historical mean.
7 The ratio of exports to imports (known as "the terms of trade") often deteriorates before a crisis.
8 Broad money growth and the ratio of M2 (a measure of money supply) to bank reserves tend to rise prior to a crisis.
9 Inflation tends to be significantly higher in pre-crisis periods compared with tranquil periods.

These factors are usually interrelated and often feed off one another. For example, in the case of the first five factors, large inflows of foreign capital occur because the financial markets have been liberalized and domestic banks have borrowed abroad. If the borrowing is denominated in a foreign currency and the domestic currency initially depreciates, the bank may have trouble servicing its debt, especially when the debt is of shorter maturity. This scenario may cause foreign investors to withdraw capital and speculators to short the currency, with their actions causing further declines in the currency. If the government is trying to maintain the currency's value, it could increase interest rates to stem capital outflows or defend its currency using direct intervention. The former action may worsen the banking industry's condition and slow down the economy. In the latter approach, the government will have to spend down its foreign currency reserves to buy its own currency in the foreign exchange markets. If the government appears unwilling or unable to defend its currency, then capital outflows and speculative attacks will increase.

The fifth through seventh factors are related because an overvalued currency may make the country's exports less competitive. With fewer exports, the country is not able to earn as much foreign currency. Other interrelationships occur because these factors often coincide.

23 The material presented here is drawn from the following: Babecký, Havránek, Matějů, Rusnák, Šmídková, and Vašíček (2013, 2014); Beim and Calomiris (2001); Bekaert and Hodrick (2012); Bussiére (2013); Daniels and VanHoose (2018); DeRosa (2009); Glick and Hutchison (2011); Reinhart and Rogoff (2009); and Weithers (2006).

Models cannot predict every crisis, and they sometimes generate false alarms. Nevertheless, an early warning system can be useful in assessing and preparing for potential negative tail risks. As with any analytical tool, the implementation of an early warning system requires integration with other analysis and judgment that cannot be easily quantified or conceptualized.

Iceland's Currency Crisis of 2008

Iceland, a country with a population of 320,000, had traditionally relied on the fishing, energy, and aluminum industries for economic growth.[24] That began to change in 2001, when the banking industry was liberalized. Three banks dominated the Icelandic banking industry: Glitnir, Kaupthing, and Landsbanki. Given Iceland's small population, these banks sought growth by offering short-term, internet-based deposit accounts to foreign investors. These accounts offered attractive interest rates and were denominated in foreign currencies. In particular, many of the depositors were British, Dutch, and other European citizens who held deposit accounts denominated in pounds and euros.

With government guarantees on their deposit accounts, the banking industry grew rapidly. The largest bank, Kaupthing, experienced asset growth of 30 times between 2000 and 2008. The three banks increased lending rapidly, with many of their loans being long term, resulting in a maturity mismatch of assets and liabilities. The banks' assets were more than 14 times the country's GDP, while foreign debt was five times GDP. The three banks constituted more than 70% of the national stock market capitalization.

The economy expanded at a real growth rate above 20% annually between 2002 and 2005, and many Icelanders left traditional industries to work in the banks. Iceland earned the nickname "Nordic Tiger" as per capita GDP approached USD 70,000 in 2007. The krona increased in value against the US dollar by 40% between 2001 and 2007. By 2007, the unemployment rate was less than 1%. Icelanders went on a shopping spree for consumer goods, in part by using loans tied to the value of foreign currencies, motivated by lower interest rates abroad. A 2002 trade surplus turned into a trade deficit in the years 2003–2007. Iceland's external debt in 2008 was more than seven times its GDP and 14 times its export revenue. Broad-based monetary aggregates grew at a rate of 14%–35% annually from 2002 to 2007. By the fall of 2008, inflation had reached 14%.

As the global financial crisis unfolded in 2008, interbank lending declined and Icelandic banks were unable to roll over their short-term debt. Anxious foreign depositors began withdrawing their funds. In the first half of 2008, the krona depreciated by more than 40% against the euro. As the Icelandic currency declined in value, it became more difficult for the banks to meet depositors' liquidity demands, while at the same time the banks' depreciating krona-denominated assets could not be used for collateral financing.

The three banks collapsed in 2008. Unfortunately for foreign depositors, because of the relative size of the banks, the government guaranteed only domestic deposits. Iceland's central bank became technically insolvent, as its EUR 2 billion in assets was dwarfed by Iceland's debt to foreign banks of EUR 50 billion. Trading in the stock market was suspended in October 2008. When it reopened several days later, the Icelandic Stock Market Index fell by more than 77% as a result of the elimination of the three banks' equity value.

The government attempted to peg the krona to the euro in October 2008 but abandoned the peg one day later. When trading in the currency was resumed later that month, the currency value fell by more than 60% and trading was eventually suspended. Iceland increased interest rates to 18% to stem outflows of krona and imposed capital controls

24 The discussion presented here is drawn from the Federal Reserve Bank of St. Louis database; Bagus and Howden (2009); Bekaert and Hodrick (2018); Daniels and VanHoose (2011); Forelle (2008a and 2008b); Forelle and Davis (2008); and Matsangou (2015).

on the selling of krona for foreign currency. The Icelandic economy contracted, and per capita GDP fell 9.2% in 2009. By the spring of 2009, unemployment was 9%. The country subsequently required a bailout from the IMF and its neighbors of USD 4.6 billion.

EXAMPLE 11

Currency Crises

Monique Kwan now turns her attention to the likelihood of crises in various emerging market currencies. She discusses this matter with a research associate, who tells her that the historical record of currency crises shows that most of these episodes were not very well anticipated by investors (in terms of their positioning), by the bond markets (in terms of yield spreads between countries), or by major credit rating agencies and economists (in terms of the sovereign credit ratings and forecasts, respectively).

1 The research associate is *most likely*:

- A correct.
- B incorrect, because most credit rating agencies and economists typically change their forecasts prior to a crisis.
- C incorrect, because investor positioning and international yield differentials typically shift prior to a crisis.

Kwan delves further into the historical record of currency crises. She concludes that even countries with relatively sound economic fundamentals can fall victim to these crisis episodes and that these attacks can occur when sentiment shifts for reasons unrelated to economic fundamentals.

2 Kwan's conclusion is *most likely*:

- A correct.
- B incorrect, because there are few historical crises involving currencies of countries with sound economic fundamentals.
- C incorrect, because there are few historical episodes in which a sudden adverse shift in market sentiment occurs that is unrelated to economic fundamentals.

To better advise the firm's clients on the likelihood of currency crises, Kwan tries to formulate an early warning system for these episodes. She recognizes that a typical currency crisis tends to be triggered by a number of economic problems, not just one.

3 Kwan's early warning system is *least likely* to indicate an impending crisis when there is:

- A an expansionary monetary policy.
- B an overly appreciated exchange rate.
- C a rising level of foreign exchange reserves at the central bank.

4 Kwan's early warning system would *most likely* be better if it:

- A had a strong record of predicting actual crises, even if it generates a lot of false signals.

B included a wide variety of economic indicators, including those for which data are available only with a significant lag.

C started flashing well in advance of an actual currency crisis to give market participants time to adjust or hedge their portfolios before the crisis hits.

Solution to 1:

A is correct. Currency crises often catch most market participants and analysts by surprise.

Solution to 2:

A is correct. Even countries with sound economic fundamentals can be subject to a currency crisis, including instances when market sentiment shifts for non-economic reasons.

Solution to 3:

C is correct. A high level of foreign exchange reserves held by a country typically decreases the likelihood of a currency crisis.

Solution to 4:

C is correct. Early warnings are a positive factor in judging the effectiveness of the system, whereas false signals and the use of lagged data would be considered negative factors.

SUMMARY

Exchange rates are among the most difficult financial market prices to understand and therefore to value. There is no simple, robust framework that investors can rely on in assessing the appropriate level and likely movements of exchange rates.

Most economists believe that there is an equilibrium level or a path to that equilibrium value that a currency will gravitate toward in the long run. Although short- and medium-term cyclical deviations from the long-run equilibrium path can be sizable and persistent, fundamental forces should eventually drive the currency back toward its long-run equilibrium path. Evidence suggests that misalignments tend to build up gradually over time. As these misalignments build, they are likely to generate serious economic imbalances that will eventually lead to correction of the underlying exchange rate misalignment.

In this reading, we have described how changes in monetary policy, fiscal policy, current account trends, and capital flows affect exchange rate trends, as well as what role government intervention and capital controls can play in counteracting potentially undesirable exchange rate movements. The reading makes the following key points:

- Spot exchange rates apply to trades for the next settlement date (usually $T + 2$) for a given currency pair. Forward exchange rates apply to trades to be settled at any longer maturity.
- Market makers quote bid and offer prices (in terms of the *price currency*) at which they will buy or sell the *base currency.*
  - The offer price is always higher than the bid price.
  - The counterparty that asks for a two-sided price quote has the option (but not the obligation) to deal at either the bid or offer price quoted.

- The bid–offer spread depends on (1) the currency pair involved, (2) the time of day, (3) market volatility, (4) the transaction size, and (5) the relationship between the dealer and client. Spreads are tightest in highly liquid currency pairs, when the key market centers are open, and when market volatility is relatively low.

- Absence of arbitrage requires the following:
  - The bid (offer) shown by a dealer in the interbank market cannot be higher (lower) than the current interbank offer (bid) price.
  - The cross-rate bids (offers) posted by a dealer must be lower (higher) than the implied cross-rate offers (bids) available in the interbank market. If they are not, then a triangular arbitrage opportunity arises.
- Forward exchange rates are quoted in terms of points to be added to the spot exchange rate. If the points are positive (negative), the base currency is trading at a forward premium (discount). The points are proportional to the interest rate differential and approximately proportional to the time to maturity.
- International parity conditions show us how expected inflation, interest rate differentials, forward exchange rates, and expected future spot exchange rates are linked. In an ideal world:
  - relative expected inflation rates should determine relative nominal interest rates;
  - relative interest rates should determine forward exchange rates; and
  - forward exchange rates should correctly anticipate the path of the future spot exchange rate.
- International parity conditions tell us that countries with high (low) expected inflation rates should see their currencies depreciate (appreciate) over time, that high-yield currencies should depreciate relative to low-yield currencies over time, and that forward exchange rates should function as unbiased predictors of future spot exchange rates.
- With the exception of covered interest rate parity, which is enforced by arbitrage, the key international parity conditions rarely hold in either the short or medium term. However, the parity conditions tend to hold over relatively long horizons.
- According to the theory of covered interest rate parity, a foreign-currency-denominated money market investment that is completely hedged against exchange rate risk in the forward market should yield exactly the same return as an otherwise identical domestic money market investment.
- According to the theory of uncovered interest rate parity, the expected change in a domestic currency's value should be fully reflected in domestic–foreign interest rate spreads. Hence, an unhedged foreign-currency-denominated money market investment is expected to yield the same return as an otherwise identical domestic money market investment.
- According to the *ex ante* purchasing power parity condition, expected changes in exchange rates should equal the difference in expected national inflation rates.
- If both *ex ante* purchasing power parity and uncovered interest rate parity held, real interest rates across all markets would be the same. This result is real interest rate parity.

- The international Fisher effect says that the nominal interest rate differential between two currencies equals the difference between the expected inflation rates. The international Fisher effect assumes that risk premia are the same throughout the world.
- If both covered and uncovered interest rate parity held, then forward rate parity would hold and the market would set the forward exchange rate equal to the expected spot exchange rate: The forward exchange rate would serve as an unbiased predictor of the future spot exchange rate.
- Most studies find that high-yield currencies do not depreciate and low-yield currencies do not appreciate as much as yield spreads would suggest over short to medium periods, thus violating the theory of uncovered interest rate parity.
- Carry trades overweight high-yield currencies at the expense of low-yield currencies. Historically, carry trades have generated attractive returns in benign market conditions but tend to perform poorly (i.e., are subject to crash risk) when market conditions are highly volatile.
- According to a balance of payments approach, countries that run persistent current account deficits will generally see their currencies weaken over time. Similarly, countries that run persistent current account surpluses will tend to see their currencies appreciate over time.
- Large current account imbalances can persist for long periods of time before they trigger an adjustment in exchange rates.
- Greater financial integration of the world's capital markets and greater freedom of capital to flow across national borders have increased the importance of global capital flows in determining exchange rates.
- Countries that institute relatively tight monetary policies, introduce structural economic reforms, and lower budget deficits will often see their currencies strengthen over time as capital flows respond positively to relatively high nominal interest rates, lower inflation expectations, a lower risk premium, and an upward revision in the market's assessment of what exchange rate level constitutes long-run fair value.
- Monetary policy affects the exchange rate through a variety of channels. In the Mundell–Fleming model, it does so primarily through the interest rate sensitivity of capital flows, strengthening the currency when monetary policy is tightened and weakening it when monetary policy is eased. The more sensitive capital flows are to the change in interest rates, the greater the exchange rate's responsiveness to the change in monetary policy.
- In the monetary model of exchange rate determination, monetary policy is deemed to have a direct impact on the actual and expected path of inflation, which, via purchasing power parity, translates into a corresponding impact on the exchange rate.
- Countries that pursue overly easy monetary policies will see their currencies depreciate over time.
- In the Mundell–Fleming model, an expansionary fiscal policy typically results in a rise in domestic interest rates and an increase in economic activity. The rise in domestic interest rates should induce a capital inflow, which is positive for the domestic currency, but the rise in economic activity should contribute to a deterioration of the trade balance, which is negative for the domestic currency. The more mobile capital flows are, the greater the likelihood that the induced inflow of capital will dominate the deterioration in trade.

- Under conditions of high capital mobility, countries that simultaneously pursue expansionary fiscal policies and relatively tight monetary policies should see their currencies strengthen over time.
- The portfolio balance model of exchange rate determination asserts that increases in government debt resulting from a rising budget deficit will be willingly held by investors only if they are compensated in the form of a higher expected return. The higher expected return could come from (1) higher interest rates and/or a higher risk premium, (2) depreciation of the currency to a level sufficient to generate anticipation of gains from subsequent currency appreciation, or (3) some combination of the two.
- Surges in capital inflows can fuel boom-like conditions, asset price bubbles, and currency overvaluation.
- Many consider capital controls to be a legitimate part of a policymaker's toolkit. The IMF believes that capital controls may be needed to prevent exchange rates from overshooting, asset price bubbles from forming, and future financial conditions from deteriorating.
- The evidence indicates that government policies have had a significant impact on the course of exchange rates. Relative to developed countries, emerging markets may have greater success in managing their exchange rates because of their large foreign exchange reserve holdings, which appear sizable relative to the limited turnover of FX transactions in many emerging markets.
- Although each currency crisis is distinct in some respects, the following factors were identified in one or more studies:
  1. Prior to a currency crisis, the capital markets have been liberalized to allow the free flow of capital.
  2. There are large inflows of foreign capital (relative to GDP) in the period leading up to a crisis, with short-term funding denominated in a foreign currency being particularly problematic.
  3. Currency crises are often preceded by (and often coincide with) banking crises.
  4. Countries with fixed or partially fixed exchange rates are more susceptible to currency crises than countries with floating exchange rates.
  5. Foreign exchange reserves tend to decline precipitously as a crisis approaches.
  6. In the period leading up to a crisis, the currency has risen substantially relative to its historical mean.
  7. The terms of trade (exports relative to imports) often deteriorate before a crisis.
  8. Broad money growth and the ratio of M2 (a measure of money supply) to bank reserves tend to rise prior to a crisis.
  9. Inflation tends to be significantly higher in pre-crisis periods compared with tranquil periods.

REFERENCES

Babecký, Jan, Tomáš Havránek, Jakub Matějů, Marek Rusnák, Kateřina Šmídková, and Bořek Vašíček. 2013. "Leading Indicators of Crisis Incidence: Evidence from Developed Countries." *Journal of International Money and Finance*, vol. 35, June: 1–19.

Babecký, Jan, Tomáš Havránek, Jakub Matějů, Marek Rusnák, Kateřina Šmídková, and Bořek Vašíček. 2014. "Banking, Debt, and Currency Crises in Developed Countries: Stylized Facts and Early Warning Indicators." *Journal of Financial Stability*, vol. 15, December: 1–17.

Bagus, Philipp, and David Howden. 2009. "Iceland's Banking Crisis: The Meltdown of an Interventionist Financial System." Mises Institute (9 June).

Beim, David O., and Charles W. Calomiris. 2001. *Emerging Financial Markets*. New York: McGraw-Hill.

Bekaert, Geert, and Robert Hodrick. 2018. *International Financial Management*, 3rd ed. Cambridge: Cambridge University Press.

Bussière, Matthieu. 2013. "Balance of Payment Crises in Emerging Markets: How Early Were the 'Early' Warning Signals?" *Applied Economics*, vol. 45, no. 12:1610–1623.

Daniels, Joseph P., and David D. VanHoose. 2018. *Global Economic Issues and Policies*, 4th ed. London: Routledge.

DeRosa, David F. 2009. *Central Banking and Monetary Policy in Emerging-Markets Nations*. Charlottesville, VA: Research Foundation of CFA Institute.

Dimson, Elroy, Paul Marsh, Paul McGinnie, Mike Staunton, and Jonathan Wilmot. 2012. *Credit Suisse Global Investment Returns Yearbook 2012*. Zurich: Credit Suisse Research Institute.

Dornbusch, Rudiger. 1976. "Expectations and Exchange Rate Dynamics." *Journal of Political Economy*, vol. 84, no. 6 (December): 1161–1176.

Fleming, J. Marcus. 1962. "Domestic Financial Policies under Fixed and Floating Exchange Rates." *IMF Staff Papers*, vol. 9, no. 3 (November): 319–379.

Forelle, Charles. 2008a. "Iceland Scrambles for Cash to Save Banks, Economy: Decision Makers Consider Selling Pension-Fund Assets Held Overseas and Seeking Aid from Nordic Central Banks." *Wall Street Journal* (6 October).

Forelle, Charles. 2008b. "The Isle That Rattled the World: Tiny Iceland Created a Vast Bubble, Leaving Wreckage Everywhere When It Popped." *Wall Street Journal* (27 December).

Forelle, Charles, and Bob Davis. 2008. "Iceland Raises Key Rate to 18% to Defend Currency." *Wall Street Journal* (29 October).

Frankel, Jeffrey, and Andrew Rose. 1996. "Currency Crashes in Emerging Markets: An Empirical Treatment." *Journal of International Economics*, vol. 41, no. 3–4: 351–366.

Glick, Reuven, and Michael Hutchison. 2011. "Currency Crises." Working Paper no. 2011-22, Federal Reserve Bank of San Francisco.

IMF. 1998. *World Economic Outlook, May 1998: Financial Crises: Causes and Indicators*. Washington, DC: International Monetary Fund.

IMF. 2007. "Exchange Rates and the Adjustment of External Imbalances." In *World Economic Outlook, April 2007: Spillovers and Cycles in the Global Economy*. Washington, DC: International Monetary Fund.

Isard, Peter, Hamid Faruqee, G. Russell Kincaid, and Martin Fetherston. 2001. "Methodology for Current Account and Exchange Rate Assessments." IMF Occasional Paper no. 209 (December).

Magud, Nichols E., Carmen M. Reinhart, and Kenneth S. Rogoff. 2011. "Capital Controls: Myth and Reality—A Portfolio Balance Approach." NBER Working Paper 16805 (February).

Matsangou, Elizabeth. 2015. "Failing Banks, Winning Economy: The Truth about Iceland's Recovery." *World Finance* (15 September).

Mundell, Robert A. 1962. "The Appropriate Use of Monetary and Fiscal Policy for Internal and External Stability." *IMF Staff Papers*, vol. 9, no. 1 (March): 70–79.

Mundell, Robert A. 1963. "Capital Mobility and Stabilization Policy under Fixed and Flexible Exchange Rates." *Canadian Journal of Economics and Political Science*, vol. 29, no. 4 (November): 475–485.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly*. Princeton, NJ: Princeton University Press.

Rosenberg, Michael R. 1996. *Currency Forecasting: A Guide to Fundamental and Technical Models of Exchange Rate Determination*. Chicago: Irwin Professional Publishing.

Rosenberg, Michael R. 2002. *Deutsche Bank Guide to Exchange Rate Determination*. Chicago: Irwin Professional Publishing.

Rosenberg, Michael R. 2003. *Exchange Rate Determination: Models and Strategies for Exchange Rate Forecasting*. New York: McGraw-Hill.

Suttle, Philip, Felix Huefner, and Robin Koepke. 2013. "Capital Flows to Emerging Market Economies." IIF Research Note, Institute of International Finance (22 January).

Suttle, Philip, Robin Koepke, Kristina Morkunaite, and Emre Tiftik. 2011. "Capital Flows to Emerging Market Economies." IIF Research Note, Institute of International Finance (25 September).

Weithers, Tim. 2006. *Foreign Exchange: A Practical Guide to the FX Markets*. Hoboken, NJ: John Wiley & Sons.

APPENDIX

Currency Codes Used in This Reading	
USD	US dollar
EUR	euro
GBP	UK pound
JPY	Japanese yen
MXN	Mexican peso
CHF	Swiss franc
CAD	Canadian dollar
SEK	Swedish krona
AUD	Australian dollar
KRW	Korean won
NZD	New Zealand dollar

PRACTICE PROBLEMS

The following information relates to Questions 1–5

Ed Smith is a new trainee in the foreign exchange (FX) services department of a major global bank. Smith's focus is to assist senior FX trader, Feliz Mehmet, CFA. Mehmet mentions that an Indian corporate client exporting to the United Kingdom wants to estimate the potential hedging cost for a sale closing in one year. Smith is to determine the premium/discount for an annual (360 day) forward contract using the exchange rate data presented in Exhibit 1.

Exhibit 1 Select Currency Data for GBP and INR

Spot (INR/GBP)	79.5093
Annual (360-day) Libor (GBP)	5.43%
Annual (360-day) Libor (INR)	7.52%

Mehmet is also looking at two possible trades to determine their profit potential. The first trade involves a possible triangular arbitrage trade using the Swiss, US and Brazilian currencies, to be executed based on a dealer's bid/offer rate quote of 0.5161/0.5163 in CHF/BRL and the interbank spot rate quotes presented in Exhibit 2.

Exhibit 2 Interbank Market Quotes

Currency Pair	Bid/Offer
CHF/USD	0.9099/0.9101
BRL/USD	1.7790/1.7792

Mehmet is also considering a carry trade involving the USD and the Euro. He anticipates it will generate a higher return than buying a one-year domestic note at the current market quote due to low US interest rates and his predictions of exchange rates in one year. To help Mehmet assess the carry trade, Smith provides Mehmet with selected current market data and his one year forecasts in Exhibit 3.

Exhibit 3 Spot Rates and Interest Rates for Proposed Carry Trade

Today's one-year Libor		Currency pair (Price/Base)	Spot rate today	Projected spot rate in one year
USD	0.80%	CAD/USD	1.0055	1.0006
CAD	1.71%	EUR/CAD	0.7218	0.7279
EUR	2.20%			

Finally, Mehmet asks Smith to assist with a trade involving a US multinational customer operating in Europe and Japan. The customer is a very cost conscious industrial company with a AA credit rating and strives to execute its currency trades at the most favorable bid/offer spread. Because its Japanese subsidiary is about to close on a major European acquisition in three business days, the client wants to lock in a trade involving the Japanese yen and the Euro as early as possible the next morning, preferably by 8:05 AM New York time.

At lunch, Smith and other FX trainees discuss how best to analyze currency market volatility from ongoing financial crises. The group agrees that a theoretical explanation of exchange rate movements, such as the framework of the international parity conditions, should be applicable across all trading environments. They note such analysis should enable traders to anticipate future spot exchange rates. But they disagree on which parity condition best predicts exchange rates, voicing several different assessments. Smith concludes the discussion on parity conditions by stating to the trainees:

> "I believe that in the current environment both covered and uncovered interest rate parity conditions are in effect."

1 Based upon Exhibit 1, the forward premium (discount) for a 360-day INR/GBP forward contract is *closest* to:

A −1.546.

B 1.546.

C 1.576.

2 Based on Exhibit 2, the *most* appropriate recommendation regarding the triangular arbitrage trade is to:

A decline the trade, no arbitrage profits are possible.

B execute the trade, buy BRL in the interbank market and sell it to the dealer.

C execute the trade, buy BRL from the dealer and sell it in the interbank market.

3 Based on Exhibit 3, the potential all-in USD return on the carry trade is *closest* to:

A 1.04%.

B 1.40%.

C 1.84%.

4 The factor *least likely* to lead to a narrow bid/offer spread for the industrial company's needed currency trade is the:

A timing of its trade.

B company's credit rating.

C pair of currencies involved.

5 If Smith's statement on parity conditions is correct, future spot exchange rates are *most likely* to be forecast by:

A current spot rates.

B forward exchange rates.

C inflation rate differentials.

The following information relates to Questions 6–12

Connor Wagener, a student at the University of Canterbury in New Zealand, has been asked to prepare a presentation on foreign exchange rates for his International Business course. Wagener has a basic understanding of exchange rates, but would like a practitioner's perspective, and he has arranged an interview with currency trader Hannah McFadden. During the interview, Wagener asks McFadden:

> "Could you explain what drives exchange rates? I'm curious as to why our New Zealand dollar was affected by the European debt crisis in 2011 and what other factors impact it."

In response, McFadden begins with a general discussion of exchange rates. She notes that international parity conditions illustrate how exchange rates are linked to expected inflation, interest rate differences, and forward exchange rates as well as current and expected future spot rates. McFadden states:

Statement 1 "Fortunately, the international parity condition most relevant for FX carry trades does not always hold."

McFadden continues her discussion:

> "FX carry traders go long (i.e. buy) high-yield currencies and fund their position by shorting, that is borrowing in, low-yield currencies. Unfortunately, crashes in currency values can occur which create financial crises as traders unwind their positions. For example, in 2008, the New Zealand dollar was negatively impacted when highly leveraged carry trades were unwound. In addition to investors, consumers and business owners can also affect currency exchange rates through their impact on their country's balance of payments. For example, if New Zealand consumers purchase more goods from China than New Zealand businesses sell to China, New Zealand will run a trade account deficit with China."

McFadden further explains:

Statement 2 "A trade surplus will tend to cause the currency of the country in surplus to appreciate while a deficit will cause currency depreciation. Exchange rate changes will result in immediate adjustments in the prices of traded goods as well as in the demand for imports and exports. These changes will immediately correct the trade imbalance."

McFadden next addresses the influence of monetary and fiscal policy on exchange rates:

> "Countries also exert significant influence on exchange rates through both the initial mix of their fiscal and monetary policies, and also by subsequent adjustments to those policies. Various models have been developed to

identify how these policies affect exchange rates. The Mundell-Fleming model addresses how changes in both fiscal and monetary policies affect interest rates and ultimately exchange rates in the short-term."

McFadden describes monetary models by stating:

Statement 3 "Monetary models of exchange rate determination focus on the effects of inflation, price level changes, and risk premium adjustments."

McFadden continues her discussion:

"So far, we've touched on balance of payments and monetary policy. The portfolio-balance model addresses the impacts of sustained fiscal policy on exchange rates. I must take a client call, but will return shortly. In the meantime, here is some relevant literature on the models I mentioned along with a couple of questions for you to consider:

Question 1 Assume an emerging market (EM) country has restrictive monetary and fiscal policies under low capital mobility conditions. Are these policies likely to lead to currency appreciation, currency depreciation, or to have no impact?

Question 2 Assume a developed market (DM) country has an expansive fiscal policy under high capital mobility conditions. Why is its currency most likely to depreciate in the long-run under an integrated Mundell-Fleming and portfolio-balance approach?"

Upon her return, Wagener and McFadden review the questions. McFadden notes that capital flows can have a significant impact on exchange rates and have contributed to currency crises in both EM and DM countries. She explains that central banks, like the Reserve Bank of New Zealand, use FX market intervention as a tool to manage exchange rates. McFadden states:

Statement 4 "Some studies have found that EM central banks tend to be more effective in using exchange rate intervention than DM central banks, primarily because of one important factor."

McFadden continues her discussion:

Statement 5 "I mentioned that capital inflows could cause a currency crisis, leaving fund managers with significant losses. In the period leading up to a currency crisis, I would predict that an affected country's:

Prediction 1 foreign exchange reserves will increase.

Prediction 2 broad money growth will increase.

Prediction 3 the exchange rate will be substantially higher than its mean level during tranquil periods.

After the interview, McFadden agrees to meet the following week to discuss more recent events on the New Zealand dollar.

6 The international parity condition McFadden is referring to in Statement 1 is:

A purchasing power parity.

B covered interest rate parity.

C uncovered interest rate parity.

7 In Statement 2, McFadden is *most likely* failing to consider the:
   A initial gap between the country's imports and exports.
   B effect of an initial trade deficit on a countries' exchange rates.
   C lag in the response of import and export demand to price changes.
8 The *least* appropriate factor used to describe the type of models mentioned in Statement 3 is:
   A inflation.
   B price level changes.
   C risk premium adjustments.
9 The best response to Question 1 is that the policies will:
   A have no impact.
   B lead to currency appreciation.
   C lead to currency depreciation.
10 The most likely response to Question 2 is a(n):
   A increase in the price level.
   B decrease in risk premiums.
   C increase in government debt.
11 The factor that McFadden is *most likely* referring to in Statement 4 is:
   A FX reserve levels.
   B domestic demand.
   C the level of capital flows.
12 Which of McFadden's predictions in Statement 5 is *least likely to be correct*?
   A Prediction 1
   B Prediction 2
   C Prediction 3

The following information relates to Question 13–20

Anna Goldsworthy is the chief financial officer of a manufacturing firm headquartered in the United Kingdom. She is responsible for overseeing exposure to price risk in both the commodity and currency markets. Goldsworthy is settling her end-of-quarter transactions and creating reports. Her intern, Scott Underwood, assists her in this process.

The firm hedges input costs using forward contracts that are priced in US dollars (USD) and Mexican pesos (MXN). Processed goods are packaged for sale under licensing agreements with firms in foreign markets. Goldsworthy is expecting to receive a customer payment of JPY 225,000,000 (Japanese yen) that she wants to convert to pounds sterling (GBP). Underwood gathers the exchange rates from Dealer A in Exhibit 1.

Exhibit 1 Dealer A's Spot Exchange Rates

	Spot Exchange Rates		
Currency Pair (Price/Base)	Bid	Offer	Midpoint
JPY/GBP	187.39	187.43	187.41
MXN/USD	17.147	17.330	17.239
GBP/EUR	0.7342	0.7344	0.7343
USD/EUR	1.1572	1.1576	1.1574
USD/GBP	1.5762	1.5766	1.5764

The firm must also buy USD to pay a major supplier. Goldsworthy calls Dealer A with specific details of the transaction and asks to verify the USD/GBP quote. Dealer A calls her back later with a revised USD/GBP bid/offer quote of 1.5760/1.5768.

Goldsworthy must purchase MXN 27,000,000 to pay an invoice at the end of the quarter. In addition to the quotes from Dealer A, Underwood contacts Dealer B, who provides a bid/offer price of GBP/MXN 0.0366/0.0372. To check whether the dealer quotes are reflective of an efficient market, Underwood examines whether the prices allow for an arbitrage profit.

In three months, the firm will receive EUR 5,000,000 (euros) from another customer. Six months ago, the firm sold EUR 5,000,000 against the GBP using a nine-month forward contract at an all-in price of GBP/EUR 0.7400. To mark the position to market, Underwood collects the GBP/EUR forward rates in Exhibit 2.

Exhibit 2 GBP/EUR Forward Rates

Maturity	Forward Points
One month	4.40/4.55
Three months	14.0/15.0
Six months	29.0/30.0

Goldsworthy also asks for the current 90-day Libors for the major currencies. Selected three-month Libors (annualized) are shown in Exhibit 3. Goldsworthy studies Exhibit 3 and says, "We have the spot rate and the 90-day forward rate for GBP/EUR. As long as we have the GBP 90-day Libor, we will be able to calculate the implied EUR 90-day Libor."

Exhibit 3 90-Day Libor

Currency	Annualized Rate
GBP	0.5800%
JPY	0.0893%
USD	0.3300%

After reading a draft report, Underwood notes, "We do not hedge the incoming Japanese yen cash flow. Your report asks for a forecast of the JPY/GBP exchange rate in 90 days. We know the JPY/GBP spot exchange rate." He asks, "Does the information we have collected tell us what the JPY/GBP exchange rate will be in 90 days?"

Goldsworthy replies, "The JPY/GBP exchange rate in 90 days would be a valuable piece of information to know. An international parity condition can be used to provide an estimate of the future spot rate."

13 Using the quotes in Exhibit 1, the amount received by Goldsworthy from converting JPY 225,000,000 will be *closest* to:

- **A** GBP 1,200,448.
- **B** GBP 1,200,576.
- **C** GBP 1,200,704.

14 Using Exhibit 1, which of the following would be the *best* reason for the revised USD/GBP dealer quote of 1.5760/1.5768?

- **A** A request for a much larger transaction
- **B** A drop in volatility in the USD/GBP market
- **C** A request to trade when both New York and London trading centers are opened

15 Using the quotes from Dealer A and B, the triangular arbitrage profit on a transaction of MXN 27,000,000 would be *closest* to:

- **A** GBP 0.
- **B** GBP 5,400.
- **C** GBP 10,800.

16 Based on Exhibits 1, 2, and 3, the mark-to-market gain for Goldsworthy's forward position is *closest* to:

- **A** GBP 20,470.
- **B** GBP 20,500.
- **C** GBP 21,968.

17 Based on Exhibit 2, Underwood should conclude that three-month EUR Libor is:

- **A** below three-month GBP Libor.
- **B** equal to three-month GBP Libor.
- **C** above three-month GBP Libor.

18 Based on the exchange rate midpoint in Exhibit 1 and the rates in Exhibit 3, the 90-day forward premium (discount) for the USD/GBP would be *closest* to:

- **A** –0.0040.
- **B** –0.0010.
- **C** +0.0010.

19 Using Exhibits 1, 2, and 3, which international parity condition would Goldsworthy *most likely* use to calculate the EUR Libor?

- **A** Real interest rate parity
- **B** Covered interest rate parity
- **C** Uncovered interest rate parity

20 The international parity condition Goldsworthy will use to provide the estimate of the future JPY/GBP spot rate is *most likely*:

- **A** covered interest rate parity.

B uncovered interest rate parity.

C relative purchasing power parity.

SOLUTIONS

1 C is correct. The equation to calculate the forward premium (discount) is:

$$F_{f/d} - S_{f/d} = S_{f/d}\left(\frac{\left[\frac{Actual}{360}\right]}{1 + i_d\left[\frac{Actual}{360}\right]}\right)\left(i_f - i_d\right)$$

$S_{f/d}$ is the spot rate with GBP the base currency or *d*, and INR the foreign currency or *f*. $S_{f/d}$ per Exhibit 1 is 79.5093, i_f is equal to 7.52% and i_d is equal to 5.43%.

With GBP as the base currency (i.e. the "domestic" currency) in the INR/GBP quote, substituting in the relevant base currency values from Exhibit 1 yields the following:

$$F_{f/d} - S_{f/d} = 79.5093\left(\frac{\left[\frac{360}{360}\right]}{1 + 0.0543\left[\frac{360}{360}\right]}\right)\left(0.0752 - 0.0543\right)$$

$$F_{f/d} - S_{f/d} = 79.5093\left(\frac{1}{1.0543}\right)\left(0.0752 - 0.0543\right)$$

$$F_{f/d} - S_{f/d} = 1.576$$

2 B is correct. The dealer is posting a bid rate to buy BRL at a price that is too high. This overpricing is determined by calculating the interbank implied cross rate for the CHF/BRL using the intuitive equation-based approach:

CHF/BRL = CHF/USD × (BRL/USD)$^{-1}$, or

CHF/BRL = CHF/USD × USD/BRL

Inverting the BRL/USD given quotes in Exhibit 2 determines the USD/BRL bid/offer rates of 0.56205/0.56211 (The bid of 0.56205 is the inverse of the BRL/USD offer, calculated as 1/1.7792; the offer of 0.56211 is the inverse of the BRL/USD bid, calculated as 1/1.7790). Multiplying the CHF/USD and USD/BRL bid/offer rates then leads to the interbank implied CHF/BRL cross rate of:

Bid: 0.9099 × 0.56205 = 0.5114

Offer: 0.9101 × 0.56211 = 0.5116

Since the dealer is willing to buy BRL at 0.5161 but BRL can be purchased from the interbank market at 0.5116, so there is an arbitrage opportunity to buy BRL in the interbank market and sell them to the dealer for a profit of 0.0045 CHF (0.5161 – 0.5116) per BRL transacted.

3 A is correct. The carry trade involves borrowing in a lower yielding currency to invest in a higher yielding one and netting any profit after allowing for borrowing costs and exchange rate movements. The relevant trade is to borrow USD and lend in Euros. To calculate the all-in USD return from a one-year EUR Libor deposit, first determine the current and one-year later USD/EUR exchange rates. Because one USD buys CAD 1.0055 today, and one CAD buys EUR 0.7218 today, today's EUR/USD rate is the product of these two numbers:

1.0055 × 0.7218 = 0.7258. The projected rate one year later is: 1.0006 × 0.7279 = 0.7283. Accordingly, measured in dollars, the investment return for the unhedged EUR Libor deposit is equal to:

$$(1.0055 \times 0.7218) \times (1 + 0.022) \times [1/(1.0006 \times 0.7279)] -1$$
$$= 0.7258 \times (1.022)(1/0.7283) -1 = 1.0184 - 1 = 1.84\%$$

However, the borrowing costs must be charged against this *gross* return to fund the carry trade investment (one-year USD Libor was 0.80%). The *net* return on the carry trade is thereby closest to: 1.84% – 0.80% = 1.04%.

4 B is correct. While credit ratings can affect spreads, the trade involves spot settlement, i.e. two business days after the trade date, so the spread quoted to this highly rated (AA) firm is not likely to be much tighter than the spread that would be quoted to a somewhat lower rated (but still high quality) firm. The relationship between the bank and client, the size of the trade, the time of day the trade is initiated, the currencies involved and the level of market volatility are likely to be more significant factors in determining the spread for this trade.

5 B is correct. By rearranging the terms of the equation defining covered interest rate parity, and assuming that uncovered interest rate parity is in effect, the forward exchange rate is equal to the expected future spot exchange rate, $F_{f/d} = S^e_{f/d}$, with the expected percentage change in the spot rate equal to the interest rate differential. Thus, the forward exchange rate is an unbiased forecast of the future spot exchange rate.

6 C is correct. The carry trade strategy is dependent upon the fact that uncovered interest rate parity does not hold in the short or medium term. If uncovered interest rate parity held, it would mean that investors would receive identical returns from either an unhedged foreign currency investment or a domestic currency investment because the appreciation/depreciation of the exchange rate would offset the yield differential. However, during periods of low volatility, evidence shows that high yield currencies do not depreciate enough and low yield currencies do not appreciate enough to offset the yield differential.

7 C is correct. McFadden states that exchange rates will *immediately* correct the trade imbalance. She is describing the Flow Supply/Demand Channel, which assumes that trade imbalances will be corrected as the deficit country's currency depreciates, causing its exports to become more competitive and its imports to become more expensive. Studies indicate that there can be long lags between exchange rate changes, changes in the prices of traded goods and changes in the trade balance. In the short-run, exchange rates tend to be more responsive to investment and financing decisions.

8 C is correct. Risk premiums are more closely associated with the portfolio-balance approach. The portfolio balance approach addresses the impact of a country's net foreign asset/liability position. Under the portfolio balance approach, investors are assumed to hold a diversified portfolio of assets including foreign and domestic bonds. Investors will hold a country's bonds as long as they are compensated appropriately. Compensation may come in the form of higher interest rates and/or higher risk premiums.

9 B is correct. The currency is likely to appreciate. The emerging market country has both a restrictive monetary policy and restrictive fiscal policy under conditions of low capital mobility. Low capital mobility indicates that interest rate changes induced by monetary and fiscal policy will not cause large changes in capital flows. Implementation of restrictive policies should result in an improvement in the trade balance, which will result in currency appreciation.

10 C is correct. Expansionary fiscal policies result in currency depreciation in the long run. Under a portfolio-balance approach, the assumption is that investors hold a mix of domestic and foreign assets including bonds. Fiscal stimulus policies result in budget deficits which are often financed by debt. As the debt level rises, investors become concerned as to how the on-going deficit will be financed. The country's central bank may need to create more money in order to purchase the debt which would cause the currency to depreciate. Or, the government could adopt a more restrictive fiscal policy, which would also depreciate the currency.

11 A is correct. EM countries are better able to influence their exchange rates because their reserve levels as a ratio to average daily FX turnover are generally much greater than those of DM countries. This means that EM central banks are in a better position to affect currency supply and demand than DM countries where the ratio is negligible. EM policymakers use their foreign exchange reserves as a kind of insurance to defend their currencies, as needed.

12 A is correct. Prediction 1 is least likely to be correct. Foreign exchange reserves tend to decline precipitously, not increase, as a currency crisis approaches. Broad money growth tends to rise in the period leading up to a currency crisis and the exchange rate is substantially higher than its mean level during tranquil periods.

13 A is correct. Goldsworthy has been given a bid–offer spread. Because she is buying the base currency—in this case, GBP—she must pay the offer price of JPY 187.43 per GBP.

$$\frac{\text{JPY } 225{,}000{,}000}{187.43 \text{ JPY/GBP}} = \text{GBP } 1{,}200{,}448$$

14 A is correct. Posted quotes are typically for transactions in 1 million units of the base currency. Larger transactions may be harder for the dealer to sell in the interbank market and would likely require the dealer to quote a wider spread (lower bid price and higher offer price).

15 A is correct. Using quotes from Dealer A, she can find

$$\frac{\text{MXN}}{\text{GBP}} = \frac{\text{MXN}}{\text{USD}} \times \frac{\text{USD}}{\text{GBP}}$$

The bid from Dealer A for MXN/GBP is effectively

$$\left(\frac{\text{MXN}}{\text{GBP}}\right)_{bid} = \left(\frac{\text{MXN}}{\text{USD}}\right)_{bid} \times \left(\frac{\text{USD}}{\text{GBP}}\right)_{bid}$$
$$= 17.147 \times 1.5762 = 27.0271$$

The offer from Dealer A is

$$\left(\frac{\text{MXN}}{\text{GBP}}\right)_{offer} = \left(\frac{\text{MXN}}{\text{USD}}\right)_{offer} \times \left(\frac{\text{USD}}{\text{GBP}}\right)_{offer}$$
$$= 17.330 \times 1.5766 = 27.3225$$

To compare with Dealer B's quote, she must take the inverse of MXN/GBP, so that she has an offer to sell MXN at a rate of 1/27.0271 = GBP 0.0370 and a bid to purchase MXN at a rate of 1/27.3225 = GBP 0.0366. Dealer A is effectively quoting MXN/GBP at 0.0366/0.0370. Although she can effectively buy pesos more cheaply from Dealer A (GBP 0.0370 from Dealer A, versus GBP 0.0372 from Dealer B), she cannot resell them to Dealer B for a higher price than GBP 0.0366. There is no profit from triangular arbitrage.

16 A is correct. Marking her nine-month contract to market six months later requires buying GBP/EUR three months forward. The GBP/EUR spot rate is 0.7342/0.7344, and the three-month forward points are 14.0/15.0. The three-month forward rate to use is 0.7344 + (15/10000) = 0.7359. Goldsworthy sold EUR 5,000,000 at 0.7400 and bought at 0.7359. The net cash flow at the settlement date will equal EUR 5,000,000 × (0.7400 – 0.7359) GBP/EUR = GBP 20,500. This cash flow will occur in three months, so we discount at the three-month GBP Libor rate of 58 bps:

$$\frac{\text{GBP } 20{,}500}{1 + 0.0058\left[\frac{90}{360}\right]} = \text{GBP } 20{,}470.32$$

17 A is correct. The positive forward points for the GBP/EUR pair shown in Exhibit 2 indicates that the EUR trades at a forward premium at all maturities, including three months. Covered interest rate parity

$$F_{f/d} = S_{f/d}\left(\frac{1 + i_f\left[\frac{\text{Actual}}{360}\right]}{1 + i_d\left[\frac{\text{Actual}}{360}\right]}\right)$$

suggests a forward rate greater than the spot rate requires a non-domestic risk-free rate (in this case, the GBP Libor) greater than the domestic risk-free rate (EUR Libor). When covered interest rate parity is violated, traders can step in and conduct arbitrage.

18 B is correct. Using covered interest rate parity, the forward rate is

$$F_{f/d} = S_{f/d}\left(\frac{1 + i_f\left[\frac{\text{Actual}}{360}\right]}{1 + i_d\left[\frac{\text{Actual}}{360}\right]}\right)$$

$$= 1.5764\left(\frac{1 + 0.0033\left[\frac{90}{360}\right]}{1 + 0.0058\left[\frac{90}{360}\right]}\right) = 1.5754$$

Because the domestic rate (Libor) is higher than the non-domestic rate, the forward rate will be less than the spot, giving a forward discount of

$$F_{f/d} - S_{f/d} = 1.5754 - 1.5764 = -0.0010$$

19 B is correct. The covered interest rate parity condition

$$F_{f/d} = S_{f/d}\left(\frac{1 + i_f\left[\frac{\text{Actual}}{360}\right]}{1 + i_d\left[\frac{\text{Actual}}{360}\right]}\right) \quad \text{(Equation 1)}$$

specifies the forward exchange rate that must hold to prevent arbitrage given the spot exchange rate and the risk-free rates in both countries. If the forward and spot exchange rates are known, as well as one of the risk-free rates, the other risk-free rate can be calculated.

20 B is correct. According to uncovered interest rate parity

$$\%\Delta S^e_{f/d} = i_f - i_d \quad \text{(Equation 2)}$$

the expected change in the spot exchange rate should reflect the interest rate spread between the two countries, which can be found in Exhibit 3. Given the spot exchange rate (from Exhibit 1) and the expected future change, she should be able to estimate the future spot exchange rate.

READING

11

Economic Growth and the Investment Decision

by Paul R. Kutasovic, PhD, CFA

Paul R. Kutasovic, PhD, CFA, is at New York Institute of Technology (USA).

LEARNING OUTCOMES

Mastery	The candidate should be able to:
☐	**a.** compare factors favoring and limiting economic growth in developed and developing economies;
☐	**b.** describe the relation between the long-run rate of stock market appreciation and the sustainable growth rate of the economy;
☐	**c.** explain why potential GDP and its growth rate matter for equity and fixed income investors;
☐	**d.** distinguish between capital deepening investment and technological progress and explain how each affects economic growth and labor productivity;
☐	**e.** forecast potential GDP based on growth accounting relations;
☐	**f.** explain how natural resources affect economic growth and evaluate the argument that limited availability of natural resources constrains economic growth;
☐	**g.** explain how demographics, immigration, and labor force participation affect the rate and sustainability of economic growth;
☐	**h.** explain how investment in physical capital, human capital, and technological development affects economic growth;
☐	**i.** compare classical growth theory, neoclassical growth theory, and endogenous growth theory;
☐	**j.** explain and evaluate convergence hypotheses;
☐	**k.** describe the economic rationale for governments to provide incentives to private investment in technology and knowledge;
☐	**l.** describe the expected impact of removing trade barriers on capital investment and profits, employment and wages, and growth in the economies involved.

1 INTRODUCTION

Forecasts of long-run economic growth are important for global investors. Equity prices reflect expectations of the future stream of earnings, which depend on expectations of future economic activity. This means that in the long term, the same factors that drive economic growth will be reflected in equity values. Similarly, the expected long-run growth rate of real income is a key determinant of the average real interest rate level in the economy, and therefore the level of real returns in general. In the shorter term, the relationship between actual and potential growth (i.e., the degree of slack in the economy) is a key driver of fixed income returns. Therefore, in order to develop global portfolio strategies and investment return expectations, investors must be able to identify and forecast the factors that drive long-term sustainable growth trends. Based on a country's long-term economic outlook, investors can then evaluate the long-term investment potential and risk of investing in the securities of companies located or operating in that country.

In contrast to the short-run fluctuations of the business cycle, the study of economic growth focuses on the long-run trend in aggregate output as measured by potential GDP. Over long periods of time, the actual growth rate of GDP should equal the rate of increase in potential GDP because, by definition, output in excess of potential GDP requires employing labor and capital beyond their optimum levels. Thus, the growth rate of potential GDP acts as an upper limit to growth and determines the economy's sustainable rate of growth. Increasing the growth rate of potential GDP is the key to raising the level of income, the level of profits, and the living standard of the population. Even small differences in the growth rate translate into large differences in the level of income over time.

What drives long-run growth? What distinguishes the "winners" from the "losers" in the long-run growth arena? Will poor countries catch up with rich countries over time? Can policies have a permanent effect on the sustainable growth rate? If so, how? If not, why not? These and other key questions are addressed in detail in this reading.

The reading is organized as follows: Section 2 examines the long-term growth record, focusing on the extent of growth variation across countries and across decades. Section 3 discusses the importance of economic growth to global investors and examines the relationship between investment returns and economic growth. Section 4 examines the factors that determine long-run economic growth. Section 5 presents the classical, neoclassical, and endogenous growth models. It also discusses whether poorer countries are converging to the higher income levels of the richer countries. Finally, Section 6 looks at the impact of international trade on economic growth. A summary and practice problems complete the reading.

2 GROWTH IN THE GLOBAL ECONOMY: DEVELOPED VS. DEVELOPING ECONOMIES

The first step in our study of long-term growth is to compare the economic performance of countries. GDP and per capita GDP are the best indicators economists have for measuring a country's standard of living and its level of economic development. Economic growth is calculated as the annual percentage change in real GDP or in real per capita GDP. Growth in real GDP measures how rapidly the total economy is expanding. Real per capita GDP reflects the average standard of living in each country—essentially the average level of material well-being. Growth in real GDP per capita (i.e., real GDP growing faster than the population) implies a rising standard of living.

Exhibit 1 presents data on the level of per capita GDP and the growth rate of GDP for various economies. Because each economy reports its data in its own currency, each one's data must be converted into a common currency, usually the US dollar. One can convert the GDP data into dollars using either current market exchange rates or the exchange rates implied by **purchasing power parity (PPP)**. Purchasing power parity is the idea that exchange rates move to equalize the purchasing power of different currencies. At the exchange rates implied by PPP, the cost of a typical basket of goods and services is the same across all economies. In other words, exchange rates should be at a level where you can buy the same goods and services with the equivalent amount of any economy's currency.

Exhibit 1 Divergent Real GDP Growth among Selected Economies

	Average Annual Real GDP Growth (percent)				Real GDP Per Capita in Dollars[a]			
	1971–1980	1981–1990	1991–2000	2001–2010	1950	1970	1990	2010
Advanced Economies	3.2	3.1	2.8	1.6				
Canada	4.0	2.8	2.4	1.8	$12,053	$19,919	$31,969	$41,288
United States	3.1	2.9	3.4	1.6	14,559	22,806	35,328	46,697
France	2.9	2.4	1.7	1.2	8,266	18,186	28,127	34,358
Germany	2.7	2.3	2.3	0.9	na	na	28,624	37,367
Ireland	4.7	3.9	7.1	2.6	5,496	9,869	18,812	36,433
Italy	3.4	2.2	1.7	0.3	5,954	16,522	27,734	31,069
Spain	3.0	3.0	2.9	2.1	3,964	11,444	21,830	30,504
United Kingdom	1.6	2.7	2.8	1.5	11,602	18,002	27,469	37,378
Japan	4.3	4.0	1.3	0.8	3,048	15,413	29,813	34,828
Singapore	10.5	7.3	7.3	5.6	4,299	8,600	27,550	56,224
South Korea	7.4	9.1	7.2	4.2	1,185	3,009	12,083	30,079
Australia	3.2	3.3	3.4	3.0	13,219	21,444	30,628	45,951
New Zealand	1.6	2.5	2.9	2.3	13,795	18,255	22,331	31,223
Developing Economies	4.3	4.2	5.4	6.3				
Developing Asia	6.2	6.9	7.4	8.5				
China	10.4	9.1	10.4	10.5	402	698	1,677	8,569
India	3.9	5.9	5.6	7.5	658	922	1,390	3,575
Indonesia	8.4	5.4	4.0	5.2	804	1,182	2,517	4,740
Pakistan	4.5	6.0	3.9	4.8	666	985	1,645	2,600
Philippines	6.6	1.7	3.0	4.7	1,296	2,136	2,660	3,672
Vietnam	4.7	5.9	7.6	7.3	689	770	1,073	3,369
Middle East	2.9	3.0	4.0	4.9				
Egypt	5.9	5.9	4.4	4.9	1,132	1,560	3,137	5,306
Turkey	4.1	5.2	3.6	4.0	2,327	4,413	7,741	11,769
Saudi Arabia	11.0	1.7	2.7	3.3	5,060	17,292	20,399	22,951
Latin America	6.5	1.6	3.3	3.4				
Argentina	2.9	–1.2	4.2	4.6	6,164	9,026	7,952	13,468
Brazil	8.8	1.5	2.5	3.6	2,365	4,324	6,959	9,589

(continued)

Exhibit 1 (Continued)

	Average Annual Real GDP Growth (percent)				Real GDP Per Capita in Dollars[a]			
	1971–1980	1981–1990	1991–2000	2001–2010	1950	1970	1990	2010
Mexico	6.6	1.8	3.5	1.8	4,180	7,634	10,754	13,710
Peru	7.6	–0.8	4.0	5.7	3,464	5,786	4,516	8,671
Venezuela	1.6	1.9	2.1	3.5	8,104	11,590	9,028	10,560
Africa	3.5	2.5	2.4	5.7				
Botswana	17.1	10.9	6.4	4.2	449	774	3,731	5,311
Ethiopia	3.0	1.9	2.9	8.4	314	479	462	749
Kenya	7.4	4.3	1.7	4.1	791	1,113	1,359	1,376
Nigeria	7.4	2.0	1.9	8.7	814	1,183	1,203	2,037
South Africa	4.1	1.5	1.8	3.5	4,361	6,959	6,595	8,716

[a] The measure of GDP per capita is in constant US dollar market prices for 2010 and adjusted for cross-economy differences in the relative prices of goods and services using purchasing power parity (PPP).
Sources: International Monetary Fund, World Economic Outlook database for growth rates, and Conference Board, Total Economy Database (September 2011).

In general, the simple method of taking a country's GDP measured in its own currency and then multiplying by the current exchange rate to express it in another currency is not appropriate. Using market exchange rates has two problems. First, market exchange rates are very volatile. Changes in the exchange rate could result in large swings in measured GDP even if there is little or no growth in the country's economy. Second, market exchange rates are determined by financial flows and flows in tradable goods and services. This ignores the fact that much of global consumption is for non-tradable goods and services. Prices of non-traded goods and services differ by country. In particular, non-traded goods are generally less expensive in developing countries than in developed countries. For example, because labor is cheaper in Mexico City than in London, the prices of labor-intensive products, such as haircuts or taxi rides, are lower in Mexico City than in London. Failing to account for differences in the prices of non-traded goods and services across countries tends to understate the standard of living of consumers in developing countries. To compare standards of living across time or across countries, we need to use a common set of prices among a wide range of goods and services. Thus, cross-country comparisons of GDP should be based on purchasing power parity rather than current market exchange rates.

The economies in Exhibit 1 are divided into two categories, developed (or advanced) economies and developing economies. Developed economies are those with high per capita GDP.[1] These include the United States, Canada, Australia, Japan, and major economies in Europe. Growth in the large, developed economies generally slowed over the last few decades, with US growth exceeding that of Europe and Japan. Also included in this group are countries such as Singapore, Ireland, and Spain, which were poor in the 1950s but now have relatively high per capita real GDPs because of high rates of growth over the past 50 years.

1 There are no universally agreed upon criteria for classifying economies as advanced or developing. The International Monetary Fund (IMF) classifies 34 economies as advanced and 150 as developing. It says that "this classification is not based on strict criteria, economic or otherwise, and has evolved over time" (IMF 2011).

The second group of countries is the developing countries of Africa, Asia, and Latin America. Per capita GDP in these countries is lower than in the advanced countries, but GDP is generally growing at a faster rate than in the developed countries. Although the growth rates of the developing countries exceed those of the advanced countries, there is significant variation in economic performance among the developing countries. China and India are growing at a rapid rate. Meanwhile, growth in Latin America, Africa, and the Middle East has lagged behind Asia.

What explains the diverse experiences among the developing countries and between the developed and developing ones? Singapore, for example, had less than half the per capita GDP of the United States in 1970 but now has per capita GDP that exceeds that of the United States. In contrast, such countries as Ethiopia and Kenya have remained poor, with little growth in per capita GDP. The literature on economic growth focuses primarily on the role of capital and labor resources and the use of technology as sources of growth. In addition to these purely economic drivers, developed and developing countries differ with respect to the presence or absence of appropriate institutions that support growth. These institutions enable developing countries to raise their standards of living and eventually move into the ranks of the developed countries. We now examine some of the key institutions and requirements for growth.

2.1 Savings and Investment

One of the major problems for some of the developing countries is a low level of capital per worker. Countries accumulate capital through private and public sector (e.g., infrastructure) investment. But increasing the investment rate may be difficult in developing countries because low levels of disposable income can make it difficult to generate significant saving. The low saving rate contributes to a vicious cycle of poverty: Low savings lead to low levels of investment, which leads to slow GDP growth, which implies persistently low income and savings. Therefore, it is very difficult to design policies to increase domestic saving and investment rates in developing countries. The good news is that the savings of domestic residents are not the only source of investment funds. A developing country can break out of the cycle of low savings by attracting foreign investment.

2.2 Financial Markets and Intermediaries

In addition to the saving rate, growth depends on how efficiently saving is allocated within the economy. A role of the financial sector in any economy is to channel funds from savers to investment projects. Financial markets and intermediaries, such as banks, can promote growth in at least three ways. First, by screening those who seek funding and monitoring those who obtain funding, the financial sector channels financial capital (savings) to projects that are likely to generate the highest risk-adjusted returns. Second, the financial sector may encourage savings and assumption of risk by creating attractive investment instruments that facilitate risk transfer and diversification and enhance liquidity. Finally, the existence of well-developed financial markets and intermediaries can mitigate the credit constraints that companies might otherwise face in financing capital investments. For example, banks can aggregate small amounts of savings into a larger pool enabling them to finance larger projects that can exploit economies of scale. Evidence suggests that countries with better-functioning financial markets and intermediaries grow at a faster rate.[2] However, not all financial

2 Levine (2005).

sector developments promote economic growth. Financial sector intermediation that results in declining credit standards and/or increasing leverage will increase risk and not necessarily increase long-run growth.

2.3 Political Stability, Rule of Law, and Property Rights

Stable and effective government, a well-developed legal and regulatory system, and respect for property rights are key ingredients for economic growth. Property rights are the legal arrangements that govern the protection of private property, including intellectual property. Clearly established property rights create the incentive for domestic households and companies to invest and save. A legal system—substantive and procedural laws[3]—is needed to establish and protect these rights. In developed countries these rights and arrangements are well established, but they may be lacking or ineffective in developing countries.

In addition, economic uncertainty increases when wars, military coups, corruption, and other sources of political instability are widespread. These factors raise investment risk, discourage foreign investment, and weaken growth. In many developing countries, especially those in Africa, the first priority in trying to enhance growth is to enact a legal system that establishes, protects, and enforces property rights.

2.4 Education and Health Care Systems

Inadequate education at all levels is a major impediment to growth for many developing countries. Many workers are illiterate, and few workers have the skills needed to use the latest technology. At the same time, many developing countries also suffer from a "brain drain," where the most highly educated individuals leave the developing country for the advanced countries. Basic education raises the skill level of the workforce and thus contributes to the country's potential for growth. In addition, because physical capital and human capital are often complementary, education can raise growth by increasing the productivity of existing physical capital. Thus, improving education, through both formal schooling and on-the-job training, is an important component of a sustainable growth strategy for a developing country. China and India are investing large amounts in education and have successfully graduated large numbers of students majoring in engineering and technology-related areas of study. This effort is significantly improving the quality of their workforces.

Empirical studies show that the allocation of education spending among different types and levels (primary, secondary, and post-secondary) of education is a key determinant of growth, especially in comparing growth in the developed countries with growth in the developing ones. The impact of education spending depends on whether the country is on the leading edge of technology and fostering innovation or simply relying on imitation as a source of growth. Typically, developed countries, such as the United States, Japan and western European nations, are on the leading edge of technology and need to invest in post-secondary education to encourage innovation and growth. For these countries, incremental spending on primary and secondary education will have a smaller impact on growth. In contrast, the developing countries, which largely apply technology developed elsewhere, should emphasize primary and secondary education. Such spending will improve growth by improving the countries' ability to absorb new technologies and to organize existing tasks more efficiently.

---

3 Substantive law focuses on the rights and responsibilities of entities and relationships among entities, and procedural law focuses on the protection and enforcement of the substantive laws.

Poor health is another obstacle to growth in the developing countries. Life expectancy rates are substantially lower in many developing countries. In Africa, tropical diseases are rampant and AIDS has had a devastating impact. As is evident in Exhibit 1, the growth rate of GDP in Botswana, a huge success story in the 1970s and 1980s, has slowed dramatically over the last two decades due, at least in part, to the AIDS epidemic.

2.5 Tax and Regulatory Systems

Tax and regulatory policies have an important impact on growth and productivity, especially at the company level. Analysis suggests that limited regulations encourage entrepreneurial activity and the entry of new companies. There is also a strong positive correlation between the entry of new companies and average productivity levels. Studies by the Organisation for Economic Co-Operation and Development (OECD) indicate that low administrative start-up cost is a key factor encouraging entrepreneurship.[4]

2.6 Free Trade and Unrestricted Capital Flows

Opening an economy to capital and trade flows has a major impact on economic growth. In an open economy, world savings can finance domestic investment. As a potential source of funds, foreign investment can break the vicious cycle of low income, low domestic savings, and low investment. Foreign investment can occur in two ways:

- Foreign companies can invest directly in a domestic economy (so-called foreign direct investment, or FDI) by building or buying property, plant, and equipment.
- Foreign companies and individuals can invest indirectly in a domestic economy by purchasing securities (equity and fixed income) issued by domestic companies.

Both of these forms of foreign investment will potentially increase the developing economy's physical capital stock, leading to higher productivity, employment and wages, and perhaps even increased domestic savings. This suggests that developing countries would benefit from policies that encourage investment from abroad, such as eliminating high tariffs on foreign imports (especially capital goods) and removing restrictions on foreign direct and indirect investments.

Brazil and India are examples of developing countries that have benefited from foreign investment. Foreign companies directly invested \$48.5 billion in Brazil in 2010, an important source of investment spending for the Brazilian economy (see Exhibit 19 in Section 6). Foreign direct investment also provides developing countries with access to advanced technology developed and used in the advanced countries. In 1999, India enacted new regulations that liberalized direct and indirect foreign investments in Indian companies. Foreign institutional and venture capital investors were given greater flexibility to invest directly in Indian entities as well as in the Indian capital markets. These changes also made it easier for foreign companies to invest in plant and equipment. These developments contributed to the acceleration in India's economic growth over the last decade (see Exhibit 1).

Capital flows are just one way that the international economy affects economic growth. The other is through trade in goods and services. In general, free trade benefits an economy by providing its residents with more goods at lower costs. Domestic

4 OECD (2003).

companies face increased competition, which limits their price discretion, but they also obtain access to larger markets. The evidence of the benefits of open markets is discussed later in the reading.

2.7 Summary of Factors Limiting Growth in Developing Countries

Developing countries differ significantly from developed countries in terms of their institutional structure and their legal and political environments. Lack of appropriate institutions and poor legal and political environments restrain growth in the developing economies and partially explain why these countries are poor and experience slow growth. Factors limiting growth include the following:

- Low rates of saving and investment
- Poorly developed financial markets
- Weak, or even corrupt, legal systems and failure to enforce laws
- Lack of property rights and political instability
- Poor public education and health services
- Tax and regulatory polices discouraging entrepreneurship
- Restrictions on international trade and flows of capital

Although these factors are not necessarily absent in developed countries, they tend to be more prevalent in developing countries. Policies that correct these issues, or mitigate their impact, enhance the potential for growth. In addition to these institutional restraints, as we will see in Section 4, growth in developing countries may be limited by a lack of physical, human, and public capital, as well as little or no innovation.

EXAMPLE 1

Why Growth Rates Matter[a]

In 1950, Argentina and Venezuela were relatively wealthy countries with per capita levels of GDP of $6,164 and $8,104, respectively. Per capita GDPs in these Latin American countries were well above those of Japan, South Korea, and Singapore, which had per capita GDPs of $3,048, $1,185, and $4,299, respectively. By 2010, however, a dramatic change occurred in the relative GDPs per capita of these countries.

Real GDP Per Capita in Dollars

	Venezuela	Argentina	Singapore	Japan	South Korea
1950	$8,104	$6,164	$4,299	$3,048	$1,185
2010	$10,560	$13,468	$56,224	$34,828	$30,079

1 Calculate the annual growth rate in per capita GDP for each of the five countries over the period 1950–2010.

2 Explain the implication of the growth rates for these countries.

3 Suppose that GDP per capita in Argentina had grown at the same rate as in Japan from 1950 to 2010. How much larger would real per capita GDP have been in Argentina in 2010?

4 Venezuela plans to stimulate growth in its economy by substantially increasing spending on infrastructure, education, and health care. Nevertheless, foreign investment is discouraged, and reforms like strengthening the legal system and encouraging private ownership have been largely ignored. Explain whether the measures described above could lead to faster economic growth.

Solution to 1:

The annual growth rates for the five countries are calculated as follows:

Argentina	$[(\$13{,}468/\$6{,}164)^{1/60}] - 1 = 1.31\%$
Venezuela	$[(\$10{,}560/\$8{,}104)^{1/60}] - 1 = 0.44\%$
Japan	$[(\$34{,}828/\$3{,}048)^{1/60}] - 1 = 4.14\%$
Singapore	$[(\$56{,}224/\$4{,}299)^{1/60}] - 1 = 4.38\%$
South Korea	$[(\$30{,}079/\$1{,}185)^{1/60}] - 1 = 5.54\%$

Solution to 2:

Differences in GDP growth rates sustained over a number of decades will significantly alter the relative incomes of countries. Nations that experience sustained periods of high growth will eventually become high-income countries and move up the income ladder. In contrast, countries with slow growth will experience relative declines in living standards. This is well illustrated in this example by a historic comparison of growth in Argentina and Venezuela with Japan, Singapore, and South Korea. In 1950, Argentina and Venezuela were relatively wealthy countries with per capita levels of GDP well above that of Japan, South Korea, and Singapore. Over the next 60 years, however, the rate of growth in per capita GDP was significantly slower in Venezuela and Argentina in comparison to the three Asian countries. This resulted in a dramatic change in the relative incomes of these countries. The per capita GDP of the three Asian countries rose sharply as each joined the ranks of developed countries. In contrast, Argentina and Venezuela stagnated and moved from the ranks of developed countries to developing country status. By 2010, per capita income in Singapore was more than five times higher than in Venezuela.

Over the long run, the rate of economic growth is an extremely important variable. Even small differences in growth rates matter because of the power of compounding. Thus, policy actions that affect the long-term growth rate even by a small amount will have a major economic impact.

Solution to 3:

Assuming Argentina had grown at the same rate as Japan since 1950, its GDP per capita in 2010 would have been $(\$6{,}164)(1+0.0414)^{60} = (\$6{,}164)(11.404) =$ \$70,294, versus \$13,468 from Exhibit 1.

If Argentina had grown at the same rate as Japan, it would have had by far the highest standard of living in the world in 2010. The question is why the growth rates in Argentina and Venezuela diverged so much from the three Asian countries.

Solution to 4:

The preconditions for economic growth are well-functioning financial markets, clearly defined property rights and rule of law, open international trade and flows of capital, an educated and healthy population, and tax and regulatory policies that encourage entrepreneurship. Investment in infrastructure would increase Venezuela's stock of physical capital, which would raise labor productivity and growth. Better education and health care would increase human capital and also increase productivity and growth. These measures would raise the growth prospects for Venezuela. However, what is missing is a legal system that could better enforce property rights, openness to international trade and foreign investment, and well-functioning capital markets. Without changes in these preconditions, a significant improvement in growth is unlikely to occur. The preconditions are summarized below:

Preconditions for Growth:	Impact of Planned Policy Action in Venezuela:
Saving and investment	Improve growth potential
Developed financial markets	No impact
Legal systems	No impact
Property rights and political stability	No impact
Education and health	Improve growth potential
Tax and regulatory polices discouraging entrepreneurship	No impact
Restrictions on international trade and flows of capital	No impact

[a] It should be noted that the global economy is evolving rapidly and past trends may or may not be sustained. Nonetheless, in order to provide concrete answers that do not require the reader to bring in additional information, our exercise solutions must assume past patterns are indicative of the future.

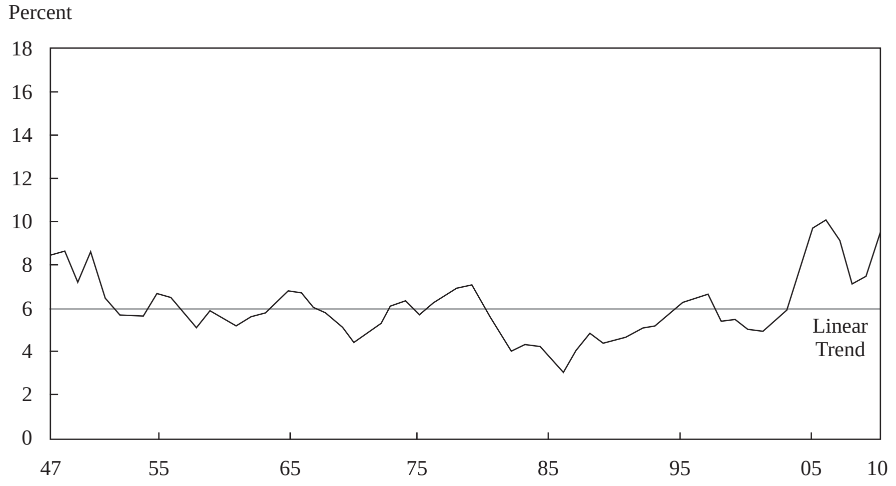
3 WHY POTENTIAL GROWTH MATTERS TO INVESTORS

The valuations of both equity and fixed income securities are closely related to the growth rate of economic activity. Anticipated growth in aggregate earnings is a fundamental driver of the equity market. Growth in an economy's productive capacity, measured by **potential GDP**, places a limit on how fast the economy can grow. The idea is that potential GDP is the maximum amount of output an economy can sustainably produce without inducing an increase in the inflation rate. A key question for equity investors, therefore, is whether earnings growth is also bounded or limited by the growth rate of potential GDP.

For earnings growth to exceed GDP growth, the ratio of corporate profits to GDP must trend upward over time. It should be clear that the share of profits in GDP cannot rise forever. At some point, stagnant labor income would make workers unwilling to work and would also undermine demand, making further profit growth unsustainable. Thus, in the long run, real earnings growth cannot exceed the growth rate of potential

GDP.[5] Exhibit 2 illustrates the long-run stability of after-tax profits as a share of GDP using US data derived from the National Income and Product Accounts (NIPA). The chart shows that since 1947, after-tax profits have ranged between 3.1 percent and 10.1 percent of GDP and have averaged around 6 percent of GDP. Note that there is neither an upward trend in the ratio of after-tax profits to GDP nor a move to a permanent increase in the ratio. The share of profits in 1947, at 8.5 percent, was essentially equal to the 9.4 percent share at the end of the period in 2010. Because there is no trend in the ratio, the same factors that limit economic growth also set the upper limit or bound on the long-run growth of aggregate earnings.

Exhibit 2 US After-Tax Corporate Profits as a Percentage of GDP



Source: FRED database, Federal Reserve Bank of St. Louis.

To examine the relationship between economic growth and stock prices, it is useful to express the aggregate value of the stock market as the product of key ratios. Letting P represent the aggregate value (price) of equities and E represent aggregate earnings, we can write:

$$P = GDP\left(\frac{E}{GDP}\right)\left(\frac{P}{E}\right)$$

This equation represents the aggregate value of equities as the product of GDP, corporate earnings as a share of GDP, and the price-to-earnings ratio for the market. Note that GDP may be interpreted as either real or nominal with a corresponding real or nominal interpretation of the other variables.

This equation can be expressed in terms of logarithmic rates of change over a time horizon T:

$$(1/T)\%\Delta P = (1/T)\%\Delta GDP + (1/T)\%\Delta(E/GDP) + (1/\mathrm{T})\%\Delta(P/E)$$

5 Earnings growth for the overall national economy can differ from the growth of earnings per share in a country's equity market composites. This is due to the presence of new businesses that are not yet included in the equity indexes and are typically growing at a faster rate than the mature companies that make up the composites. Thus, the earnings growth rate of companies making up the composites should be lower than the earnings growth rate for the overall economy.

Thus, the percentage change in stock market value equals the percentage change in GDP plus the percentage change in the share of earnings (profit) in GDP plus the percentage change in the price-to-earnings multiple.[6] Over short to immediate horizons, all three of these factors contribute to appreciation or depreciation of the stock market. In the long run, however, the growth rate of GDP must dominate. As noted above, the ratio of earnings to GDP cannot rise forever. It cannot decline forever either because unprofitable businesses will disappear. Hence, the second term in the equation above must be approximately zero over long horizons (T). Similarly, the price-to-earnings ratio cannot grow or contract forever because investors will not pay an arbitrarily large price for a unit of earnings, nor will they give away earnings for nothing. Hence, the third term must also be approximately zero over long horizons. The conclusion is that the drivers of potential GDP are ultimately the drivers of stock market price performance.

Exhibit 3 shows the close relationship between economic growth and equity market appreciation over long horizons. Over the period 1946–2007, the S&P 500 Index returned 10.82 percent per year, of which 7.15 percent per year came from price appreciation. The price appreciation was almost exactly equal to the 6.95 percent growth rate of US nominal GDP (real GDP growth plus inflation). Changes in the earnings-to-GDP and price-to-earnings ratios contributed only a combined 0.20 percent per year. As shown in the last column of the exhibit, these two ratios contributed much more to the volatility of the market than to its return.[7]

Exhibit 3 Decomposition of S&P 500 Returns: Log Returns, 1946–2007

	Annual Return/Growth Rate	Standard Deviation
S&P 500 return	10.82%	15.31%
Real GDP growth	3.01	2.97
Inflation	3.94	3.29
EPS/GDP	–0.12	17.62
P/E	0.32	23.80
Dividend yield	3.67	1.49
Total	10.82	

Source: Stewart, Piros, and Heisler (2011).

Estimates of potential GDP and its growth rate are widely available. For example, both the OECD and the International Monetary Fund (IMF) provide such estimates as a basis for their intermediate-term and long-term forecasts of economic growth by country. In addition, central banks regularly make projections of potential GDP. The methods used to estimate potential GDP are examined later in the reading. The data in Exhibit 1 illustrate that simply extrapolating past GDP growth into the future may

6 For simplicity, we have not explicitly incorporated issuance or repurchasing of shares. To do so, we would simply need to distinguish between aggregate stock market value and price per share. However, this would not alter our conclusions. Similarly, we could incorporate the dividend payout ratio into our argument, but again, this would not alter our conclusions.

7 It should be noted that the 1946–2007 time period was chosen because both endpoints correspond to fairly normal economic and market conditions. Selecting endpoints that correspond to crisis or bubble conditions would distort the role played by the various components of return.

produce an incorrect forecast. A country's GDP growth rate can and does change over time. GDP growth can either slow down, as was the case for Japan (compare 1971–1990 with 1991–2010), or speed up, as was the case for Brazil over the last decade. Factors or policies that cause potential growth to increase or decrease by even a small amount will have a large impact on living standards and the future level of economic activity. The effect is analogous to the rate of return on a portfolio, where small differences in return compounded over many years result in a substantially higher or lower value for the portfolio. Being able to recognize these changes is critical for the global investor.

Estimates of an economy's growth potential are also relevant for global fixed income investors. One of the uses of potential GDP is to gauge inflationary pressures in the economy. Actual GDP growth above (below) the potential growth rate puts upward (downward) pressure on inflation, which puts corresponding pressure on nominal interest rates and bond prices.[8]

The growth rate of potential GDP is also an important determinant of the level of real interest rates, and therefore real asset returns in general, in the economy. The real interest rate is essentially the real return that consumers/savers demand in exchange for postponing consumption. Faster growth in potential GDP means that consumers expect their real income to rise more rapidly. This implies that an extra unit of future income/consumption is less valuable than it would be if income were expected to grow more slowly. Hence, all else equal, the real interest rate will have to be higher in order to induce the savings needed to fund required capital accumulation. Thus, higher rates of potential GDP growth translate into higher real interest rates and higher expected real asset returns in general.

Potential GDP and its growth rate enter into fixed income analysis in other ways as well. Among them are the following:

1 A higher rate of potential GDP growth improves the general credit quality of fixed income securities because most such securities are ultimately backed by a flow of income even if the lender has a claim on specific underlying assets.

2 Central banks frequently explain their monetary policy decisions by referring to the level of "resource utilization" and the degree of "slack in the economy." In other words, monetary policy decisions are affected by the difference between an economy's estimated potential output and its actual operating level (referred to as the output gap) and by growth of actual GDP relative to the sustainable growth rate. Thus, fixed income investors need to closely monitor the output gap and growth rates of actual and potential GDP to assess the likelihood of a change in central bank policy.

3 Credit rating agencies use the growth rate of potential GDP as an input in evaluating the credit risk of sovereign debt or government-issued debt. All else equal, slower estimated potential GDP growth raises the perceived risk of these bonds.

4 Government budget deficits typically increase during recessions and decrease during expansions. In examining fiscal policy, actual fiscal positions are often judged relative to structural or cyclically adjusted deficits—the budgetary balance that would exist if the economy were operating at potential GDP.

8 Note that this is an argument about cyclical variations in growth and inflation around the economy's long-term potential growth rate. It does not imply that there is a long-run trade-off between growth and inflation.

EXAMPLE 2

Impact on Equity and Fixed Income Investors

Your firm subscribes to asset class risk and return estimates generated by a large pension consultant. The equity market return estimates are based primarily on long-term average index returns. Following a multi-year period of very high equity returns driven by unusually high earnings growth and expanding P/E multiples, capital's share of total income as well as valuation multiples are near all-time highs. Based on the latest data, the vendor projects that your domestic equity market will return 13.5 percent per year—11 percent annual appreciation and 2.5 percent dividend yield—forever.

Your firm also subscribes to a macroeconomic forecasting service that provides, in addition to shorter-term projections, estimates of the long-term growth rate of potential GDP and the long-term inflation rate. This service forecasts 3.25 percent real growth in the future and 3.75 percent inflation, down from 4.0 percent and 5.0 percent, respectively, over the last 75 years.

1 Why might you have greater confidence in the macroeconomic service's forecasts than in the pension consultant's equity market return forecast?
2 Assuming the macroeconomic forecasts are accurate, what implicit assumptions underlie the pension consultant's forecast of 11 percent equity market appreciation?
3 Assuming the macroeconomic forecasts are accurate, what would be a more reasonable forecast for long-term equity returns?
4 In addition to its long-term potential GDP forecast, the macroeconomic forecasting service estimates sluggish 1.5 percent GDP growth for the next year. Based on this short-term GDP forecast, the bond analyst at your firm recommends that the firm increase its fixed income investments. What assumptions underlie the bond analyst's forecast?

Solution to 1:

High volatility makes equity returns very hard to predict based on their own history. As illustrated in Exhibit 3, the high volatility of equity returns is due to the underlying volatility of earnings as a share of GDP and valuation ratios. Long-term real GDP growth rates tend to be far less volatile, especially for developed economies, such as the United States or the euro area, because long-term potential growth is governed by fundamental economic forces that tend to evolve slowly over time. Similarly, for countries with prudent monetary policies, inflation rates are much less volatile than stock prices. Thus, one could reasonably place much higher confidence in forecasts of long-term real and nominal (real growth plus inflation) GDP growth than in equity market return forecasts based on historical equity returns.

Solution to 2:

We can decompose the equity market appreciation rate into components due to (a) nominal GDP growth, (b) expansion/contraction of the share of profits in GDP, and (c) expansion/contraction of the P/E. The macroeconomic forecast indicates that nominal GDP will grow at 7 percent (3.25% real + 3.75% inflation). So the pension consultant's forecast of 11 percent equity market appreciation implies a 4 percent per year combined contribution from expansion in the P/E multiple and/or the profit share of GDP—*forever*.

Solution to 3:

Neither the P/E nor the profit share of GDP can grow at a non-negligible rate forever. A much more reasonable forecast of long-term equity market appreciation would be the projected 7 percent growth rate of nominal GDP.

Solution to 4:

With forecasted actual GDP growth well below the growth in potential GDP, the bond analyst assumes a growing output gap or slack in the economy. This slack may place downward pressure on inflation and reduce inflationary expectations. To close this gap, the central bank may need to lower short-term interest rates and ease policy. In such an environment, bond prices should rise.

DETERMINANTS OF ECONOMIC GROWTH

4

What are the forces driving long-run economic growth? The following sections discuss labor, physical and human capital, technology, and other factors, such as natural resources and public infrastructure, as inputs to economic growth and production functions and how changes in such inputs affect growth. Section 4.1 begins the discussion by presenting one of the simplest useful models of the production function.

4.1 Production Function

A production function is a model of the quantitative link between the inputs (factors of production), technology, and output. A two-factor aggregate production function with labor and capital as the inputs can be represented as:

$$Y = AF(K,L) \tag{1}$$

where Y denotes the level of aggregate output in the economy, L is the quantity of labor or number of workers or hours worked in the economy, and K is an estimate of the capital services provided by the stock of equipment and structures used to produce goods and services. The function $F(\ )$ embodies the fact that capital and labor can be used in various combinations to produce output.

In the production function above, A is a multiplicative scale factor referred to as **total factor productivity (TFP)**. Note that an increase in TFP implies a proportionate increase in output for any combination of inputs. Hence, TFP reflects the general level of productivity or technology in the economy. The state of technology embodies the cumulative effects of scientific advances, applied research and development, improvements in management methods, and ways of organizing production that raise the productive capacity of factories and offices.

It is worth noting that both the function $F(\ )$ and the scale factor A reflect technology. An innovation that makes it possible to produce the same output with the same amount of capital but fewer workers would be reflected in a change in the function $F(\ )$ because the relative productivity of labor and capital has been altered. In contrast, an increase in TFP does not affect the relative productivity of the inputs. As is standard in the analysis of economic growth, *unless stated otherwise, the level of "technology" should be interpreted as referring to TFP.*

In order to obtain concrete results, it is useful to use a specific functional form for the production function. The **Cobb–Douglas production function**, given by

$$F(K,L) = K^{\alpha}L^{1-\alpha} \tag{2}$$

is widely used because it is easy to analyze and does a good job of fitting the historic data relating inputs and output. The parameter α determines the shares of output (factor shares) paid by companies to capital and labor and is assumed to have a value between 0 and 1. The reason for this follows from basic microeconomics. In a competitive economy, factors of production are paid their marginal product. Profit maximization requires that the marginal product of capital equal the **rental price of capital** and the marginal product of labor equal the (real) wage rate. In the case of capital, the marginal product of capital (MPK) for the Cobb–Douglas production function is[9]

$$\text{MPK} = \alpha A K^{\alpha-1} L^{1-\alpha} = \alpha Y/K$$

Setting the MPK equal to the rental price (r) of capital,

$$\alpha Y/K = r$$

If we solve this equation for α, we find that it equals the ratio of capital income, rK to output or GDP, Y. Thus, α *is the share of GDP paid out to the suppliers of capital.* A similar calculation shows that 1 – α is the share of income paid to labor. This result is important because it is easy to estimate α for an economy by simply looking at capital's share of income in the national income accounts.

The Cobb–Douglas production function exhibits two important properties that explain the relationship between the inputs and the output. First, the Cobb–Douglas production function exhibits **constant returns to scale**. This means that if all the inputs into the production process are increased by the same percentage, then output rises by that percentage. Under the assumption of constant returns to scale, we can modify the production function (Equation 1) and examine the determinants of the quantity of output per worker. Multiplying the production function by $1/L$ gives

$$Y/L = AF(K/L, L/L) = AF(K/L, 1)$$

Defining $y = Y/L$ as the output per worker or (average) **labor productivity** and $k = K/L$ as the capital-to-labor ratio, the above expression becomes

$$y = AF(k, 1)$$

Specifying the Cobb–Douglas production function in output per worker terms, where again lower case letters denote variables measured on a per capita basis, we get

$$y = Y/L = A(K/L)^{\alpha}(L/L)^{1-\alpha} = Ak^{\alpha} \quad (3)$$

This equation tells us that the amount of goods a worker can produce (labor productivity) depends on the amount of capital available for each worker (capital-to-labor ratio), technology or TFP, and the share of capital in GDP (α). It is important to note that there are two different measures of productivity or efficiency in this equation. Labor productivity measures the output produced by a unit of labor and is measured by dividing the output (GDP) by the labor input used to produce that output ($y = Y/L$). TFP is a scale factor that multiplies the impact of the capital and labor inputs. Changes in TFP are estimated using a growth accounting method discussed in the next section.

A second important property of the model is the relation between an individual input and the level of output produced. The Cobb–Douglas production function exhibits **diminishing marginal productivity** with respect to each individual input. Marginal productivity is the extra output produced from a one-unit increase in an input keeping the other inputs unchanged. It applies to any input as long as the other inputs are held constant. For example, if we have a factory of a fixed size and we add more workers to the factory, the marginal productivity of labor measures how much additional output each additional worker will produce. Diminishing marginal

9 The marginal product of capital is simply the derivative of output with respect to capital. This can be approximated as $\Delta Y/\Delta K \approx [A(K + \Delta K)^{\alpha} L^{1-\alpha} - AK^{\alpha} L^{1-\alpha}]/\Delta K \approx [A\alpha K^{\alpha-1}\Delta K L^{1-\alpha}]/\Delta K = A\alpha K^{\alpha-1}L^{1-\alpha} = \alpha Y/K$. The approximation becomes exact for very small increments, ΔK.

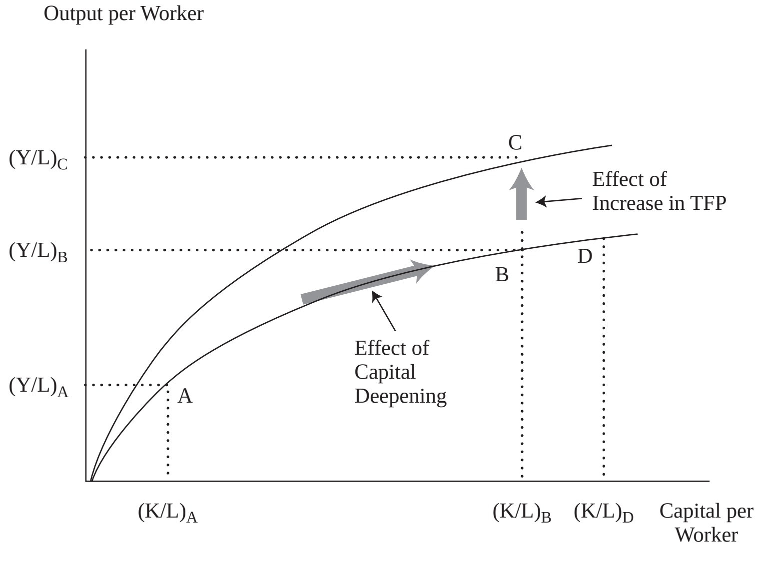
productivity means that at some point the extra output obtained from each additional unit of the input will decline. To continue our example, if we hire more workers at the existing factory (fixed capital input in this case) each additional worker adds less to output than the previously hired worker does and average labor productivity (y) falls.

The significance of diminishing marginal returns in the Cobb–Douglas production function depends on the value of α. *A value of α close to zero means diminishing marginal returns to capital are very significant and the extra output made possible by additional capital declines quickly as capital increases.* In contrast, a value of α close to one means that the next unit of capital increases output almost as much as the previous unit of capital. In this case, diminishing marginal returns still occur but the impact is relatively small. Note that the exponents on the K and L variables in the Cobb–Douglas production function sum to one, indicating constant returns to scale—that is, there are no diminishing marginal returns if both inputs are increased proportionately.

4.2 Capital Deepening vs. Technological Progress

The property of diminishing marginal returns plays an important role in assessing the contribution of capital and technology to economic growth. Exhibit 4 shows the relationship between per capita output and the capital-to-labor ratio. It shows that adding more and more capital to a fixed number of workers increases per capita output but at a decreasing rate. Looking at Equation 3 and Exhibit 4, we can think of growth in per capita output coming from two sources: capital deepening and an improvement in technology, often referred to as technological progress.

Exhibit 4 Per Capita Production Function Capital Deepening vs. Technological (TFP) Progress



Capital deepening, an increase in the capital-to-labor ratio, is reflected in the exhibit by a move along the production function from point A to B. The increase in the capital-to-labor ratio reflects rising investment in the economy. The ratio will increase as long as the growth rate of capital (net investment) exceeds the growth rate of labor. However, once the capital-to-labor ratio becomes very high, as at point

B, further additions to capital have relatively little impact on per capita output (e.g., moving to point D). This occurs because the marginal product of capital declines as more capital is added to the labor input.

At the point where the marginal product of capital equals its marginal cost, profit maximizing producers will stop adding capital (i.e., stop increasing the capital-to-labor ratio).[10] As we will discuss in Section 5, this point is very significant in the neoclassical model of growth because per capita growth in the economy will come to a halt. Once the economy reaches this steady state, capital deepening cannot be a source of sustained growth in the economy. Only when the economy is operating below the steady state and when the marginal product of capital exceeds its marginal cost can capital deepening raise per capita growth.

The neoclassical model's stark implication that more rapid capital accumulation—that is, higher rates of investment—cannot result in a permanently higher rate of per capita growth is somewhat disappointing. As we will see in our discussion of endogenous growth, capital accumulation can result in a permanently higher growth rate if the investment results not just in *more* capital (i.e., pure capital deepening) but also in new, innovative products and processes. That is, if the additional capital embodies new, more efficient methods of production or previously unavailable products, then more rapid capital accumulation can result in a permanently higher growth rate of per capita output.

In contrast to moves along a given production function, an improvement in TFP causes a proportional upward shift in the entire production function. As a result, the economy can produce higher output per worker for a given level of capital per worker. This is shown in Exhibit 4 by the move from B to C. Technological progress also increases the marginal product of capital relative to its marginal cost. This makes additional capital investments profitable and tends to mitigate the limits imposed on growth by diminishing marginal returns. In addition, continued growth in per capita output is possible even in the steady state as long as there is ongoing technological progress (increases in TFP). In summary, *sustained growth in per capita output requires progress in TFP.*

EXAMPLE 3

Capital Deepening vs. Technological Progress

One of main differences between developed and developing countries is the amount of capital available for each worker. Country A is an advanced economy with $100,000 of capital available for each worker and thus a high capital-to-labor ratio. In contrast, Country B is a developing country with only $5,000 of capital available for each worker. What impact will the following developments have on the growth rate of potential GDP?

1 An increase in business investment in both countries

2 An increase in the amount of spending on university research in both countries

3 An elimination of restrictions in Country B on the inflow of foreign investment

10 To avoid confusion later, we must note that once technological progress (TFP growth) is introduced, the capital-to-labor ratio will have to keep increasing just to keep the marginal productivity of capital equal to its marginal cost. But the point remains: Once that equality is attained, companies will not increase the capital-to-labor ratio faster than is necessary to maintain that equality.

Solution to 1:

An increase in business investment will raise the capital-to-labor ratio in both countries. It results in capital deepening and a movement along the per worker production function. However, the impact on growth will be significantly different for the two countries. Country B will experience an increase in output per worker and thus in the growth rate of potential GDP. This is because Country B operates at a low level of capital per worker, at a point like A in Exhibit 4. Diminishing returns to capital are small, so any addition to capital has a major impact on growth. Country A operates at a point like B in Exhibit 4, so additions to capital have little impact on growth because of diminishing returns.

Solution to 2:

An increase in spending on university research will increase TFP and cause an upward shift in the production function in both countries. This can be seen in the move from point B to point C in Exhibit 4. The shift in the production function will raise growth in both countries and offset the negative impact of diminishing returns. This result shows that developing countries have the potential to grow through both capital deepening and technological progress, whereas improvement in potential GDP growth in developed countries is largely driven by technological progress.

Solution to 3:

The elimination of restrictions will result in higher foreign investment, which has the same impact as an increase in domestic business investment. This is again a movement along the production function such as from point A to B in Exhibit 4. With diminishing returns insignificant at low levels of capital to labor, the higher level of foreign investment will boost growth of potential GDP in Country B.

4.3 Growth Accounting

Since the publication of Solow's seminal work in 1957,[11] growth accounting has been used to analyze the performance of economies. The growth accounting equation is essentially the production function written in the form of growth rates. It starts with the Cobb–Douglas production function and decomposes the percentage change in output into components attributable to capital, labor, and technology:

$$\Delta Y/Y = \Delta A/A + \alpha\Delta K/K + (1 - \alpha)\Delta L/L \qquad (4)$$

The **growth accounting equation** states that the growth rate of output equals the rate of technological change plus α times the growth rate of capital plus (1 – α) times the growth rate of labor. Because a 1 percent increase in capital leads to an α% increase in output, α is the elasticity of output with respect to capital. Similarly, (1 – α) is the elasticity of output with respect to labor. Thus, in the Cobb–Douglas production function, the exponents α and (1 – α) play dual roles as both output elasticities and the shares of income paid to each factor. Note that the impact of any unspecified inputs (e.g., natural resources) is subsumed into the TFP component.

Data on output, capital, labor, and the elasticities of capital and labor are available for most developed countries. The rate of technological change is not directly measured and must therefore be estimated. The elasticities of capital and labor in the growth accounting equation are the relative shares of capital (α) and labor (1 – α) in national income and are estimated from the GDP accounts. For the United States, the relative shares of labor and capital are approximately 0.7 and 0.3, respectively. This means that

11 See Solow (1957).

an increase in the growth rate of labor will have a significantly larger impact—roughly double—on potential GDP growth than will an equivalent increase in the growth rate of capital, holding all else equal. For example, because capital's share in GDP in the US economy is 0.3, a 1 percent increase in the amount of capital available for each worker increases output by only 0.3 percent. An equivalent increase in the labor input would boost growth by 0.7 percent.

The growth accounting equation has a number of uses in studying an economy. First, Solow used the equation to estimate the contribution of technological progress to economic growth. Solow estimated the growth in TFP as a residual in the above equation by plugging in $\Delta Y/Y$, $\Delta K/K$, $\Delta L/L$, and α and solving for $\Delta A/A$. This residual measures the amount of output that cannot be explained by growth in capital or labor and can thus be regarded as progress in TFP.

Second, the growth accounting equation is used to empirically measure the sources of growth in an economy. In such studies, the growth accounting equation is used to quantify the contribution of each factor to long-term growth in an economy and answer such questions as the following: How important are labor and demographic factors to growth? What is the contribution of capital, and how important is capital deepening as a source of growth? What is the impact of TFP? The growth accounting equation can be expanded by considering different forms of capital and labor inputs, such as human capital and knowledge capital, and by considering the quality of the inputs as well.

Finally, the growth accounting equation is used to measure potential output. Potential GDP is estimated using Equation 4 with trend estimates of labor and capital and α estimated as one minus the labor share of GDP. The difficult task is estimating the growth rate of TFP, which, by definition, is a residual in the growth accounting equation.[12] The standard methodology treats TFP as exogenous and estimates its growth rate using various time-series models.

An alternative method of measuring potential GDP is the **labor productivity growth accounting equation**. It is very similar to the Solow approach but is simpler and models potential GDP as a function of the labor input and the productivity of the labor input. It avoids the need to estimate the capital input and the difficulty associated with computing total factor productivity. The disadvantage is that it incorporates both capital deepening and TFP progress in the productivity term in a way that can be difficult to analyze and to predict over long periods of time. Under this approach, the equation for estimating potential GDP is

$$\begin{aligned}\text{Growth rate in potential GDP} &= \text{Long-term growth rate of labor force}\\ &\quad + \text{Long-term growth rate in labor productivity}\end{aligned} \tag{5}$$

Thus, potential GDP growth is a combination of the long-term growth rate of the labor force and the long-term growth rate of labor productivity. If the labor force is growing at 1 percent per year and productivity per worker is rising at 2 percent per year, then potential GDP is rising at 3 percent per year.

12 TFP is computed as the growth in output less the growth in the factor inputs. These would include labor and capital in the traditional Solow two-factor production model. If the production function is expanded by including more inputs, the weighted growth rates of these inputs would also be subtracted from the growth in output.

4.4 Extending the Production Function

As a simplification, the production function in Equation 1 focused on only the labor and capital inputs. A more complete specification of the production function expands the list of inputs to include the following:

- Raw materials: natural resources such as oil, lumber, and available land (N)
- Quantity of labor: the number of workers in the country (L)
- Human capital: education and skill level of these workers (H)
- Information, computer, and telecommunications (ICT) capital: computer hardware, software, and communication equipment (K_{IT})
- Non-ICT capital: transport equipment, metal products and plant machinery other than computer hardware and communications equipment, and non-residential buildings and other structures (K_{NT})
- Public capital: infrastructure owned and provided by the government (K_P)
- Technological knowledge: the production methods used to convert inputs into final products, reflected by total factor productivity (A)

The expanded production function is expressed mathematically as:

$$Y = AF(N,L,H,K_{IT},K_{NT},K_P)$$

The impact of each of these inputs on economic growth is addressed in the following sections.

4.5 Natural Resources

Raw materials, including everything from available land to oil to water, are an essential input to growth. There are two categories of natural resources:

1. **Renewable resources** are those that are replenished, such as a forest. For example, if a tree is cut, a seedling can be planted and a new forest harvested in the future.
2. **Non-renewable resources** are finite resources that are depleted once they are consumed. Oil and coal are examples.

Although it seems intuitive that countries with more natural resources will be wealthier, the relation between resource endowment and growth is not so straightforward. Natural resources do account for some of the differences in growth among countries. Today, Middle Eastern countries and such countries as Brazil and Australia have relatively high per capita incomes because of their resource base. Countries in the Middle East have large pools of oil. Brazil has an abundance of land suitable for large-scale agricultural production, allowing it to be a major exporter of coffee, soybeans, and beef.

Even though *access* to natural resources (e.g., via trade) is important, *ownership and production of natural resources is not necessary for a country to achieve a high level of income*. Countries in East Asia, such as Japan and South Korea, have experienced rapid economic growth but have few natural resources. In contrast, both Venezuela and Saudi Arabia have large oil reserves and are major producers of oil, yet both countries have experienced subpar growth in comparison to the natural-resource-poor countries of Singapore, Japan, and South Korea. As was examined in Example 1, economic growth in Venezuela over the last 60 years was well below that of Singapore, Japan, and South Korea.

For some countries, the presence of natural resources may even restrain growth, resulting in a "resource curse." Venezuela and Nigeria are two examples of countries blessed with resources yet with sluggish economic growth. There are two main reasons why this may occur. First, countries rich in natural resources may fail to develop the economic institutions necessary for growth. Second, countries rich in resources may suffer the **Dutch disease**, where currency appreciation driven by strong export demand for resources makes other segments of the economy, in particular manufacturing, globally uncompetitive.[13] In this situation, the manufacturing sector contracts and the country does not participate in the TFP progress that occurs in countries with more vigorous manufacturing sectors.

In contrast, there is a long-standing concern that non-renewable natural resources will eventually limit growth. The idea is that a combination of rapid economic growth and a fixed stock of resources will cause resource depletion as the available pool of resources is used up. These concerns are probably overstated. Technological progress (TFP from all sources) enables the economy to use fewer resources per unit of output and to develop substitutes. The growing scarcity of specific resources will increase their price and encourage a shift toward more plentiful substitutes. Finally, the share of national income going to land and resources has been declining for most countries, especially as the composition of output in the global economy shifts toward the use of more services.

EXAMPLE 4

Impact of Natural Resources

The table below shows the share of world proved oil reserves as of 1990 for each of the 34 economies shown in Exhibit 1, along with the growth rate of real per capita GDP from 1990 to 2010. The simple correlation between the share of oil reserves and subsequent growth is not statistically different from zero.

	Percent of World Proved Oil Reserves: 1990	Real Per Capita GDP Growth (%) 1990–2010		Percent of World Proved Oil Reserves: 1990	Real Per Capita GDP Growth (%) 1990–2010
Saudi Arabia	25.75	3.00	Germany	0.04	1.60
Venezuela	5.85	2.80	France	0.02	1.45
Mexico	5.64	2.65	New Zealand	0.01	2.60
United States	2.62	2.50	Pakistan	0.01	4.35
China	2.40	10.45	Japan	0.01	1.05
Nigeria	1.60	5.25	Spain	0.00	2.50
Indonesia	0.82	4.60	Philippines	0.00	3.85
India	0.75	6.55	Botswana	0.00	5.29
Canada	0.61	2.10	Ethiopia	0.00	5.61
Egypt	0.45	4.65	Ireland	0.00	4.83
United Kingdom	0.43	2.15	Kenya	0.00	2.89
Brazil	0.28	3.05	Singapore	0.00	6.45

13 Following the discovery of large natural gas fields in the Netherlands, the Dutch guilder appreciated and the manufacturing sector contracted.

	Percent of World Proved Oil Reserves: 1990	Real Per Capita GDP Growth (%) 1990–2010		Percent of World Proved Oil Reserves: 1990	Real Per Capita GDP Growth (%) 1990–2010
Argentina	0.23	4.40	South Africa	0.00	2.65
Australia	0.17	3.20	South Korea	0.00	5.69%
Italy	0.07	1.00	Vietnam	0.00	7.45%
Turkey	0.05	3.80			
Peru	0.04	4.85			

Sources: US Energy Information Administration (www.eia.gov) and Exhibit 1.

What might account for the fact that real per capita GDP growth appears to be unrelated to oil reserves, perhaps the single most economically important natural resource (aside from water)?

Solution:

Energy is a vital input for any economy. Thus, *access* to energy resources is critical. *Ownership* of raw energy resources, however, is not. Countries that are not self-sufficient in oil or other resources acquire what they need through trade. It should be noted that countries that lack oil may possess other types of energy resources, such as natural gas, coal, hydropower, or geothermal energy. In addition, countries can grow by emphasizing less energy intensive products, especially services, and adopting more energy efficient production methods. In sum, natural resources are important but not necessary for growth.

4.6 Labor Supply

As noted above, economic growth is affected by increases in inputs, mainly labor and capital. Growth in the number of people available for work (quantity of workforce) is an important source of economic growth and partially accounts for the superior growth performance of the United States among the advanced economies—in particular, relative to Europe and Japan. Most developing countries, such as China, India, and Mexico, have a large potential labor supply. We can measure the potential size of the labor input as the total number of hours available for work. This, in turn, equals the labor force times the average hours worked per worker. The **labor force** is defined as the working age population (ages 16 to 64) that is either employed or available for work but not working (i.e., unemployed). Thus, growth in the labor input depends on four factors: population growth, labor force participation, net migration, and average hours worked.

4.6.1 *Population Growth*

Long-term projections of the labor supply are largely determined by the growth of the working age population. Population growth is determined by fertility rates and mortality rates. Population growth rates are significantly lower in the developed countries than in the developing countries. As a result, there is an ongoing decline in the developed countries' share of the world's population. Note that although population growth may increase the growth rate of the overall economy, it has no impact on the rate of increase in *per capita* GDP.

The age mix of the population is also important. The percentage of the population over the age of 65 and the percentage below the age of 16 are key considerations. Some of the developed countries, especially European countries, Japan, and South Korea, are facing a growing demographic burden as the portion of non-working elders (over 65) grows as a share of the population. In contrast, growth in many developing countries will receive a demographic boost as the fraction of the population below the age of 16 begins to decline. Interestingly, China is similar to the advanced economies, with a growing proportion of the population over age 65.

Exhibit 5 Population Data for Selected Countries (millions)

	2000	2005	2010	Annual Growth (%) 2000–2010
France	59.1	61.2	63.0	0.64
Germany	82.2	82.3	81.6	–0.07
Ireland	3.8	4.0	4.5	1.71
Spain	40.3	43.4	46.1	1.35
United Kingdom	58.9	59.4	61.3	0.40
Russia	146.7	142.8	142.9	–0.26
Japan	126.9	127.8	127.6	0.06
United States	282.2	295.6	309.1	0.91
Mexico	98.4	103.9	108.4	0.97
China	1,267.4	1,307.6	1,341.4	0.57
India	1,024.3	1,110.0	1,190.5	1.52

Source: OECD Stat Extracts.

4.6.2 *Labor Force Participation*

In the short run, the growth rate of the labor force may differ from population growth because of changes in the participation rate. The **labor force participation rate** is defined as the percentage of the working age population in the labor force. It has trended upward in most countries over the last few decades because of rising participation rates among women. In contrast to population, an increase in the participation rate may raise the growth of per capita GDP. In many southern European countries, such as Greece and Italy, the participation rate among women is well below the rates in the United States and northern European countries (see Exhibit 6). Thus, rising participation rates among women in these countries could increase growth in the labor force and in potential GDP. This has been the case for Spain, where the female labor force participation rate rose from 52.0 percent in 2000 to 66.1 percent in 2010. It should be noted, however, that rising or falling labor force participation is likely to represent a transition to a new higher or lower level of participation rather than a truly permanent rate of change. Thus, although trends in participation may contribute to or detract from potential growth for substantial periods, one should be cautious in extrapolating such trends indefinitely.

Exhibit 6 Labor Force Data for Selected Countries (2010)

	Percent of Population under Age 15	Percent of Population over Age 65	Participation Rate: Male (%)	Participation Rate: Female (%)
France	18.3	16.7	72.6	67.3
Germany	13.1	21.0	82.3	71.5
Greece	14.3	18.6	76.5	55.0
Ireland	21.5	11.4	79.5	63.4
Italy	14.1	20.0	73.5	51.5
Spain	15.0	17.0	80.4	66.1
Sweden	16.6	18.3	84.5	77.8
UK	17.7	15.9	82.8	70.5
Japan	13.3	22.7	93.8	68.5
US	20.1	13.1	79.2	69.8
Mexico	28.1	5.9	83.2	48.3
Turkey	26.4	7.0	73.9	28.8

Source: OECD Stat Extracts.

EXAMPLE 5

Impact of the Age Distribution on Growth: Mexico vs. Germany

Exhibits 5 and 6 provide population data for selected countries. The data show that the rate of population growth and the age composition vary significantly among countries. Thus, demographic factors can be expected to have a significant impact on relative growth rates across countries. This is very clear in the cases of Mexico and Germany. There was essentially zero growth in the population of Germany from 2000 to 2010, while the population of Mexico increased by 0.97 percent annually. The age composition of the two countries is also very different. How will the age distribution impact growth over the next decade?

Solution:

What is important for growth is the number of workers available to enter the workforce. Over the next decade, Mexico will receive a demographic benefit because of the high percentage of young people entering the workforce. This is because 28.1 percent of the population in 2010 was below the age of 15. In contrast, only 13.1 percent of the German population was below the age of 15. In addition, Germany is facing a demographic challenge given the high and growing share of its population over the age of 65. In Mexico, only 5.9 percent of the population is above the age of 65, compared with 21 percent in Germany. In sum, the lack of population growth and a rapidly aging population in Germany will limit its potential rate of growth. Germany must rely on high labor productivity growth, increase its workforce participation rate, or encourage immigration if it is to increase its near-term potential rate of growth. Meanwhile, potential GDP growth in Mexico should receive a boost from its favorable population trends.

4.6.3 *Net Migration*

Another factor increasing economic and population growth, especially among the developed countries, is immigration. Heightened immigration is a possible solution to the slowing labor force growth being experienced by many developed countries with low birthrates within the native population. The growth rate of the labor force in Ireland, Spain, the United Kingdom, and the United States has increased over the last decade because of immigration. As Exhibit 5 shows, the population growth rates for Ireland and Spain for the period 2000–2010, at 1.71 percent and 1.35 percent, respectively, were well above the population growth rates in other European countries. As shown in Exhibit 7, this is due to the impact of immigration. The open-border policies of both countries led to a significant population of immigrants that contributed to a large increase in labor input for both countries. As a consequence, both countries experienced GDP growth above the European average during this period (see Exhibit 1).

Exhibit 7 Ireland and Spain: Net Migration

	2000–2007	2008	2009	2010	Total 2000–2010
Ireland	357,085	38,502	–7,800	–12,200	375,587
Spain	4,222,813	460,221	181,073	111,249	4,975,356

Source: OECD Stat Extracts.

EXAMPLE 6

Potential Growth in Spain: Labor Input

The Investment Policy Committee of Global Invest Inc. reviewed a report on the growth prospects for Spain and noted that, with total hours worked growing at a 1.2 percent annual rate between 2000 and 2010, labor input had been a major source of growth for the economy. As of 2011, some members expected the growth rate of labor to slow considerably given projection from the OECD and IMF that immigration into Spain will fall to essentially zero over the next few years. A research assistant at the firm gathered demographic data on Spain from Exhibits 5–7 and other sources. The data are presented in the following table:

	2000	2010	Annual Growth (2000–2010)
Population (millions)	40.3	46.1	1.35%
Immigration since 2000 (millions)		4.975	
Percent of population under 15		15.0	
Percent of population over 65		17.0	
Male participation rate		80.4%	
Female participation rate		66.1%	
Unemployment rate		20.1%	

Using this information for Spain and Exhibits 5 and 6 for relevant comparison data, determine the following:

1 Whether a change in the trend growth rate of the labor input is likely over the next few years.
2 How the high unemployment rate of 20.1 percent is likely to affect the growth rate of the labor force.

Solution to 1:

The growth in the labor input depends on a number of factors, including the population growth rate, the labor force participation rate, and the percentage of the population below the age of 15. The labor force in Spain expanded sharply between 2000 and 2010 mainly because of a large 5.8 million person increase in the population, going from 40.3 million in 2000 to 46.1 million in 2010. Looking ahead, growth in the labor force is set to slow substantially for a number of reasons:

- The population increase between 2000 and 2010 is very misleading and not likely to be repeated in the future. Between 2000 and 2010, immigration raised the population of Spain by nearly 5 million people. Without the immigrants, the population of Spain between 2000 and 2010 would have grown by only about 825,000 or at an annual rate of 0.2 percent. With immigration, the population growth rate was 1.35 percent. The pace of immigration that occurred between 2000 and 2010 is not sustainable and is likely to slow. This will result in slower growth in the population and the labor force.
- In the short run, the growth rate of the labor force may differ from population growth because of changes in the participation rate. Looking at the data, the male participation rate in Spain, at 80.4 percent, is very high and, as shown in Exhibit 6, is above the male participation rates in France, Greece, and Italy and slightly below that of Germany. The female participation rate is low in comparison to northern European countries, such as Sweden. But it is higher than in Greece and Italy, which are probably a better comparison. Thus, little increase is likely in the male or female participation rates.
- Only 15 percent of the Spanish population is below the age of 15. The comparable figure from Exhibit 6 for the United Kingdom is 17.7 percent, for France 18.3 percent, for the United States 20.1 percent, and for Mexico 28.1 percent. Thus, Spain does not appear poised for a notable surge in young adults entering the labor force.

In summary, growth in the labor input in Spain should slow over the next few years and the growth rate of potential GDP should do the same.

Solution to 2:

Reducing the unemployment rate would mitigate some of the negative demographic factors because a reduction in the number of unemployed workers would boost utilization of the existing labor supply. This would represent a transition to a higher level of employment rather than a permanent increase in the potential growth rate. Nonetheless, it could boost potential growth for a substantial period.

4.6.4 *Average Hours Worked*

The contribution of labor to overall output is also affected by changes in the average hours worked per worker. Average hours worked is highly sensitive to the business cycle. However, the long-term trend in average hours worked has been toward a shorter workweek in the advanced countries. This development is the result of legislation, collective bargaining agreements, the growth of part-time and temporary work, and the impact of both the "wealth effect" and high tax rates on labor income, which cause workers in high-income countries to value leisure time relatively more highly than labor income.

Exhibit 8 provides data on average hours worked per year per person in the labor force for selected years since 1995. For most countries, the average number of hours worked per year has been declining. There is also a significant difference in hours worked across countries. In 2010, average hours worked per year in South Korea, at 2,193 hours, were 54.5 percent more than the 1,419 average hours worked per year in Germany. The increase in female labor force participation rates may be contributing to the shorter average workweek because female workers disproportionately take on part-time, rather than full-time, jobs.

Exhibit 8 Average Hours Worked per Year per Person in Selected Countries

	1995	2000	2005	2010
France	1,651	1,591	1,559	1,594
Germany	1,534	1,473	1,435	1,419
Greece	2,123	2,121	2,081	2,109
Ireland	1,875	1,719	1,654	1,664
Italy	1,859	1,861	1,819	1,778
Spain	1,733	1,731	1,688	1,663
Sweden	1,609	1,574	1,607	1,624
UK	1,743	1,711	1,676	1,647
Japan	1,884	1,821	1,775	1,733
South Korea	2,658	2,520	2,364	2,193
Canada	1,761	1,768	1,738	1,702
US	1,840	1,832	1,795	1,778
Mexico	1,857	1,888	1,909	1,866
Turkey	1,876	1,937	1,918	1,877

Source: OECD Stat Extracts.

4.7 Labor Quality: Human Capital

In addition to the quantity of labor, the quality of the labor force is an important source of growth for an economy. **Human capital** is the accumulated knowledge and skills that workers acquire from education, training, or life experience. In general, better-educated and more skilled workers will be more productive and more adaptable to changes in technology or other shifts in market demand and supply.

An economy's human capital is increased through investment in education and on-the-job training. Like physical capital, investment in education is costly, but studies show that there is a significant return on that investment. That is, people with more education earn higher wages. In addition, education may also have a spillover or externality impact. Increasing the educational level of one person raises not only the output of that person but also the output of those around that person. The spillover effect operates through the link between education and advances in technology. Education not only improves the quality of the labor force, and thus the stock of human capital, but also encourages growth through innovation. Importantly, increased education, obtained both formally and via on-the-job training, could result in a permanent increase in the growth rate of an economy if the more educated workforce results in more innovations and a faster rate of technological progress. Investment in the health of the population is also a major contributor to human capital, especially in the developing countries.

4.8 Capital: ICT and Non-ICT

The physical capital stock increases from year to year as long as net investment (gross investment less the depreciation of the capital) is positive. Thus, countries with a higher rate of investment should have a growing physical capital stock and a higher rate of GDP growth.[14] Exhibit 9 shows the level of gross non-residential investment as a share of GDP. The exhibit shows significant variation across countries, with the investment share in the United States being low in comparison to other developed countries.

Exhibit 9 Business Investment as a Percentage of GDP

	ICT Percent of GDP			Investment Percent of GDP			
	1990	2000	2008	1990	2000	2008	2010
Developed Countries							
France	2.6	3.7	3.2	21.5	19.5	21.1	19.1
Germany	3.1	3.6	2.7	22.8	21.5	19.4	17.3
Ireland	1.1	2.3	1.2	20.8	23.9	21.7	11.0
Italy	2.4	2.8	2.1	22.0	20.3	21.1	20.2
Spain	3.4	3.6	3.7	25.3	26.2	28.8	23.0
UK	3.3	5.1	4.2	20.5	17.1	16.6	15.0
Australia	3.6	5.5	4.6	23.6	22.0	29.4	27.6
Japan	4.9	3.7	3.1	32.5	25.4	23.2	20.2
South Korea	2.4	5.1	4.8	35.7	30.6	31.2	29.1
Singapore	3.2	5.4	4.9	35.2	33.1	30.2	23.8
Canada	2.8	3.9	3.6	21.3	19.2	22.6	22.2
US	4.1	6.6	5.1	17.4	19.9	18.1	15.8
Developing Countries							
Brazil	NA	NA	NA	14.0	18.3	20.7	19.3
China	NA	NA	NA	24.9	35.1	44.0	48.2
India	NA	NA	NA	21.8	24.3	34.9	36.8
Mexico	NA	NA	NA	17.9	25.5	26.9	25.0
South Africa	NA	NA	NA	19.1	15.1	22.5	19.3

Source: OECD StatLink.

14 The impact on growth of *per capita* GDP will be somewhat smaller if the population is growing because a proportion of net investment simply provides the capital needed to maintain the capital-to-labor ratio.

The correlation between economic growth and investment is high. Countries that devote a large share of GDP to investment, such as China, India, and South Korea, have high growth rates. The fastest-growing countries in Europe over the last decade, Ireland and Spain, have the highest investment-to-GDP ratios. Countries that devote a smaller share of GDP to investment, such as Brazil and Mexico, have slower growth rates. The data show why the Chinese economy has expanded at such a rapid rate: annual GDP growth rate in excess of 10 percent over the last two decades. Investment spending in China on new factories, equipment, and infrastructure as a percentage of GDP is the highest in the world. In recent years, China devoted over 40 percent of its GDP to investment spending.

As we discussed in Section 4.2, long-term sustainable growth cannot rely on pure capital deepening. How can we reconcile this notion with the strong correlation between investment spending and economic growth across countries? First, although diminishing marginal productivity will eventually limit the impact of capital deepening, investment-driven growth may last for a considerable period of time, especially in countries that start with relatively low levels of capital per worker.

A second, and closely related, explanation is that the impact of investment spending on available capital depends on the existing physical capital stock. As with the share of GDP devoted to investment, the stock of capital available per worker varies significantly across countries. In 2000, the average US worker had $148,091 worth of capital, compared with $42,991 in Mexico and $6,270 in India.[15] The wide difference in physical capital per worker suggests that the positive impact of changes in the physical capital stock on growth is very significant in developing countries. Mexican workers have relatively little access to machinery or equipment, so adding even a little can make a big percentage difference. In developed countries, such as the United States, Japan, Germany, France, and the United Kingdom, the physical capital stock is so large that positive net investment in any given year has only a small percentage effect on the accumulated capital stock. For the developed countries, a sustained high level of investment over many years is required to have a meaningful relative impact on the physical capital stock even though the absolute size of the increase in any given year is still larger than in the developing countries.

Third, because physical capital is not really homogeneous, the composition of investment spending and the stock of physical capital matters for growth and productivity. Insights obtained from the endogenous theory of growth (discussed in Section 5) and from studies attempting to obtain a more accurate measure of TFP show that the composition of the physical capital stock is very important. These studies suggest that capital spending could be separated into two categories. The first is spending on information, computers, and telecommunications equipment (ICT investment). Capital spending on these goods is a measure of the impact of the information technology sector on economic growth. One of the key drivers of growth in the developed countries over the last decade has been the IT sector. Growth in the IT sector has been driven by technological innovation that has caused the price of key technologies, such as semiconductors, to fall dramatically. The steep decline in the price of high-technology capital goods has encouraged investment in IT at the expense of other assets.

The IT sector has grown very rapidly and has made a significant contribution to increasing the rate of economic and productivity growth. The greater use of IT equipment in various industries has resulted in **network externalities**. Computers allow people to interconnect through the internet and by e-mail, enabling them to work more productively. *The more people in the network, the greater the potential productivity gains.* The effects of the network externalities are largely captured in TFP rather than observed as a distinct, direct effect. The share of ICT investment in

15 Heston, Summers, and Aten (2009).

GDP is shown in Exhibit 9. The data show that in most countries, the IT sector is still relatively small and that between 2000 and 2008, IT spending declined as a share of GDP as the global recession disproportionately affected high-technology spending.

The other category of investment, non-ICT capital spending, includes non-residential construction, transport equipment, and machinery. High levels of capital spending for this category should eventually result in capital deepening and thus have less impact on potential GDP growth. In contrast, a growing share of ICT investments in the economy, through their externality impacts, may actually boost the growth rate of potential GDP.[16]

4.9 Technology

The most important factor affecting growth of per capita GDP is technology, especially in developed countries. Technology allows an economy to overcome some of the limits imposed by diminishing marginal returns and results in an upward shift in the production function, as we noted in Exhibit 4. Technological progress makes it possible to produce more and/or higher-quality goods and services with the same resources or inputs. It also results in the creation of new goods and services. Technological progress can also be one of the factors improving how efficiently businesses are organized and managed.

Technological change can be embodied in human capital (knowledge, organization, information, and experience base) and/or in new machinery, equipment, and software. Therefore, high rates of investment are important, especially investment in ICT goods. Countries can also innovate through expenditures, both public and private, on research and development (R&D). Expenditures on R&D and the number of patents issued, although not directly measuring innovation, provide some useful insight into innovative performance. Exhibit 10 shows R&D spending as a share of GDP for various countries. The developed countries spend the highest percentage of GDP on R&D because they must rely on innovation and the development of new products and production methods for growth.[17] In contrast, developing countries spend less on R&D because these countries can acquire new technology through imitation or copying the technology developed elsewhere. The embodiment of technology in capital goods can enable relatively poor countries to narrow the gap relative to the technology leaders.

Exhibit 10 Research and Development as a Percentage of GDP in Selected Countries

	1990	2000	2009
France	2.3	2.2	2.2
Germany	2.6	2.5	2.8

(continued)

16 It is worthwhile to note that there have been important "transformational technologies" at various stages of history. One need only think about the impact of the steam engine, the internal combustion engine, powered flight, atomic energy, vaccination, and so on, to realize that revolutionary advances are not unique to information, computers, and telecommunications. All of these are, to some extent, "general purpose technologies" (GPT) that affect production and/or innovation in many sectors of the economy. ICT capital clearly embodies this GPT characteristic. Nanotechnology could well become the next "super GPT," at which point investing in ICT may begin to look like mere capital deepening.

17 The relationship between economic growth and R&D spending is not clear cut. Although technological innovation resulting from high R&D spending raises output and productivity in the long run, it may result in a cyclical slowing of growth as companies and workers are displaced by the new technologies. This is the Schumpeterian concept of creative destruction, which captures the double-edged nature of technological innovation.

Exhibit 10 (Continued)

	1990	2000	2009
Ireland	0.8	1.2	1.8
Italy	1.2	1.0	1.3
Spain	0.8	1.0	1.4
UK	2.1	1.8	1.9
Australia	1.3	1.5	2.2
Japan	3.0	3.0	3.4
South Korea	1.7	2.3	3.1
Singapore	1.1	1.9	2.9
Canada	1.5	1.9	2.0
US	2.6	2.7	2.9
China	NA	1.0	1.7
India	NA	0.8	0.8
Mexico	NA	0.3	0.4

Source: OECD Stat Extracts.

The state of technology, as reflected by total factor productivity, embodies the cumulative effects of scientific advances, applied research and development, improvements in management methods, and ways of organizing production that raise the productive capacity of factories and offices. Because it is measured as a residual, TFP estimates are very sensitive to the measurements of the labor and capital inputs. Empirical work at the Conference Board and the OECD accounts for changes in the composition and quality of both the labor and capital inputs. The resulting measure of TFP should capture the technological and organizational improvements that increase output for a given level of inputs. Exhibit 11 provides data for the periods 1995–2005 and 2005–2009 on the growth rate in labor productivity and total factor productivity.[18] Labor productivity growth depends on both capital deepening and technological progress. The contribution of capital deepening can be measured as the difference between the growth rates of labor productivity and total factor productivity. For example, from 2005 to 2009, Ireland's labor productivity grew by 0.8 percent per year, of which 2.9 percent [0.8% – (–2.1%)] came from capital deepening, which offset the –2.1 percent decline in TFP. The larger the difference between the productivity growth measures, the more important capital deepening is as a source of economic growth. As we discussed previously, however, growth in per capita income cannot be sustained perpetually by capital deepening.

Exhibit 11 Labor and Total Factor Productivity

	Growth in Hours Worked[a] (%)	Growth in Labor Prod. (%)	Growth in TFP (%)	Growth Due to Capital Deepening (%)	Growth in GDP (%)	Productivity Level 2010; GDP per Hour Worked ($)
Germany						53.6
1995–2005	–0.3	1.6	0.9	0.7	1.3	

18 Data for the developing countries are from 1995–2005 and 2005–2008. TFP data are not available for 2009 for these countries.

Exhibit 11 (Continued)

	Growth in Hours Worked[a] (%)	Growth in Labor Prod. (%)	Growth in TFP (%)	Growth Due to Capital Deepening (%)	Growth in GDP (%)	Productivity Level 2010; GDP per Hour Worked ($)
2005–2009	0.2	0.2	0.1	0.1	0.4	
Ireland						50.3
1995–2005	3.2	4.1	1.7	2.4	7.3	
2005–2009	–0.8	0.8	–2.1	2.9	0.0	
United States						60.3
1995–2005	0.9	2.4	0.9	1.5	3.3	
2005–2009	–0.8	1.5	–0.5	2.0	0.7	
Japan						40.7
1995–2005	–1.0	2.1	0.4	1.7	1.1	
2005–2009	–1.3	0.8	–0.6	1.4	–0.5	
South Korea						27.9
1995–2005	0.0	4.3	2.4	1.9	4.3	
2005–2009	–0.5	2.8	2.0	0.8	2.3	
China						8.6
1995–2005	1.1	6.7	1.5	5.2	7.8	
2005–2008	1.2	10.3	4.2	6.1	11.5	
India						5.3
1995–2005	2.1	4.2	1.9	2.3	6.3	
2005–2008	2.2	6.0	2.4	3.6	8.2	
Brazil						10.4
1995–2005	2.1	0.3	–0.3	0.6	2.4	
2005–2008	2.0	2.9	–0.5	3.4	4.9	
Mexico						16.8
1995–2005	2.2	1.4	0.4	1.0	3.6	
2005–2008	1.8	0.8	–0.1	0.9	2.6	

[a] Total hours worked is the preferred measure of labor quantity. However, this measure is not available for most developing countries (including China, India, Brazil, and Mexico). For these countries, total employment is used assuming that the change in total hours worked equals the change in employment. In this case, labor productivity is measured as output per worker, but for the developed countries labor productivity is output per hour.

Source: Conference Board Total Economy Database.

Exhibit 11 also provides data on the *level* of labor productivity or the amount of GDP produced per hour of work. The level of productivity depends on the accumulated stock of human and physical capital and is much higher among the developed countries. For example, China has a population of over 1.3 billion people, compared with slightly over 300 million people in the United States. Although the United States has significantly fewer workers than China because of its smaller population, its economy as measured by real GDP is much larger. This is because US workers have historically been more productive than Chinese workers as measured by GDP per hour worked, as shown in Exhibit 11. In contrast to the *level* of productivity, the *growth rate* of productivity will typically be higher in the developing countries, where human and physical capital are scarce but growing rapidly and the impact of diminishing marginal returns is relatively small.

An understanding of productivity trends is critical for global investors. A permanent increase in the rate of labor productivity growth will increase the sustainable rate of economic growth and raise the upper boundary for earnings growth and the potential return on equities. In contrast, a low growth rate of labor productivity, if it persists over a number of years, suggests poor prospects for equity prices. A slowdown in productivity growth lowers both the long-run potential growth rate of the economy and the upper limit for earnings growth. Such a development would be associated with slow growth in profits and correspondingly low equity returns.

EXAMPLE 7

Why the Sluggish Growth in the Japanese Economy?

As shown in Exhibit 1, annual growth in real GDP in Japan averaged 0.8 percent for 2000–2010 and a weak 1.3 percent in the prior decade. This growth is in sharp contrast to the 4.2 percent annual growth rate experienced from 1971 to 1990. The sluggish growth in Japan over the last decade should not be surprising. The economy of Japan is growing at its potential rate of growth which is limited by the following:

1. The labor input is not growing. Population growth has been essentially zero since 2000 (Exhibit 5), and average hours worked per year per person is declining (Exhibit 8).
2. There has been a lack of technological innovation. The lack of growth in the labor input could be offset through higher productivity derived from innovation and more efficient use of available inputs. However, this is not occurring in Japan. Total factor productivity (Exhibit 11) increased at a sluggish 0.4 percent annual rate from 1995 to 2005 and declined between 2005 and 2009.
3. Diminishing returns to capital are very significant. Despite the negative growth in TFP, labor productivity growth remained relatively high. This means that all the growth in labor productivity in Japan was due to capital deepening (Exhibit 11). The problem for Japan, as discussed in Section 4.2, is that once the capital-to-labor ratio becomes high, further additions to capital have little impact on per capita output. Thus, the growth in labor productivity should slow.

Use the data for 2005–2009 and the labor productivity growth accounting equation to estimate the growth rate in potential GDP for Japan.

Solution:

To estimate the growth rate in potential GDP, we use Equation 5, given by

Growth rate of potential GDP = Long-term growth rate of labor force + Long-term growth rate in labor productivity

To use this equation, we need to project the growth rate in the labor input and labor productivity.

The hours worked data in Exhibit 11 are a potential source to use to estimate the growth rate of the labor input. Exhibit 11 shows the labor input for Japan declining by 1.3 percent per year between 2005 and 2009. The problem here is that the decline in hours worked is overstated because of the negative impact of the global recession on hours worked. As an alternative, the labor input should grow at the same rate as the population plus the net change in immigration. The population data in Exhibit 5 show essentially zero population growth in Japan for the period 2000–2010. This trend is likely to continue. Thus, a reasonable estimate for potential GDP growth in Japan is around 0.8 percent. We get this estimate by assuming no growth in the labor input and a 0.8 percent annual increase in labor productivity (using data from Exhibit 11 for 2005–2009).

4.10 Public Infrastructure

The final expansion of the definition of the capital input is public infrastructure investment. Roads, bridges, municipal water, dams, and, in some countries, electric grids are all examples of public capital. They have few substitutes and are largely complements to the production of private sector goods and services. Ashauer (1990) found that infrastructure investment is an important source of productivity growth and should be included as an input in the production function. As with R&D spending, the full impact of government infrastructure investment may extend well beyond the direct benefits of the projects because improvements in the economy's infrastructure generally boost the productivity of private investments.

4.11 Summary

Long-term sustainable growth is determined by the rate of expansion of real potential GDP. Expansion of the supply of factors of production (inputs) and improvements in technology are the sources of growth. The factors of production include human capital, ICT and non-ICT capital, public capital, labor, and natural resources. Data for the sources of growth are available from the OECD and the Conference Board. Exhibit 12 provides data from the Conference Board on the sources of output growth for various countries. These estimates are based on the growth accounting formula.[19]

Exhibit 12 Sources of Output Growth

	Contribution from:					
	Labor Quantity (%)	Labor Quality (%)	Non-ICT Capital (%)	ICT Capital (%)	TFP (%)	Growth in GDP (%)
Germany						
1995–2005	–0.2	0.1	0.3	0.2	0.9	1.3
2005–2009	–0.6	0.1	0.5	0.3	0.1	0.4
Ireland						
1995–2005	2.0	0.3	2.6	0.7	1.7	7.3
2005–2009	–0.2	0.1	1.8	0.4	–2.1	0.0
United States						
1995–2005	0.6	0.3	0.7	0.8	0.9	3.3

(continued)

19 A standard growth accounting model (expanded version of Equation 4) is used to compute the contribution of each input to aggregate output (GDP) growth. The inputs include both the quantity and quality of labor and ICT and non-ICT capital. Each input is weighted by its share in national income, and TFP captures all sources of growth that are left unexplained by the labor and capital inputs.

Exhibit 12 (Continued)

	Contribution from:					
	Labor Quantity (%)	Labor Quality (%)	Non-ICT Capital (%)	ICT Capital (%)	TFP (%)	Growth in GDP (%)
2005–2009	0.1	0.1	0.5	0.5	–0.5	0.7
Japan						
1995–2005	–0.6	0.4	0.6	0.3	0.4	1.1
2005–2009	–0.6	0.1	0.4	0.2	–0.6	–0.5
South Korea						
1995–2005	–0.5	0.8	1.1	0.5	2.4	4.3
2005–2009	–0.7	0.0	0.8	0.2	2.0	2.3
China						
1995–2005	0.5	0.2	4.5	1.1	1.5	7.8
2005–2008	0.3	0.2	5.5	1.3	4.2	11.5
India						
1995–2005	1.0	0.2	2.7	0.5	1.9	6.3
2005–2008	1.1	0.1	3.7	0.9	2.4	8.2
Brazil						
1995–2005	0.8	0.1	1.1	0.7	–0.3	2.4
2005–2008	0.8	0.2	1.9	2.5	–0.5	4.9
Mexico						
1995–2005	1.2	0.2	1.4	0.4	0.4	3.6
2005–2008	1.1	0.1	1.3	0.2	–0.1	2.6

Source: Conference Board Total Economy Database.

EXAMPLE 8

The Irish Economy

As shown in Exhibit 1, economic growth in Ireland since 1970 has been significantly higher than that experienced in the major European economies of Germany, France, and the United Kingdom. In 1970, the per capita GDP of Ireland, at $9,869, was 45.2 percent below the per capita GDP of the United Kingdom. In 2010, per capita GDP in Ireland, at $36,433, was only 2.5 percent below the United Kingdom's $37,371 per capita GDP. Like most of the global economy, Ireland fell into a deep recession in 2009, with GDP contracting by over 7 percent. To understand the factors driving the Irish economy and the prospects for future equity returns, use the data in Exhibits 11 and 12 and the population data below to address the following questions:

1 Using the growth accounting framework data, evaluate the sources of growth for the Irish economy from 1995 to 2009.
2 What is likely to happen to the potential rate of growth for Ireland? What are the prospects for equity returns?

	2000	2010	Annual Growth Rate
Population (millions)	3.8	4.5	1.71%
Net immigration total (2000–2010)		375,587	
Net immigration total (2009–2010)		–20,000	
Population less immigrants (millions)	3.8	4.1	0.8%

Solution to 1:

The sources of growth for an economy include labor quantity, labor quality, non-ICT capital, ICT capital, and TFP. The growth accounting data in Exhibit 12 indicate that economic growth in Ireland from 1995 to 2009 is explained by the following factors:

Input	Contribution: 1995–2005		Contribution: 2005–2009	
Labor		**2.3%**		**–0.1%**
Labor quantity	2.0%		–0.2%	
Labor quality	0.3%		0.1%	
Capital/Investment		**3.3%**		**2.2%**
Non-ICT capital	2.6%		1.8%	
ICT capital	0.7%		0.4%	
TFP		**1.7%**		**–2.1%**
Total: GDP growth		**7.3%**		**0.0%**

In sum, the main driver of growth for the Irish economy since 1995 has been capital spending. It accounted for over 45 percent of growth in 1995–2005 and has been the only factor contributing to growth in the Irish economy since 2005, offsetting the negative contribution from labor and TFP. Another way to look at growth in Ireland for the period 2005–2009 is that all the growth is through capital deepening. As shown in Exhibit 11, capital deepening added 2.9 percent to growth and by offsetting the decline in TFP caused an increase in labor productivity of 0.8 percent.

Solution to 2:

Looking forward, prospects for the economy are not as favorable as in the past. To estimate the growth rate in potential GDP, we use Equation 5, given by

$$\text{Growth rate of potential GDP} = \text{Long-term growth rate of labor force} + \text{Long-term growth rate in labor productivity}$$

To use this equation, we need to project the growth rate in the labor input and labor productivity. The total hours worked data in Exhibit 11 are one potential source to use to estimate the growth rate of the labor input. Exhibit 11 shows the labor input declining by 0.8 percent between 2005 and 2009. The problem here is that the decline in hours worked is overstated because of the negative impact of the recession on hours worked. As an alternative, the labor input should grow at the same rate as the population plus the net change due to immigration. The population data for Ireland (given above) show that over half of the population growth between 2000 and 2010 was due to immigration. Since 2009, however, outward migration has replaced inward migration, reducing the growth rate in the labor input. Thus, if the 2000–2010 influx of immigrants is reversed over the next decade, a reasonable, perhaps somewhat conservative, estimate for labor force growth is zero. We also assume:

1 There is no increase in labor productivity coming from capital deepening as investment slows (resulting in essentially no growth in net investment and the physical capital stock).
2 TFP growth reverts to its average growth rate of 1.7 percent in the 1995–2005 time period (see Exhibit 11).
3 Labor productivity grows at the same rate as TFP.

Thus, growth in potential GDP is 0.0% + 1.7% = 1.7%.

In summary, despite the projected rebound in TFP growth, overall potential growth in Ireland is likely to decline as labor input growth and capital deepening no longer contribute to overall growth. As discussed in Section 3 of the reading, slower growth in potential GDP will limit potential earnings growth and equity price appreciation.

EXAMPLE 9

Investment Outlook for China and India

The Investment Policy Committee at Global Invest Inc. is interested in increasing the firm's exposure to either India or China because of their rapid rates of economic growth. Economic growth in China has been close to 10 percent over the last few years, and India has grown over 7 percent. You are asked by the committee to do the following:

1 Determine the sources of growth for the two economies and review the data on productivity and investment using information from Exhibits 5, 9, 10, 11, and 12. Which of the two countries looks more attractive based on the sources of growth?
2 Estimate the long-term sustainable earnings growth rate using data from 1995 to 2008.
3 Make an investment recommendation.

Solution to 1:

The sources of economic growth include size of labor force, quality of labor force (human capital), ICT and non-ICT capital, natural resources, and technology. Looking at the sources of growth in Exhibit 12, we get the following:

Input	Percent Contribution: 1995–2005	Percent Contribution: 2005–2008
India		
Labor quantity	1.0	1.1
Labor quality	0.2	0.1
Non-ICT capital	2.7	3.7
ICT capital	0.5	0.9
TFP	1.9	2.4
Total: GDP growth	6.3	8.2

Input	Percent Contribution: 1995–2005	Percent Contribution: 2005–2008
China		
Labor quantity	0.5	0.3
Labor quality	0.2	0.2
Non-ICT capital	4.5	5.5
ICT capital	1.1	1.3
TFP	1.5	4.2
Total: GDP growth	7.8	11.5

- The contribution of the labor quantity input is more important to growth in India than in China. Labor quantity contributed 1 percent to India's GDP growth over 1995–2005 and 1.1 percent over 2005–2008. The equivalent numbers for China are 0.5 percent and 0.3 percent, respectively. Looking ahead, labor is likely to be a major factor adding to India's growth. The population of India (Exhibit 5) is growing at a faster rate than that of China. The annual growth rate in population from 2000 to 2010 was 1.52 percent in India versus 0.57 percent in China. Also, hours worked in India (Exhibit 11) are growing at a faster rate than in China. Therefore, the workforce and labor quantity input should grow faster in India. The edge here goes to India.
- The contribution to GDP made by the quality of the labor force is essentially identical in the two countries (0.2 percent in China versus 0.2 percent in India between 1995 and 2005 and 0.2 percent in China and 0.1 percent in India between 2005 and 2008). This factor is a tie.
- The contribution of non-ICT capital investment is significantly higher in China (4.5 percent in China versus 2.7 percent in India between 1995 and 2005 and 5.5 percent in China and 3.7 percent in India between 2005 and 2008). The edge goes to China.
- The contribution of ICT capital investment is significantly higher in China (1.1 percent in China versus 0.5 percent in India between 1995 and 2005 and 1.3 percent in China and 0.9 percent in India between 2005 and 2008). The edge goes to China.
- Both countries spend a high percentage of GDP on capital investment (Exhibit 9). In 2010, investment spending as a percentage of GDP was 48.2 percent in China and 36.8 percent in India. The Chinese share is higher, and this provides China with an edge unless diminishing marginal returns to capital deepening become an issue. However, this is not likely for a while given the low level of capital per worker in China. China and India still have a long way to go to converge with the developed economies. The advantage goes to China.
- The contribution of technological progress is measured by TFP. Comparing the two countries, TPF growth was higher in India over the period 1995–2005 (1.9 percent in India versus 1.5 percent in China). For the period 2005–2009, however, TFP growth was significantly higher in China (4.2 percent versus 2.4 percent). In addition, expenditures on

R&D for 2009 (Exhibit 10) as a percentage of GDP were higher in China (1.7 percent in China and 0.8 percent in India). The edge here goes to China.

- Finally, growth in overall labor productivity (Exhibit 11) is considerably higher in China than India (10.3 percent in China versus 6.0 percent in India between 2005 and 2008). This is due to a greater increase in the capital-to-labor ratio in China (because of the high rate of investment, the physical capital stock is growing faster than the labor input) and due to faster technological progress in China. The edge here goes to China.

In sum, based on the sources of growth, China appears to be better positioned for growth in the future.

Solution to 2:

Estimates of potential GDP using the inputs from Exhibit 11 for China and India are

Growth rate in potential GDP = Long-term growth rate of labor force (equals growth in hours worked in Exhibit 11) + Long-term growth rate in labor productivity.

China (using 1995–2008)[a]

Growth in potential = 1.1% + 7.5% = 8.6%

India (using 1995–2008)

Growth in potential = 2.1% + 4.6% = 6.7%

Solution to 3:

Growth prospects in both countries are very attractive. However, China's growth potential is higher because of its greater level of capital spending and the greater contribution of technological progress toward growth. Long-term earnings growth is closely tied to the growth rate in potential GDP. Therefore, based on the previous calculations, earnings in China would be projected to grow at an annual rate of 8.6 percent, compared with 6.7 percent in India. Over the next decade, ignoring current valuation, the Chinese equity market would be projected to outperform the Indian market as its higher rate of sustainable growth translates into a higher rate of appreciation in equity values.[b]

[a] Calculated as geometric mean growth rates using data for the 1995–2005 and 2005–2008 subperiods.
[b] It bears repeating that the global economy is evolving rapidly and past trends may or may not be sustained. This is especially true of China and India. To provide concrete answers that do not require the reader to bring in additional information, our exercise solutions must assume past patterns are indicative of the future.

5 THEORIES OF GROWTH

The factors that drive long-term economic growth and determine the rate of sustainable growth in an economy are the subject of much debate among economists. The academic growth literature includes three main paradigms with respect to per capita growth in an economy—the classical, neoclassical, and endogenous growth models. Per capita economic growth under the classical model is only temporary because an exploding population with limited resources brings growth to an end. In the neoclassical model, long-run per capita growth depends solely on exogenous technological progress. The final model of growth attempts to explain technology within the model itself—thus the term endogenous growth.

5.1 Classical Model

Classical growth theory was developed by Thomas Malthus in his 1798 publication *Essay on the Principle of Population.* Commonly referred to as the Malthusian theory, it is focused on the impact of a growing population in a world with limited resources. The concerns of resource depletion and overpopulation are central themes within the Malthusian perspective on growth. The production function in the classical model is relatively simple and consists of a labor input with land as a fixed factor. The key assumption underlying the classical model is that population growth accelerates when the level of per capita income rises above the subsistence income, which is the minimum income needed to maintain life. This means that technological progress and land expansion, which increase labor productivity, translate into higher population growth. But because the labor input faces diminishing marginal returns, the additional output produced by the growing workforce eventually declines to zero. Ultimately, the population grows so much that labor productivity falls and per capita income returns back to the subsistence level.

The classical model predicts that in the long run, the adoption of new technology results in a larger but not richer population. Thus, the standard of living is constant over time even with technological progress, and there is no growth in per capita output. As a result of this gloomy forecast, economics was labeled the "dismal science."

The prediction from the Malthusian model failed for two reasons:

1 The link between per capita income and population broke down. In fact, as the growth of per capita income increased, population growth slowed rather than accelerating as predicted by the classical growth model.

2 Growth in per capita income has been possible because technological progress has been rapid enough to more than offset the impact of diminishing marginal returns.

Because the classical model's pessimistic prediction never materialized, economists changed the focus of the analysis away from labor to capital and to the neoclassical model.

5.2 Neoclassical Model

Robert Solow devised the mainstream neoclassical theory of growth in the 1950s.[20] The heart of this theory is the Cobb–Douglas production function discussed in Section 4.1. As before, the potential output of the economy is given by

$$Y = AF(K,L) = AK^{\alpha}L^{1-\alpha}$$

20 Solow (1957).

where K is the stock of capital, L is the labor input, and A is total factor productivity.[21] In the neoclassical model, both capital and labor are variable inputs each subject to diminishing marginal productivity.

The objective of the neoclassical growth model is to determine the long-run growth rate of output per capita and relate it to (a) the savings/investment rate, (b) the rate of technological change, and (c) population growth.

5.2.1 *Balanced or Steady State Rate of Growth*

As with most economic models, the neoclassical growth model attempts to find the equilibrium position toward which the economy will move. In the case of the Solow model, this equilibrium is the balanced or **steady state rate of growth** that occurs when the output-to-capital ratio is constant. Growth is balanced in the sense that capital per worker and output per worker grow at the same rate.

We begin the analysis by using the per capita version of the Cobb–Douglas production function given earlier in Equation 3:

$$y = Y/L = Ak^{\alpha}$$

where $k = K/L$. Using their definitions, the rates of change of capital per worker and output per worker are given by[22]

$$\Delta k/k = \Delta K/K - \Delta L/L$$

and

$$\Delta y/y = \Delta Y/Y - \Delta L/L$$

From the production function, the growth rate of output per worker is also equal to

$$\Delta y/y = \Delta A/A + \alpha \Delta k/k \tag{6}$$

The physical capital stock in an economy will increase because of gross investment (I) and decline because of depreciation. In a closed economy, investment must be funded by domestic saving. Letting s be the fraction of income (Y) that is saved, gross investment is given by $I = sY$. Assuming the physical capital stock depreciates at a constant rate, δ, the change in the physical capital stock is given by

$$\Delta K = sY - \delta K$$

Subtracting labor supply growth, $\Delta L/L \equiv n$, and rearranging gives

$$\Delta k/k = sY/K - \delta - n \tag{7}$$

In the steady state, the growth rate of capital per worker is equal to the growth rate of output per worker. Thus,

$$\Delta k/k = \Delta y/y = \Delta A/A + \alpha \Delta k/k$$

from which we get

$$\Delta y/y = \Delta k/k = (\Delta A/A)/(1 - \alpha)$$

---

21 Our exposition of the neoclassical model with technological progress reflected in total factor productivity corresponds to what is known as "Hicks neutral" technical change. The neoclassical model is usually presented with "Harrod neutral" or "labor augmenting" technical change. In that formulation, the production function is given by $Y = F(K, BL)$, where B represents technological change and (BL) is interpreted as the "effective" labor supply. In general, this is not equivalent to our formulation using TFP. However, they are equivalent if, as we assume here, the function $F()$ has the Cobb–Douglas form. To see this, note that $[K^{\alpha}(BL)^{1-\alpha}] = [B^{1-\alpha}(K^{\alpha}L^{1-\alpha})] = [A(K^{\alpha}L^{1-\alpha})]$, where $A \equiv B^{1-\alpha}$ is total factor productivity.

22 Strictly speaking, these and other rate of change equations are exact only for changes over arbitrarily short periods ("continuous time").

Letting θ denote the growth rate of TFP (i.e., $\Delta A/A$), we see that the equilibrium sustainable growth rate of output per capita (= Growth rate of capital per worker) is a constant that depends only on the growth rate of TFP (θ) and the elasticity of output with respect to capital (α). Adding back the growth rate of labor (n) gives the sustainable growth rate of output.

$$\text{Growth rate of output per capita} = \frac{\theta}{1-\alpha} \quad (8)$$

$$\text{Growth rate of output} = \frac{\theta}{1-\alpha} + n$$

This is the key result of the neoclassical model. Note that $[\theta/(1-\alpha)]$ is the steady state growth rate of labor productivity, so Equation 8 is consistent with the labor productivity growth accounting equation discussed in Section 4.3.

Substituting $[\theta/(1-\alpha)]$ into the left-hand side of Equation 7 and rearranging gives the equilibrium output-to-capital ratio, denoted by the constant Ψ.

$$\frac{Y}{K} = \left(\frac{1}{s}\right)\left[\left(\frac{\theta}{1-\alpha}\right) + \delta + n\right] \equiv \Psi \quad (9)$$

In the steady state, the output-to-capital ratio is constant and the capital-to-labor ratio (k) and output per worker (y) grow at the same rate, given by $[\theta/(1-\alpha)]$. On the steady state growth path, the marginal product of capital is also constant and, given the Cobb–Douglas production function, is equal to $\alpha(Y/K)$. The marginal product of capital is also equal to the real interest rate in the economy. Note that even though the capital-to-labor ratio (k) is rising at rate $[\theta/(1-\alpha)]$ in the steady state, the increase in the capital-to-labor ratio (k) has no impact on the marginal product of capital, which is not changing. *Capital deepening is occurring, but it has no effect on the growth rate of the economy or on the marginal product of capital once the steady state is reached.*

EXAMPLE 10

Steady State Rate of Growth for China, Japan, and Ireland

Earlier examples generated estimates of potential growth for China (11.5 percent), Japan (0.8 percent), and Ireland (1.7 percent). Given the data below,

1 Calculate the steady state growth rates from the neoclassical model for China, Japan, and Ireland.

2 Compare the steady state growth rates to the growth rates in potential GDP estimated in Examples 7–9 and explain the results.

	Labor Cost in Total Factor Cost (%)	TFP Growth (%)	Labor Force Growth (%)
China	46.5	2.5	1.2
Japan	57.3	0.2	0.0
Ireland	56.7	0.8	0.0

Sources: Conference Board Total Economy Database; labor cost and TFP growth are based on 1995–2009 data for Japan and Ireland and 1995–2008 data for China. Labor force growth estimates are from earlier examples.

Solution to 1:

Using Equation 8, the steady state growth rate in the neoclassical model is given by

$\Delta Y/Y = (\theta)/(1 - \alpha) + n$ = Growth rate of TFP scaled by labor factor share + Growth rate in the labor force

Using the above equation and data, steady state growth rates for the three countries are estimated as follows:

China: The labor share of output $(1 - \alpha)$ is given by the average of the labor cost as a percentage of total factor cost, which is equal to 0.465 for China. The growth rate in the labor force is 1.2 percent, and the growth rate of TFP is 2.5 percent.

Steady state growth rate = 2.5%/0.465 + 1.2% = 6.58%

Japan: The labor share of output $(1 - \alpha)$ for Japan is 0.573. The growth rate in the labor force is 0.0 percent, and TFP growth is 0.2 percent.

Steady state growth rate = 0.2%/0.573 + 0.0% = 0.35%

Ireland: The labor share of output $(1 - \alpha)$ is 0.567 percent for Ireland. The growth rate in the labor force is 0.0 percent, and TFP growth is 0.8 percent.

Steady state growth rate = 0.8%/0.567 + 0.0% = 1.41%

Solution to 2:

The growth rate in potential GDP for China (8.6 percent, estimated in Example 9) is significantly above the estimated 6.58 percent steady state growth rate. The reason for this is that the economy of China is still in the process of converging to the higher income levels of the United States and the major economies in Europe. The physical capital stock is below the steady state, and capital deepening is a significant factor increasing productivity growth (see Exhibit 11) and the growth in potential GDP.

This is not the case for Japan and Ireland. Both countries are operating at essentially the steady state. The estimated growth rate in potential GDP for Japan (0.8 percent, from Example 7) is only slightly above its 0.35 percent steady state growth rate. Likewise, the estimated growth rate in potential GDP for Ireland (1.7 percent, from Example 8) is effectively equal to its estimated steady state growth rate of 1.4 percent. Operating close to the steady state means that capital investment in these countries, which results in an increasing capital-to-labor ratio, has no significant effect on the growth rate of the economy. Only changes in the growth rates of TFP and labor and in the labor share of output have an impact on potential GDP growth.

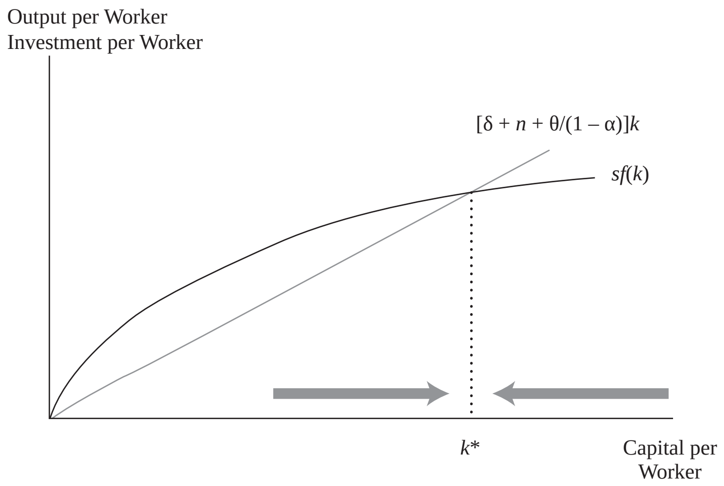
An intuitive way to understand the steady state equilibrium given in Equation 9 is to transform it into a savings/investment equation:

$$sy = \left[\left(\frac{\theta}{1-\alpha}\right) + \delta + n\right]k$$

Steady state equilibrium occurs at the output-to-capital ratio where the savings and actual gross investment per worker generated in the economy (sy) are just sufficient to (1) provide capital for new workers entering the workforce at rate n, (2) replace plant and equipment wearing out at rate δ, and (3) deepen the physical capital stock at the rate $[\theta/(1 - \alpha)]$ required to keep the marginal product of capital equal to the rental price of capital.

Exhibit 13 shows the steady state equilibrium graphically. The straight line in the exhibit indicates the amount of investment required to keep the physical capital stock growing at the required rate. Because the horizontal axis is capital per worker, the slope of the line is given by $[\delta + n + \theta/(1 - \alpha)]$. The curved line shows the amount of actual investment per worker and is determined by the product of the saving rate and the production function. It is curved because of diminishing marginal returns to the capital input in the production function. The intersection of the required investment and actual investment lines determines the steady state. Note that *this exhibit is a snapshot at a point in time*. Over time, the capital-to-labor ratio rises at rate $[\theta/(1 - \alpha)]$ as the actual saving/investment curve $[sf(k)]$ shifts upward because of TFP growth, and *the equilibrium moves upward and to the right along the straight line.*

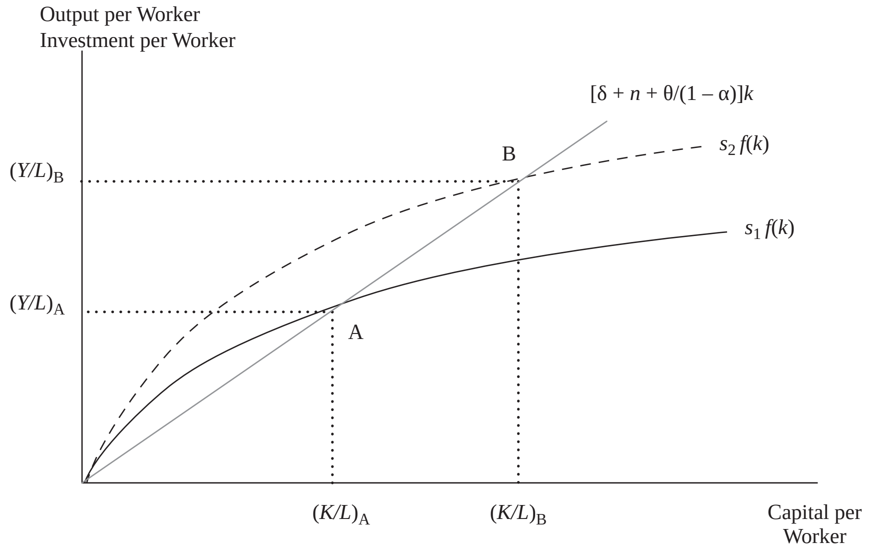
Exhibit 13 Steady State in the Neoclassical Model



The impact of the various parameters in the model on the steady state can also be seen in the exhibit. At any point in time when the economy is on its steady state growth path, the exogenous factors—labor supply and TFP—are fixed. We would like to know what effect each of the parameters in the model has on the steady state capital-to-labor ratio and therefore on output per worker. For example, if there are two economies that differ only with respect to one parameter, what does that imply about their per capita incomes? All else the same, we can say the following regarding the impact of the parameters:

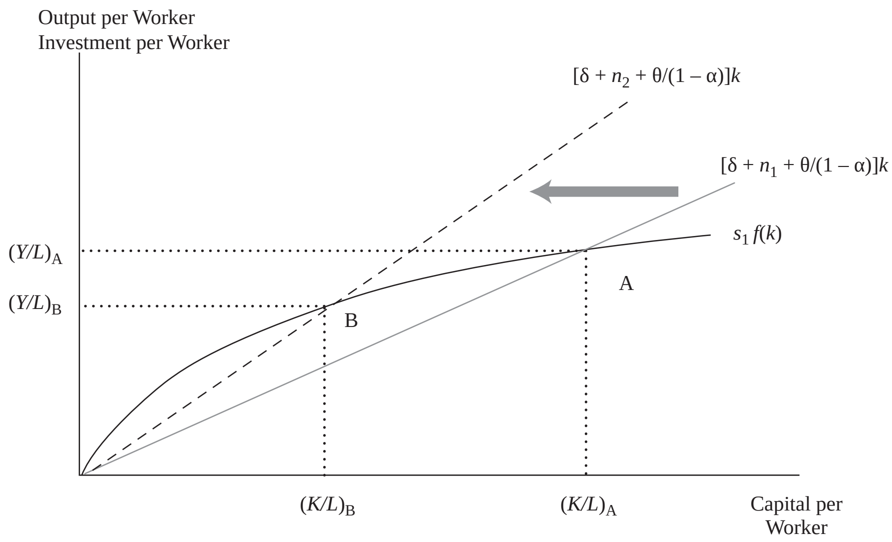
- *Saving rate (s)*: An increase in the saving rate implies a higher capital-to-labor ratio (k) and higher output per worker (y) because a higher saving rate generates more saving/investment at every level of output. In Exhibit 14, the saving/investment curve $[sf(k)]$ shifts upward from an initial steady state equilibrium at point A to a new equilibrium at point B. At the new equilibrium point, it intersects the required investment line $[\delta + n + \theta/(1 - \alpha)]$ at higher capital-to-labor and output per worker ratios. Note that although the higher saving rate increases both k and y, it has no impact on the steady state growth rates of output per capita or output (Equation 8).

Exhibit 14 Impact on the Steady State: Increase in the Saving Rate



- *Labor force growth (n)*: An increase in the labor force growth rate reduces the equilibrium capital-to-labor ratio because a corresponding increase in the steady state growth rate of capital is required. Given the gross saving/investment rate, this can only be achieved at a lower capital-to-labor ratio. Output per worker is correspondingly lower as well. In Exhibit 15, the higher population growth rate increases the slope of the required investment line. This shifts the steady state equilibrium from point A to point B, where it intersects the supply of saving/investment curve at lower capital-to-labor and output per worker ratios.
- *Depreciation rate (δ)*: An increase in the depreciation rate reduces the equilibrium capital-to-labor and output per worker ratios because a given rate of gross saving generates less net capital accumulation. Graphically, it increases the slope of the required investment line and affects the steady state equilibrium in the same way as labor force growth (Exhibit 15).
- *Growth in TFP (θ)*: An increase in the growth rate of TFP reduces the steady state capital-to-labor ratio and output per worker for given levels of labor input and TFP. This result must be interpreted with care. Raising the growth rate of TFP means that output per worker will grow faster in the future (Equation 8), but at a given point in time, a given supply of labor, and a given *level* of TFP, output per worker is lower than it would be with a slower TFP growth rate. In effect, the economy is on a steeper trajectory off a lower base of output per worker. Graphically, it is identical to Exhibit 15 in that faster TFP growth steepens the required investment line (increases the slope), which intersects with the available saving/investment curve at lower capital-to-labor and investment per worker ratios.

Exhibit 15 Impact on the Steady State: Increase in Labor Force Growth



In sum, such factors as the saving rate, the growth rate of the labor force, and the depreciation rate change the *level* of output per worker but do not permanently change the *growth rate* of output per worker. A permanent increase in the growth rate in output per worker can only occur if there is a change in the growth rate of TFP.

So far we have focused on the steady state growth path. What happens if the economy has not yet reached the steady state? During the transition to the steady state growth path, the economy can experience either faster or slower growth relative to the steady state. Using Equations 6, 7, and 9, we can write the growth rates of output per capita and the capital-to-labor ratio as, respectively,

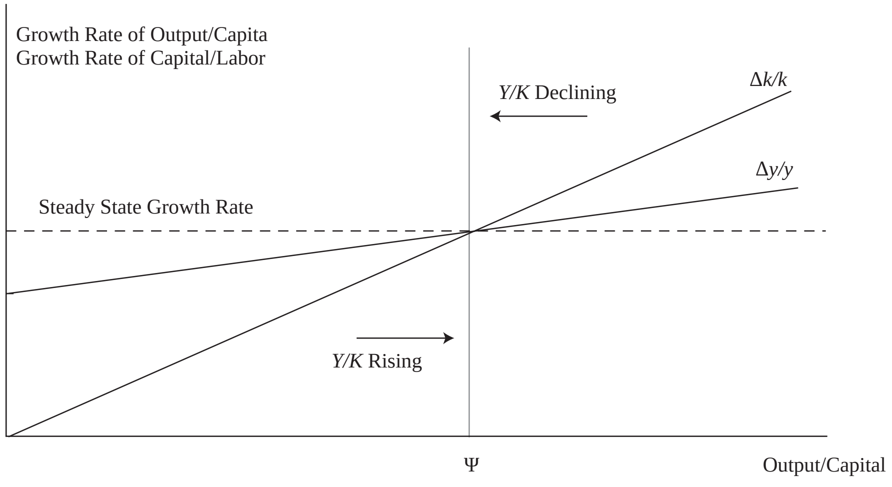
$$\frac{\Delta y}{y} = \left(\frac{\theta}{1-\alpha}\right) + \alpha s\left(\frac{Y}{K} - \Psi\right) = \left(\frac{\theta}{1-\alpha}\right) + \alpha s\left(y/k - \Psi\right) \tag{10}$$

and

$$\frac{\Delta k}{k} = \left(\frac{\theta}{1-\alpha}\right) + s\left(\frac{Y}{K} - \Psi\right) = \left(\frac{\theta}{1-\alpha}\right) + s\left(y/k - \Psi\right), \tag{11}$$

where the second equality in each line follows from the definitions of y and k, which imply $(Y/K) = y/k$. These relationships are shown in Exhibit 16.

Exhibit 16 Dynamics in the Neoclassical Model



If the output-to-capital ratio is above its equilibrium level (ψ), the second term in Equations 10 and 11 is positive and the growth rates of output per capita and the capital-to-labor ratio are above the steady state rate $[\theta/(1-\alpha)]$. This corresponds to a situation in which actual saving/investment exceeds required investment and above-trend growth in per capita output is driven by an above-trend rate of capital deepening. This situation usually reflects a relatively low capital-to-labor ratio but could, at least in principle, arise from high TFP. Because $\alpha < 1$, capital is growing faster than output and the output-to-capital ratio is falling. Over time, the growth rates of both output per capita and the capital-to-labor ratio decline to the steady state rate.

Of course, the converse is true if the output-to-capital ratio is below its steady state level. Actual investment is insufficient to sustain the trend rate of growth in the capital-to-labor ratio, and both output per capita and the capital-to-labor ratio grow more slowly. This situation usually corresponds to a relatively high and unsustainable capital-to-labor ratio, but could reflect relatively low TFP and hence relatively low output. Over time, output grows faster than capital, the output-to-capital ratio rises, and growth converges to the trend rate.

5.2.2 *Implications of the Neoclassical Model*

There are four major groups of conclusions from the neoclassical model:

1 Capital Accumulation

- **a** Capital accumulation affects the level of output but not the growth rate in the long run.
- **b** Regardless of its initial capital-to-labor ratio or initial level of productivity, a growing economy will move to a point of steady state growth.
- **c** In a steady state, the growth rate of output equals the rate of labor force growth plus the rate of growth in TFP scaled by labor's share of income $[n + \theta/(1-\alpha)]$.[23] The growth rate of output does not depend on the accumulation of capital or the rate of business investment.

23 Readers who are familiar with the "labor-augmenting" technical change formulation of the neoclassical model should note that in that formulation, the rate of labor-augmenting technical change is also the growth rate of labor productivity. In our formulation, the growth rate of labor productivity is $[\theta/(1-\alpha)]$. So both formulations imply that long-run growth equals the growth rate of the labor supply (n) plus a constant growth rate of labor productivity.

2 Capital Deepening vs. Technology

- **a** Rapid growth that is above the steady state rate of growth occurs when countries first begin to accumulate capital; but growth will slow as the process of accumulation continues (see Exhibit 16).
- **b** Long-term sustainable growth cannot rely solely on capital deepening investment—that is, on increasing the stock of capital relative to labor. If the capital-to-labor ratio grows too rapidly (i.e., faster than labor productivity), capital becomes less productive, resulting in slower rather than faster growth.
- **c** More generally, increasing the supply of some input(s) too rapidly relative to other inputs will lead to diminishing marginal returns and cannot be the basis for sustainable growth.
- **d** In the absence of improvements in TFP, the growth of labor productivity and per capita output would eventually slow.
- **e** Because of diminishing marginal returns to capital, the only way to sustain growth in potential GDP per capita is through technological change or growth in total factor productivity. This results in an upward shift in the production function—the economy produces more goods and services for any given mix of labor and capital inputs.

3 Convergence

- **a** Given the relative scarcity and hence high marginal productivity of capital and potentially higher saving rates in developing countries, the growth rates of developing countries should exceed those of developed countries.
- **b** As a result, there should be a convergence of per capita incomes between developed and developing countries over time.

4 Effect of Savings on Growth

- **a** The initial impact of a higher saving rate is to temporarily raise the rate of growth in the economy.[24] In response to the higher saving rate, growth exceeds the steady state growth rate during a transition period. However, the economy returns to the balanced growth path after the transition period.
- **b** During the transition period, the economy moves to a higher level of per capita output and productivity.
- **c** Once an economy achieves steady state growth, the growth rate does not depend on the percentage of income saved or invested. Higher savings cannot permanently raise the growth rate of output.
- **d** However, countries with higher saving rates will have a higher level of per capita output, a higher capital-to-labor ratio, and a higher level of labor productivity.

24 Mathematically, this can be seen as follows: Equation 9 indicates that an increase in the saving rate (s) reduces the steady state output-to-capital ratio (ψ). This makes the last term in Equations 10 and 11 positive, raising the growth rates of output per capita (y) and the capital-to-labor ratio (k) above the steady state rate.

EXAMPLE 11

Comparative Statics and Transitional Growth in the Neoclassical Model

Beginning in steady state equilibrium, an economy's saving rate suddenly increases from 20 percent of income to 30 percent of income. Other key parameters describing the economy are as follows:

Growth rate of TFP (θ)	= 0.02
Income share of capital (α)	= 0.35
Depreciation rate (δ)	= 0.10
Labor force growth rate (n)	= 0.01

The following table shows the output-to-capital ratio that will prevail in this economy at various points in time after the increase in the saving rate.

Years after Saving Rate Increase	Output-to-Capital Ratio
5	0.5947
10	0.5415
25	0.4857
50	0.4708
100	0.4693
New steady state	??

By rearranging the Cobb–Douglas production function (Equation 3), the proportional impact of the saving rate change on the capital-to-labor ratio can be expressed in terms of the proportional impact on the output-to-capital ratio. The proportional impact on per capita income can then be determined from the production function (Equation 3). Labeling the paths with and without the change in saving rate as "new" and "old" respectively, at each date we have:[a]

$$\frac{k_{new}}{k_{old}} = \left[\frac{\left(Y/K\right)_{new}}{\left(Y/K\right)_{old}}\right]^{\frac{1}{\alpha-1}}$$

and

$$\frac{y_{new}}{y_{old}} = \left(\frac{k_{new}}{k_{old}}\right)^{\alpha}$$

1 Using Equations 8 and 9, calculate the steady state growth rate of per capita income and the steady state output-to-capital ratio both before and after the change in the saving rate. What happens to the capital-to-labor ratio and output per capita?

2 Use the output-to-capital ratios given in the table above along with Equation 10 and your answers to Question 1 to determine the growth rate of per capita income that will prevail immediately following the change in the saving rate and at each of the indicated times after the change. Explain the pattern of growth rates.

3 Using the output-to-capital ratios given in the table above, calculate the proportional impact of the increased saving rate on the capital-to-labor ratio and on per capita income over time. With respect to these variables, how will the new steady state compare with the old steady state?

Solution to 1:

From Equation 8, the steady state growth rate of per capita income, both before and after the increase in the saving rate, is $\Delta y/y = \theta/(1 - \alpha) = 0.02/(1 - 0.35) = 0.0308$, or 3.08 percent. From Equation 9, the steady state output-to-capital ratio is

$$\frac{Y}{K} = \left(\frac{1}{s}\right)\left[\left(\frac{\theta}{1-\alpha}\right) + \delta + n\right] \equiv \Psi$$

Using the parameter values given above, $[\theta/(1 - \alpha) + \delta + n] = (0.0308 + 0.10 + 0.01) = 0.1408$, so the steady state output-to-capital ratio is $(0.1408/0.2) = 0.7040$ with the initial 20 percent saving rate and $(0.1408/0.30) = 0.4693$ with the new 30 percent saving rate. As shown in Exhibit 14, both the capital-to-labor ratio and output per worker are at higher *levels* in the new steady state. But once the new steady state is achieved, they do not grow any faster than they did in the steady state with the lower saving rate.

Solution to 2

According to Equation 10, the growth rate of per capita income is given by

$$\frac{\Delta y}{y} = \left(\frac{\theta}{1-\alpha}\right) + \alpha s\left(y/k - \Psi\right)$$

Immediately following the increase in the saving rate, the relevant value of ψ becomes the new steady state output-to-capital ratio (0.4693). The actual output-to-capital ratio does not change immediately, so y/k is initially still 0.7040. Plugging these values into the growth equation above gives the growth rate of per capita income:

$$\Delta y/y = 0.0308 + (0.35)(0.30)(0.7040 - 0.4693) = 0.0554, \text{ or } 5.54 \text{ percent}$$

Similar calculations using the output-to-capital ratios in the table above give the following:

Years after Saving Rate Increase	Output-to-Capital Ratio	Growth Rate of Per Capita Income (%)
0	0.7040	5.54
5	0.5947	4.39
10	0.5415	3.84
25	0.4857	3.25
50	0.4708	3.09
100	0.4693	3.08
New steady state	0.4693	3.08

The growth rate "jumps" from the steady state rate of 3.08 percent to 5.54 percent when the saving rate increases because the increase in saving/investment results in more rapid capital accumulation. Over time, the growth rate slows because the marginal productivity of capital declines as the capital-to-labor ratio increases. In addition, as the capital-to-labor ratio increases and the output-to-capital ratio declines, a greater portion of savings is required to maintain the capital-to-labor ratio, leaving a smaller portion for continued capital deepening. Roughly two-thirds of the growth acceleration has dissipated after 10 years.

Solution to 3:

Using the output-to-capital ratio that will prevail five years after the saving rate increase, the proportional impact on the capital-to-labor ratio and on per capita income will be

$$\frac{k_{new}}{k_{old}} = \left[\frac{\left(Y/K\right)_{new}}{\left(Y/K\right)_{old}}\right]^{\frac{1}{\alpha-1}} = \left[\frac{0.5947}{0.7040}\right]^{\frac{-1}{0.65}} = 1.2964$$

and

$$\frac{y_{new}}{y_{old}} = \left(\frac{k_{new}}{k_{old}}\right)^{\alpha} = 1.2964^{0.35} = 1.0951$$

Thus, after five years, the capital-to-labor ratio will be 29.64 percent higher than it would have been without the increase in the saving rate and per capita income will be 9.51 percent higher. Similar calculations for the other time periods give the following:

Years after Saving Rate Increase	Proportionate Increase (%) in: Capital-to-Labor Ratio	Per Capita Income
0	0.00	0.00
5	29.64	9.51
10	49.74	15.18
25	77.01	22.12
50	85.71	24.19
100	86.68	24.42
New steady state	86.68	24.42

In the new steady state, the capital-to-labor ratio will be 86.68 percent higher at every point in time than it would have been in the old steady state. Per capita income will be 24.42 percent higher at every point in time. Both variables will be growing at the same rate (3.08 percent) as they would have been in the old steady state.

[a] Note that the output-to-capital ratio would have been constant on the original steady state path. Because of the impact of total factor productivity, the capital-to-labor ratio and output per capita are not constant even in steady state. In comparing "paths" for these variables, we isolate the impact of the saving rate change by canceling out the effect of TFP growth. Mathematically, we cancel out A in Equation 3 to get the equations shown here.

5.2.3 *Extension of the Neoclassical Model*

Solow (1957) used the growth accounting equation to determine the contributions of each factor to economic growth in the United States for the period 1909–1949. He reached the surprising conclusion that over 80 percent of the per capita growth in the United States was due to TFP. Denison (1985) authored another study examining US growth for the period 1929–1982 using the Solow framework. His findings were similar to Solow's, with TFP explaining nearly 70 percent of US growth. The problem with these findings is that the neoclassical model provides no explicit explanation of the economic determinants of technological progress or how TFP changes over time. Because technology is determined outside the model (i.e., exogenously), critics argue that the neoclassical model ignores the very factor driving growth in the economy. Technology is simply the residual or the part of growth that cannot be explained by other inputs, such as capital and labor. This lack of an explanation for technology led to growing dissatisfaction with the neoclassical model.

The other source of criticism of the neoclassical model is the prediction that the steady state rate of economic growth is unrelated to the rate of saving and investment. Long-run growth of output in the Solow model depends only on the rates of growth of the labor force and technology. Higher rates of investment and savings have only a transitory impact on growth. Thus, an increase in investment as a share of GDP from 10 percent to 15 percent of GDP will have a positive impact on the near-term growth rate but will not have a permanent impact on the ultimately sustainable percentage growth rate. This conclusion makes many economists uncomfortable. Mankiw (1995) provided evidence rebutting this hypothesis and showed that saving rates and growth rates are positively correlated across countries. Finally, the neoclassical model predicts that in an economy where the stock of capital is rising faster than labor productivity, the return to investment should decline with time. For the advanced countries, the evidence does not support this argument because returns have not fallen over time.

Critiques of the neoclassical model led to two lines of subsequent research on economic growth. The first approach, which was originated by Jorgenson (1966, 2000), is termed the augmented Solow approach. It remains in the neoclassical tradition in that diminishing marginal returns are critical and there is no explanation for the determinants of technological progress. Instead, this approach attempts to reduce empirically the portion of growth attributed to the unexplained residual labeled technological progress (TFP). The idea is to develop better measures of the inputs used in the production function and broaden the definition of investment by including human capital, research and development, and public infrastructure. In addition, the composition of capital spending is important. Higher levels of capital spending on high-technology goods will boost productivity more than spending on machine tools or structures.

By adding inputs like human capital to the production function, the augmented Solow model enables us to more accurately measure the contribution of technological progress to growth. However, the economy still moves toward a steady state growth path because even broadly defined capital is assumed to eventually encounter diminishing marginal returns. In essence, this line of research uses the growth accounting methodology and increases the number of inputs in the production function in order to provide a more accurate measure of technological progress. The second approach is the endogenous growth theory, which we examine in the next section.

5.3 Endogenous Growth Theory

The alternative to the neoclassical model is a series of models known as endogenous growth theory. These models focus on explaining technological progress rather than treating it as exogenous. In these models, self-sustaining growth emerges as a natural consequence of the model and the economy does not necessarily converge to a steady state rate of growth. Unlike the neoclassical model, there are *no diminishing marginal returns to capital for the economy as a whole* in the endogenous growth models. So increasing the saving rate permanently increases the rate of economic growth. These models also allow for the possibility of increasing returns to scale.

Romer (1986) provided a model of technological progress and a rationale for why capital does not experience diminishing marginal returns. He argued that capital accumulation is the main factor accounting for long-run growth, once the definition of capital is broadened to include such items as human or knowledge capital and research and development (R&D). R&D is defined as investment in new knowledge that improves the production process. In endogenous growth theory, knowledge or human capital and R&D spending are factors of production, like capital and labor, and have to be paid for through savings.

Companies spend on R&D for the same reason they invest in new equipment and build new factories: to make a profit. R&D spending is successful if it leads to the development of a new product or method of production that is successful in the marketplace. However, there is a fundamental difference between spending on new equipment and factories and on R&D. The final product of R&D spending is ideas. These ideas can potentially be copied and used by other companies in the economy. Thus, R&D expenditures have potentially large positive externalities or spillover effects. This means that spending by one company has a positive impact on other companies and increases the overall pool of knowledge available to all companies. Spending by companies on R&D and knowledge capital generates benefits to the economy as a whole that exceed the private benefit to the individual company making the R&D investment. Individual companies cannot fully capture all the benefits associated with creating new ideas and methods of production. Some of the benefits are external to the company, and so are the social returns associated with the investment in R&D and human capital.

This distinction between the private and social returns or benefits to capital is important because it solves an important microeconomic issue. The elimination of the assumption of diminishing marginal returns to capital implies constant returns to capital and increasing returns to all factors taken together. If individual companies could capture these scale economies, then all industries would eventually be dominated by a single company—a monopoly. There is simply no empirical evidence to support this implication. Separating private returns from social returns solves the problem. If companies face constant returns to scale for all private factors, there is no longer an inherent advantage for a company being large. But the externality or social benefit results in increasing returns to scale across the entire economy as companies benefit from the private spending of the other companies.

The role of R&D spending and the positive externalities associated with this spending have important implications for economic growth. In the endogenous growth model, the economy does not reach a steady growth rate equal to the growth of labor plus an exogenous rate of labor productivity growth. Instead, saving and investment decisions can generate self-sustaining growth at a permanently higher rate. This situation is in sharp contrast to the neoclassical model, in which only a transitory increase in growth above the steady state is possible. The reason for this difference is that because of the externalities on R&D, diminishing marginal returns to capital do not set in. The production function in the endogenous growth model is a straight line given by

$$y_e = f(k_e) = ck_e \tag{12}$$

where output per worker (y_e) is proportional to the stock of capital per worker (k_e), c is the (constant) marginal product of capital in the aggregate economy, and the subscript e denotes the endogenous growth model. In contrast, the neoclassical production function is a curved line that eventually flattens out (see Exhibit 4).

To understand the significance of introducing constant returns to aggregate capital accumulation, note that in this model the output-to-capital ratio is fixed (= c) and therefore output per worker (y_e) always grows at the same rate as capital per worker (k_e). Thus, faster or slower capital accumulation translates one for one into faster or slower growth in output per capita. Substituting Equation 12 into Equation 7 gives an equation for the growth rate of output per capita in the endogenous growth model:

$$\Delta y_e/y_e = \Delta k_e/k_e = sc - \delta - n$$

Because all the terms on the right-hand side of this equation are constant, this is both the long-run and short-run growth rate in this model. Examination of the equation shows that *a higher saving rate (s) implies a permanently higher growth rate*. This is the key result of the endogenous growth model.

The positive externalities associated with spending on R&D and knowledge capital suggest that spending by private companies on these inputs may be too low from an overall societal point of view. This is an example of a market failure where private companies under-invest in the production of these goods. In this case, there may be a role for government intervention to correct for the market failure by direct government spending on R&D and/or providing tax breaks and subsidies for private production of knowledge capital. Higher levels of spending on knowledge capital could translate into faster economic growth even in the long run. Finally, according to the endogenous growth theory, there is *no reason why the incomes of developed and developing countries should converge*. Because of constant or even increasing returns associated with investment in knowledge capital, the developed countries can continue to grow as fast as, or faster than, the developing countries. As a result, there is no reason to expect convergence of income over time. We now turn to the convergence debate in more detail.

EXAMPLE 12

Neoclassical vs. Endogenous Growth Models

Consider again an economy with per capita income growing at a constant 3.08 percent rate and with a 20 percent saving rate, an output-to-capital ratio (c in the endogenous growth model, Equation 12) of 0.7040, a depreciation rate (δ) of 10 percent, and a 1 percent labor force growth (n).

1. Use the endogenous growth model to calculate the new steady state growth rate of per capita income if the saving rate increases to 23.5 percent.
2. How much higher will per capita income be in 10 years because of the higher saving rate? How does this compare with the impact calculated in Example 11 using the neoclassical model? What accounts for the difference?
3. In an effort to boost growth, the government is considering two proposals. One would subsidize all private companies that increase their investment spending. The second would subsidize only investments in R&D and/or implementation of new technologies with potential for network externalities. Interpret these proposals in terms of the neoclassical and endogenous growth models and assess their likely impact on growth. (Focus only on "supply-side" considerations here.)

Solution to 1:

In the endogenous growth model the new growth rate of per capita income is

$$\Delta y_e/y_e = sc - \delta - n = (0.235)(0.7040) - 0.10 - 0.01 = 0.0554, \text{ or } 5.54 \text{ percent}$$

This is the same as the growth rate immediately following the increase in the saving rate (to 30 percent in that case) in the earlier example using the neoclassical model (Example 11). Unlike in the neoclassical model, in the endogenous growth model this higher growth rate will be sustained.

Solution to 2:

According to the endogenous growth model, per capita income will grow 2.46 percent (= 5.54% – 3.08%) faster with the higher saving rate. After 10 years, the cumulative impact of the faster growth rate will be

$$\exp(0.0246 \times 10) = \exp(0.246) = 1.2789$$

So, per capita income will be almost 28 percent higher than it would have been at the lower saving rate. This increase is substantially larger than the 15.18 percent cumulative increase after 10 years found in Example 11 assuming a much larger increase in the saving rate (to 30 percent instead of 23.5 percent) in the neoclassical model. The difference arises because the endogenous growth model assumes that capital accumulation is not subject to diminishing returns. Therefore, the growth rate is permanently, rather than temporarily, higher.

Solution to 3:

Subsidizing all private investment would tend to have a significant, pure capital deepening component. That is, companies would be encouraged to buy more, but not necessarily better, plant and equipment. The neoclassical model indicates that this is likely to result in a temporary surge in growth, but even if the higher rate of investment/saving is sustained, growth will again decline over time. On the positive side, this proposal is very likely to succeed, at least for a while, because it does not require investment in unproven technologies or ill-defined network effects. The impact of the other proposal is more uncertain but potentially much more powerful. If the investments in R&D and/or new technologies lead to new knowledge, greater efficiency, new products and methods, and/or network externalities, then the endogenous growth model suggests that growth is likely to be permanently enhanced.

5.4 Convergence Debate

As is evident in Exhibit 1, a wide gap separates the living standards in the developed and developing nations of the world. The question is, Will this difference persist forever or will the per capita income levels of the developing countries converge to those of the developed countries? Convergence means that countries with low per capita incomes should grow at a faster rate than countries with high per capita incomes. Thus, over time the per capita income of developing countries should converge toward that of the developed countries. Whether convergence occurs has major implications for the future growth prospects of developed versus developing countries. It also has important investment implications.

Neoclassical growth theory predicts two types of convergence: absolute convergence and conditional convergence. **Absolute convergence** means that developing countries, regardless of their particular characteristics, will eventually catch up with the developed countries and match them in per capita output. The neoclassical model assumes that all countries have access to the same technology. As a result, per capita income in all countries should eventually grow at the same rate. Thus, the model implies convergence of per capita *growth rates* among all countries. It does not, however, imply that the *level* of per capita income will be the same in all countries regardless of underlying characteristics; that is, it does not imply absolute convergence.

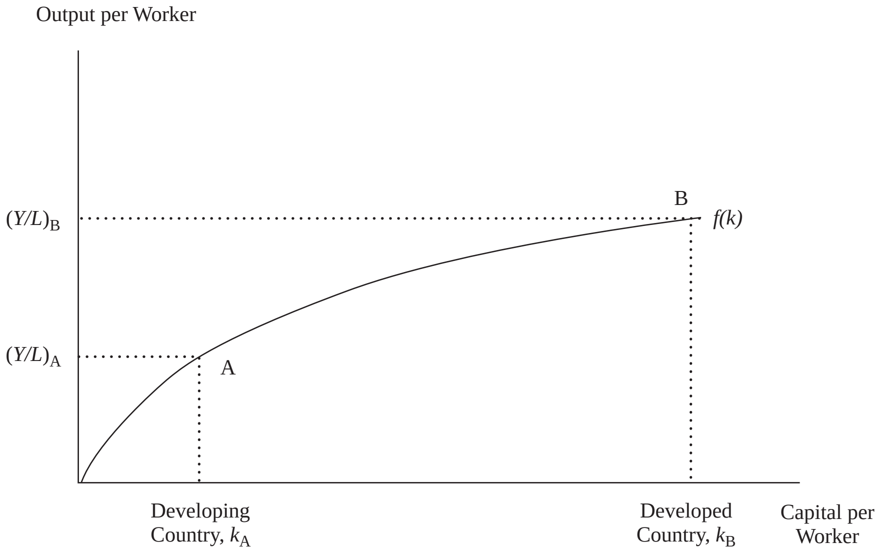
Conditional convergence means that convergence is conditional on the countries having the same saving rate, population growth rate, and production function. If these conditions hold, the neoclassical model implies convergence to the same *level* of per capita output as well as the same steady state growth rate. In terms of Exhibit 13, these economies would have the same k^* and thus the same steady state. If they start with different capital-to-labor ratios, their growth rates will differ in the transition to the steady state. The economy with a lower capital-to-labor ratio will experience more rapid growth of productivity and per capita income, but the differential will diminish until they finally converge. Countries with different saving rates or population growth rates and thus different steady state values for k^* will have different steady state *levels* of per capita income, but their growth rates of per capita output will still converge.

The data (see Exhibit 18) indicate that some of the poorer countries are diverging rather than converging to the income levels of the developed countries. Thus, in addition to the first two convergence concepts, we have the notion of **club convergence**, where only rich and middle-income countries that are members of the club are converging to the income level of the richest countries in the world. This means that the countries with the lowest per capita income in the club grow at the fastest rate. In contrast, countries outside the club continue to fall behind. Poor countries can join the club if they make appropriate institutional changes, such as those summarized in Section 2.7. Finally, countries may fall into a **non-convergence trap** if they do not implement necessary institutional reforms. For example, failure to reform labor markets has undermined growth in some European countries that have experienced weak growth in employment and high rates of unemployment over the last two decades. Certain institutional arrangements that initially enhance growth may later generate non-convergence traps if maintained too long. Import substitution policies enabled the Latin American countries to grow rapidly in the 1950s and 1960s but caused them to stagnate in the 1970s and 1980s.

If convergence, and especially club convergence, does occur, investing in countries with lower per capita incomes that are members of the club should, over long periods of time, provide a higher rate of return than investing in higher-income countries. Convergence means that the rate of growth of potential GDP should be higher in developing countries that have made the institutional changes that are a precondition for growth and that enable these countries to become members of the convergence club. With higher long-term growth in these economies, corporate profits should also grow at a faster rate. Given the faster rate of growth in earnings, stock prices may also rise at a faster rate. Of course, risk is also likely to be higher in these markets. Nonetheless, it is reasonable to conclude that long-term investors should allocate a risk-tolerance-appropriate portion of their assets to those developing economies that have become members of the convergence club.

Convergence between the developed and developing countries can occur in two ways. First, convergence takes place through capital accumulation and capital deepening. Exhibit 17 illustrates the difference between developed and developing countries using the per capita neoclassical production function. The developed countries operate at point B, so increases in capital have almost no impact on productivity. In contrast, developing countries operate at point A, where increases in capital significantly boost labor productivity.

Exhibit 17 Per Capita Production Function Developed vs. Developing Countries



A second source of convergence is that developing countries can imitate or adopt technology already widely utilized in the advanced countries. Developing countries can learn from advanced countries as scientific and management practices spread with globalization. By importing technology from the advanced countries, the developing countries can achieve faster economic growth and converge to the income of the advanced countries. Technology transfers will narrow the income gap between developed and developing countries only if the poor countries invest the resources to master the technology and apply it to their economies. This spending is similar to R&D spending and allows the country to join the convergence club. The steady state rate of growth for members of the convergence club will be determined by the global rate of technological progress. Without such spending, the country will be left out and will continue to fall behind the developed countries.

In contrast to the neoclassical model, the endogenous growth model makes no prediction that convergence should occur. This model allows for countries that start with high per capita income and more capital to grow faster and stay ahead of the developing countries. If the externalities associated with knowledge and human capital are large, the higher-income country can maintain its lead through high rates of investment in these capital inputs.

If the convergence hypothesis is correct, there should be an inverse relation between the initial level of per capita real GDP and the growth rate in per capita GDP. Exhibit 18 shows the countries in Exhibit 1 in descending order of per capita income in 1950. If incomes are converging across countries, the poor countries in 1950 should have a higher growth rate between 1950 and 2010 than the rich countries.

Exhibit 18 Real Per Capita GDP by Selected Economy

	Real GDP Per Capita in Dollars				Average Annual Growth in Real Per Capita GDP (%)			
	1950	**1970**	**1990**	**2010**	**1950–1970**	**1970–1990**	**1990–2010**	**1950–2010**
United States	14,559	22,806	35,328	46,697	2.27	2.21	1.40	1.96
New Zealand	13,795	18,255	22,331	31,223	1.41	1.01	1.69	1.37

Exhibit 18 (Continued)

	Real GDP Per Capita in Dollars				Average Annual Growth in Real Per Capita GDP (%)			
	1950	1970	1990	2010	1950–1970	1970–1990	1990–2010	1950–2010
Australia	13,219	21,444	30,628	45,951	2.45	1.80	2.05	2.10
Canada	12,053	19,919	31,196	41,288	2.54	2.27	1.41	2.07
United Kingdom	11,602	18,002	27,469	37,378	2.22	2.14	1.55	1.97
France	8,266	18,186	28,127	34,358	4.02	2.20	1.01	2.40
Venezuela	8,104	11,590	9,028	10,560	1.81	−1.24	0.79	0.44
Argentina	6,164	9,026	7,952	13,468	1.93	−0.63	2.67	1.31
Italy	5,954	16,522	27,734	31,069	5.24	2.62	0.57	2.79
Ireland	5,496	9,869	18,812	36,433	2.97	3.28	3.36	3.20
Saudi Arabia	5,060	17,292	20,399	22,951	6.34	0.83	0.59	2.55
South Africa	4,361	6,959	6,595	8,716	2.36	−0.27	1.40	1.16
Singapore	4,299	8,600	27,550	56,224	3.53	5.99	3.63	4.38
Mexico	4,180	7,634	10,754	13,710	3.06	1.73	1.22	2.00
Spain	3,964	11,444	21,830	30,504	5.44	3.28	1.69	3.46
Peru	3,464	5,786	4,516	8,671	2.60	−1.23	3.32	1.54
Japan	3,048	15,413	29,813	34,828	8.44	3.35	0.78	4.14
Brazil	2,365	4,324	6,959	9,589	3.06	2.41	1.62	2.36
Turkey	2,327	4,413	7,741	11,769	3.25	2.85	2.12	2.74
Philippines	1,296	2,136	2,660	3,672	2.53	1.10	1.63	1.75
South Korea	1,185	3,009	12,083	30,079	4.77	7.20	4.67	5.54
Egypt	1,132	1,560	3,137	5,306	1.62	3.55	2.66	2.61
Nigeria	814	1,183	1,203	2,037	1.89	0.08	2.67	1.54
Indonesia	804	1,182	2,517	4,740	1.95	3.85	3.22	3.00
Kenya	791	1,113	1,359	1,376	1.72	1.00	0.06	0.93
Vietnam	689	770	1,073	3,369	0.56	1.67	5.89	2.68
Pakistan	666	985	1,645	2,600	1.98	2.60	2.32	2.29
India	658	922	1,390	3,575	1.70	2.07	4.84	2.86
Botswana	449	774	3,731	5,311	2.76	8.18	1.78	4.20
China	402	698	1,677	8,569	2.80	4.48	8.50	5.23
Ethiopia	314	479	462	749	2.13	−0.18	2.45	1.46

The results for the convergence hypothesis are mixed. The economies with the highest per capita income in 1950 were the United States, New Zealand, Australia, and Canada. The markets with the fastest growth rate over the period 1950–2010 were South Korea and China, each growing at a rate above 5 percent. This result strongly supports convergence because the per capita incomes of these economies in 1950 were well below that of the United States. In addition, the results for Japan, Singapore, and Spain showed a convergence to the level of income in the advanced economies. In total, 21 of the 30 economies grew faster than the United States over the period. However, Ethiopia, Kenya, Nigeria, the Philippines, Peru, South Africa, Argentina, Venezuela, and New Zealand fell further behind the United States. Interestingly, since 1990 convergence has been relatively strong overall, with 24 countries (75 percent) growing faster than the United States—including Ethiopia, Nigeria, the Philippines, Peru, South Africa, Argentina, and New Zealand—but has not continued among the most advanced economies as France, Japan, and Italy lagged the United States, Canada, and Australia.

The evidence seems to suggest that poorer countries may converge if they develop the appropriate legal, political, and economic institutions as discussed in Section 2.7. In addition, trade policy is an important factor, which we address in the next section.

6 GROWTH IN AN OPEN ECONOMY

The Solow model discussed in Section 5.2 assumed a closed economy in which domestic investment equals domestic savings and there is no international trade or capital flows. Opening up the economy to trade and financial flows can significantly affect the rate of growth in an economy for the following reasons:

1 A country can borrow or lend funds in global markets, and domestic investment can be funded by global savings. Thus, investment is not constrained by domestic savings.
2 Countries can shift resources into industries in which they have a comparative advantage and away from industries in which they are relatively inefficient, thereby increasing overall productivity.
3 Companies have access to a larger, global market for their products, allowing them to better exploit any economies of scale and increasing the potential reward for successful innovation.
4 Countries can import technology, thus increasing the rate of technological progress.
5 Global trade increases competition in the domestic market, forcing companies to produce better products, improve productivity, and keep costs low.

According to the neoclassical model, convergence should occur more quickly if economies are open and there is free trade and international borrowing and lending. Opening up the economy should increase the rate at which countries' capital-to-labor ratios converge. The dynamic adjustment process can be described as follows:

1 Developing countries have less capital per worker, and as a result, the marginal product of capital is higher. Thus, the rate of return on investments should be higher in countries with low capital-to-labor ratios and lower in countries with high capital-to-labor ratios.
2 Global savers, seeking higher returns on investments, will invest in the capital-poor countries. In an open economy, capital should flow from countries with high capital-to-labor ratios to those that are capital poor.
3 Because of the capital inflows, the physical capital stock in the developing countries should grow more rapidly than in rich countries even if the saving rate is low in the poorer countries. Faster capital growth will result in higher productivity growth, causing per capita incomes to converge.
4 Because capital flows must be matched by offsetting trade flows, capital-poor countries will tend to run a trade deficit as they borrow globally to finance domestic investment. In contrast, the developed countries will tend to run trade surpluses as they export capital.

5 During the transition to the new steady state, the inflows of capital will temporarily raise the rate of growth in the capital-poor country above the steady state rate of growth. At the same time, growth in the capital-exporting countries will be below the steady state.

6 Over time, the physical capital stock will rise in the capital-poor country, reducing the return on investments. As a result, the rate of investment and size of the country's trade deficit will decline. Growth will slow and approach the steady state rate of growth. If investment falls below the level of domestic savings, the country will eventually shift from a trade deficit to a trade surplus and become an exporter of capital.

7 In the Solow model, after the reallocation of world savings, there is no permanent increase in the rate of growth in an economy. Both the developed and developing countries grow at the steady state rate of growth.

In contrast to the neoclassical model, endogenous growth models predict that a more open trade policy will permanently raise the rate of economic growth. In these models, international trade increases global output through the following:

1 A selection effect, where increased competition from foreign companies forces less efficient domestic companies to exit and more efficient ones to innovate and raises the efficiency of the overall national economy.

2 A scale effect that allows producers to more fully exploit economies of scale by selling to a larger market.

3 A backwardness effect arising from less advanced countries or sectors of an economy catching up with the more advanced countries or sectors through knowledge spillovers.

Open trade also affects the innovation process by encouraging higher levels of spending on R&D and on human capital as companies invest to take advantage of access to larger markets and the greater flow of ideas and knowledge among countries. The rate of return to new investment increases, as does the rate of economic growth. In general, most countries gain from open trade, with the scale effect benefiting smaller countries and the backwardness effect benefiting the poorer, less developed countries. But trade can also retard growth in some cases, especially in small countries that lag behind the technology leaders. Opening these countries to trade may discourage domestic innovation because companies will recognize that, even if they innovate, they may lose out to more efficient foreign companies.

EXAMPLE 13

The Entry of China and India into the Global Economy

China and India effectively entered the global economy in the 1980s as they shifted toward more market-oriented policies and opened up to global trade. Their impact on global growth was significant. In 2010, China and India accounted for 13.6 percent and 5.5 percent of world GDP, respectively, whereas the two countries combined for only 4.2 percent of global output in 1980. The entry of these two countries significantly increased the global supply of skilled and unskilled labor receiving relatively lower wages. As a result of the surge in available labor, global potential GDP increased sharply. Economic theory suggests that the supply-side increase in the global capacity to produce goods and services would increase global output and put downward pressure on prices.

The neoclassical model of growth can provide us with some further insights into the impact of China and India entering the global economy. At the time, China and India had relatively lower wages and capital compared to the United States and Europe. One would expect that the rate of return on capital would be higher in China and India and that capital would flow from the developed countries to China and India. Hence, both China and India would be expected to run trade deficits. This has been the case for India but, contrary to the prediction of the model, China has run trade surpluses. These surpluses stem mainly from China's very high domestic saving rate.

Nonetheless, China has experienced large foreign direct investment (see Exhibit 19) inflows, which have reinforced its already high private investment rate. As China and India accumulate capital, their capital-to-labor ratios, real wage levels, and per capita income should converge toward those of the advanced economies. Depending on global aggregate demand conditions, wages might even have to fall in the developed countries in the process of shifting wealth and income to the developing economies. Because of the surge in the global supply of labor, the overall share of labor in global income should decline relative to capital. In addition, global productivity should rise as China and India account for a rising share of global output. In sum, over the long run, the growing share of global GDP going to China and India will benefit the global economy as more efficient utilization of resources allows global potential GDP to grow more rapidly for an extended period.

Although both the neoclassical and endogenous models of growth show the benefits of open markets, over the last 50 years developing countries have pursued two contrasting strategies for economic development:

- *Inward-oriented policies* attempt to develop domestic industries by restricting imports. Instead of importing goods and services, these policies encourage the production of domestic substitutes, despite the fact that it may be more costly to do so. These policies are also called import substitution policies.
- *Outward-oriented policies* attempt to integrate domestic industries with those of the global economy through trade and make exports a key driver of growth.

Many African and Latin American countries pursued inward-oriented policies from the 1950s to the 1980s that resulted in poor GDP growth and inefficient industries producing low-quality goods. In contrast, many East Asian countries, such as Singapore and South Korea, pursued outward-oriented polices during this same time period, which resulted in high rates of GDP growth and convergence with developed countries. These countries also benefited from the positive effects of foreign direct investment, which suggests that more open and trade-oriented economies will grow at a faster rate. The evidence strongly supports this case.

In Example 1, we compared the economic performance of Argentina and Venezuela with that of Japan, South Korea, and Singapore. In 1950, the per capita GDP of the two Latin American countries was well above that of the three East Asian countries. By 2010, however, the per capita GDPs of the three Asian countries was well above those of Argentina and Venezuela. The difference in the growth rates between Argentina and Venezuela and the three Asian countries is explained largely by the openness of their economies. Argentina and Venezuela were relatively closed economies, whereas the Asian countries relied on foreign investment and open markets to fuel growth.

The good news is that many African and Latin American countries have removed trade barriers and are pursuing more outward-oriented policies, and they experienced better growth. Brazil is a good example. Exports of goods and services increased from

$64.6 billion in 2000 to $249.9 billion in 2010, an increase of over 286 percent. As shown in Exhibit 19, exports as a share of GDP rose from 5.2 percent to 10.7 percent over this period.

Exhibit 19 Exports and Foreign Direct Investment of Selected Countries

	1980	1990	2000	2010
Brazil				
Exports as percent of GDP	5.1	4.5	5.2	10.7
Inflows of foreign direct investment ($ billions)	NA	NA	$32.8	$48.5
China				
Exports as percent of GDP	5.9	7.5	9.3	17.2
Inflows of foreign direct investment ($ billions)	NA	NA	$38.4	$185.0
India				
Exports as percent of GDP	2.7	3.5	3.8	7.2
Inflows of foreign direct investment ($ billions)	NA	NA	$3.6	$24.6
Ireland				
Exports as percent of GDP	32.1	59.6	88.3	128.5
Inflows of foreign direct investment ($ billions)	NA	NA	$25.8	$26.3
Mexico				
Exports as percent of GDP	12.8	16.2	15.7	19.1
Inflows of foreign direct investment ($ billions)	NA	NA	$18.0	$19.6
South Africa				
Exports as percent of GDP	NA	11.6	10.7	16.1
Inflows of foreign direct investment ($ billions)	NA	NA	$0.9	$1.56
South Korea				
Exports as percent of GDP	NA	26.3	25.1	35.0
Inflows of foreign direct investment ($ billions)	NA	NA	$9.3	–$0.2
United States				
Exports as percent of GDP	8.8	9.1	10.9	12.7
Inflows of foreign direct investment ($ billions)	NA	NA	$159.2	$351.3

Source: OECD StatLink.

EXAMPLE 14

Why Some Countries Converge and Others Do Not

As evident from the high rates of growth between 1950 and 2010 shown in Exhibit 18, China and South Korea are converging toward the income levels of the advanced economies. In contrast, the economies of Mexico and South Africa have not converged. Using the data in Exhibits 9 and 19, give some reasons why this has occurred.

Solution:

Two reasons largely account for the difference. First, growth in the Chinese and South Korean economies has been driven by high rates of business investment. As shown in Exhibit 9, investment as a share of GDP in 2010 was 48.2 percent in China, almost double the rate of 25.0 percent in Mexico and more than double the rate of 19.3 percent in South Africa. Although investment as a share of GDP in South Korea is lower than in China, it is well above that of Mexico and South Africa.

Second, both China and South Korea have pursued an aggressive export-driven, outward-oriented policy focusing on manufactured goods. In 2010, exports were 35 percent of GDP for South Korea and 17.2 percent of GDP for China (Exhibit 19). In addition, foreign direct investment is a major factor underlying growth in China.

The comparable export numbers for Mexico and South Africa are 19.1 percent and 16.1 percent of GDP, respectively. Despite the North American Free Trade Agreement (NAFTA), Mexico's exports as a share of GDP rose only modestly from 1990 to 2010. In contrast, exports as a share of GDP for China have nearly doubled since 2000. In addition, Mexico and South Africa attracted only a combined $21.2 billion in foreign direct investment in 2010, significantly less than that of Ireland—a smaller but much wealthier and very open country—and the $185 billion inflow of foreign investment into China. The upshot is that Mexico and South Africa have been more inward-oriented economies. These trends are changing, however, as many African and Latin American countries are increasingly relying on growing exports and foreign investment to increase GDP growth.

EXAMPLE 15

Investment Prospects in Spain: Estimating the Sustainable Growth Rate

You are a financial analyst at Global Invest Inc., an investment management firm that runs a number of global mutual funds with a significant exposure to Spain. The Madrid General Index, which reached a crisis-induced low of 716 in March 2009, remains far below its November 2007 peak of 1725. The members of the investment policy committee at the firm believe the equity market in Spain is attractive and is currently being depressed by temporary problems in the banking and real estate markets of Spain, which they feel are overstated. They believe that higher profits will ultimately drive the market higher but are concerned about the long-term prospects and the sustainable rate of growth for Spain. One of the research assistants at the firm gathers the data shown in Exhibit 20 from the OECD and the Conference Board.

Exhibit 20 Growth Data for Spain

	GDP in Billions of USD Adjusted for PPP	Gross Capital Spending as Percentage of GDP	Consumption of Fixed Capital (percent of GDP)	Labor Cost as Percentage of Total Factor Cost	Total Hours Worked (millions)	Output per Hour Worked in 2009 USD Adjusted for PPP	Growth in Total Factor Productivity (%)
2000	1,156.4	26.3	13.6	64.40	28,402	40.7	–0.87
2001	1,198.6	26.4	13.7	63.66	29,232	41.0	–0.78
2002	1,231.0	26.6	14.1	62.97	29,836	41.3	–0.57
2003	1,269.1	27.4	14.4	62.69	30,495	41.6	–0.21
2004	1,310.6	28.3	14.9	61.71	31,274	41.9	–0.58
2005	1,357.9	29.5	15.3	60.87	32,132	42.3	–0.65
2006	1,412.5	31.0	15.6	60.66	33,146	42.6	–0.28
2007	1,462.9	30.9	15.7	60.04	33,757	43.3	–0.07
2008	1,475.6	29.1	16.2	60.23	33,830	43.6	–1.63
2009	1,420.6	24.4	16.9	59.47	31,705	44.8	–1.61

Sources: OECD Stat Extracts and the Conference Board Total Economy Database.

From the Conference Board website, the physical capital stock for Spain was estimated at \$2,177.2 billion (adjusted for purchasing power parity) in 1999. The research analyst calculated the physical capital stock (K) for Spain for the years 2000–2009 using the following equation:

$$K_t = K_{t-1} + I - D$$

where I is gross investment or gross capital spending and D is the depreciation or the consumption of fixed capital. So for 2000 and 2001, the physical capital stock is calculated as:

$$K_{2000} = \$2,177.2 + \$1,156.4\ (0.263 - 0.136) = \$2,324.1 \text{ billion}$$

$$K_{2001} = \$2,324.1 + \$1,198.6\ (0.264 - 0.137) = \$2,476.3 \text{ billion}$$

The physical capital stock for the remaining years is calculated in the same way and given by Exhibit 21.

Exhibit 21 Estimated Physical Capital Stock (USD billions)

2000	\$2,324.1
2001	2,476.3
2002	2,630.2
2003	2,795.1
2004	2,970.8
2005	3,163.6
2006	3,381.1
2007	3,603.5
2008	3,793.8
2009	3,900.4

You are requested by the investment policy committee to use the above data to address the following:

1 Calculate the potential growth rate of the Spanish economy using the production function or growth accounting method (Equation 4), and determine the amount of growth attributed to each source.
2 Calculate the potential growth rate of the Spanish economy using the labor productivity method (Equation 5).
3 How significant are capital deepening and technology in explaining growth for Spain?
4 What is the steady state growth rate for Spain according to the neoclassical model?
5 Assess the implications of the growth analysis for future economic growth and equity prices in Spain.

Solution to 1:

The production function or growth accounting method estimates the growth in GDP using Equation 4:

$$\text{Growth in potential GDP} = \alpha\Delta K/K + (1-\alpha)\Delta L/L + \Delta A/A$$

The annual growth rate in capital is calculated from Exhibit 21 as[a]

$$(3{,}900.4/2{,}324.1)^{1/9} - 1 = 5.92\%$$

The labor input is measured by the growth rate in total hours worked in the economy (Exhibit 20) and given by

$$(31{,}705/28{,}402)^{1/9} - 1 = 1.23\%$$

The growth rate in total factor productivity (Exhibit 20) is calculated by using a geometric average of the growth rates for 2000–2009 and is equal to –0.73 percent. Finally, the labor share of output is given by the average of the labor cost as a percentage of total factor cost, which is 61.7 percent for 2000–2009 (Exhibit 20). Thus, the share of capital (α) is 1 – 0.617 = 38.3%.

Using these numbers, the growth in potential GDP is

$$\begin{aligned}\text{Growth in potential GDP} &= \alpha\Delta K/K + (1-\alpha)\Delta L/L + \Delta A/A \\ &= (0.383)0.0592 + (0.617)0.0123 + (-0.0073) \\ &= 2.30\%\end{aligned}$$

Sources of growth for Spain over the period 2000–2009 were:

Capital	(0.383) × (0.0592) = 2.27%
Labor	(0.617) × (0.0123) = 0.76%
TFP	= –0.73%

Solution to 2:

The labor productivity method estimates the growth in GDP using Equation 5:

$$\begin{aligned}\text{Growth rate in potential GDP} = {} & \text{Long-term growth rate of labor force} \\ & + \text{Long-term growth rate in labor productivity}\end{aligned}$$

As before, we use the growth in total hours worked to measure the growth in the labor force. The growth in labor productivity per hour worked is

$$(44.8/40.7)^{1/9} - 1 = 1.07\%$$

$$\text{Growth in potential GDP} = 1.23\% + 1.07\% = 2.3\%$$

Note that the estimate of potential GDP growth using the labor productivity approach is the same as that obtained from the growth accounting method. In general, the two methods are likely to give somewhat different estimates because they rely on different data inputs. The growth accounting method requires measurements of the physical capital stock and TFP. As discussed in Section 4.3, TFP is estimated using various time-series or econometric models of the component of growth that is not accounted for by the explicit factors of production. As a result, the estimate of TFP reflects the average (or "smoothed") behavior of the growth accounting residual. The labor productivity approach is simpler, and it avoids the need to estimate the capital input and TFP. In contrast to the estimated value of TFP, labor productivity is measured as a pure residual; that is, it is the part of GDP growth that is not explained by the labor input (and only the labor input). The cost of the simplification is that the labor productivity approach does not allow a detailed analysis of the drivers of productivity growth.

Solution to 3:

Capital deepening occurs in an economy when there is an increase in the capital-to-labor ratio. The labor input for Spain is measured in terms of total hours worked in the economy. Thus, the capital-to-labor ratio for Spain is calculated by dividing the physical capital stock in Exhibit 21 by total hours worked in Exhibit 20. The results, shown in Exhibit 22, indicate that capital deepening was very significant in Spain: The amount of capital per hour worked increased from \$81.83 in 2000 to \$123.02 in 2009. In terms of the growth rate, the capital-to-labor ratio increased at an annual rate of 4.6 percent.

The contribution of TFP is measured by the growth in total factor productivity. In contrast to capital deepening, TFP made a negative contribution to growth; the average rate of growth for TFP from 2000 to 2009 was –0.73 percent. However, TFP is estimated using various statistical techniques, and given the uncertainty around these estimates, it should be viewed with some caution.

Exhibit 22	Estimated Capital-to-Labor Ratio ($ millions/hour worked)
2000	$81.83
2001	84.71
2002	88.16
2003	91.66
2004	94.99
2005	98.46
2006	102.01
2007	106.75
2008	112.14
2009	123.02

Solution to 4:

The steady state growth rate in the neoclassical model is estimated by (see Equation 8):

$\Delta Y/Y = (\theta)/(1 - \alpha) + n$ = Growth rate of TFP scaled by labor factor share + Growth rate in the labor force

Steady state growth rate = –0.73%/(1 – 0.383) + 1.23% = 0.05%

As expected, the growth rate in potential GDP (calculated as in the solutions to 1 and 2) is above the steady state growth rate. The reason for this is that the economy of Spain is still in the process of converging to the higher income levels of the United States and the major economies in Europe. The physical capital stock is below the steady state, and capital deepening is a significant factor increasing productivity growth and the growth in potential GDP. Steady state growth may be somewhat underestimated in our analysis given that TFP growth is likely to revert to the 1 percent annual rate of increase exhibited in other major developed economies. This is likely to be offset by a lower growth rate in the labor input (see Example 6).

Solution to 5:

The results suggest that potential GDP growth in Spain is approximately 2.3 percent. As we saw in Exhibit 1, the growth rate of actual GDP since early 2000 has been 2.1 percent per year, close to the previous estimate of potential but well above the steady state. The problem is that all the growth in potential GDP is due to the increase in the labor and capital inputs, with capital deepening being very significant as the capital-to-labor ratio is increasing at a 4.6 percent annual rate. The neoclassical model would suggest that the impact of capital deepening will decline over time and the economy will move toward a steady state rate of growth. Thus, growth based on capital deepening should not be sustainable over time. The other major question raised is whether the labor input can continue to grow at an annual rate of 1.2 percent. We examined this question in Example 6. In sum, potential GDP growth is likely to fall over time given Spain's reliance on capital deepening and the strong possibility that growth in the labor input is likely to slow. However, the reversion of TFP growth to levels more typical of other European economies should mitigate the decline. Even if TFP does rebound, slower growth in potential GDP in Spain will likely restrain future stock price increases.

[a] Using the 1999 capital stock as a base instead of the 2000 capital stock would give almost the same growth rate: $(3{,}900.4/2{,}177.2)^{1/10} - 1 =$ 6.00%.

SUMMARY

This reading focuses on the factors that determine the long-term growth trend in the economy. As part of the development of global portfolio equity and fixed-income strategies, investors must be able to determine both the near-term and the sustainable rates of growth within a country. Doing so requires identifying and forecasting the factors that determine the level of GDP and that determine long-term sustainable trends in economic growth.

- The sustainable rate of economic growth is measured by the rate of increase in the economy's productive capacity or potential GDP.
- Growth in real GDP measures how rapidly the total economy is expanding. Per capita GDP, defined as real GDP divided by population, measures the standard of living in each country.
- The growth rate of real GDP and the level of per capita real GDP vary widely among countries. As a result, investment opportunities differ by country.

- Equity markets respond to anticipated growth in earnings. Higher sustainable economic growth should lead to higher earnings growth and equity market valuation ratios, all other things being equal.
- The best estimate for the long-term growth in earnings for a given country is the estimate of the growth rate in potential GDP.
- In the long run, the growth rate of earnings cannot exceed the growth in potential GDP. Labor productivity is critical because it affects the level of the upper limit. A permanent increase in productivity growth will raise the upper limit on earnings growth and should translate into faster long-run earnings growth and a corresponding increase in stock price appreciation.
- For global fixed-income investors, a critical macroeconomic variable is the rate of inflation. One of the best indicators of short- to intermediate-term inflation trends is the difference between the growth rate of actual and potential GDP.
- Capital deepening, an increase in the capital-to-labor ratio, occurs when the growth rate of capital (net investment) exceeds the growth rate of labor. In a graph of output per capita versus the capital-to-labor ratio, it is reflected by a move along the curve (i.e., the production function).
- An increase in total factor productivity (TFP) causes a proportional upward shift in the entire production function.
- One method of measuring sustainable growth uses the production function and the growth accounting framework developed by Solow. It arrives at the growth rate of potential GDP by estimating the growth rates of the economy's capital and labor inputs plus an estimate of total factor productivity.
- An alternative method measures potential growth as the long-term growth rate of the labor force plus the long-term growth rate of labor productivity.
- The forces driving economic growth include the quantity and quality of labor and the supply of non-ICT and ICT capital, public capital, raw materials, and technological knowledge.
- The labor supply is determined by population growth, the labor force participation rate, and net immigration. The physical capital stock in a country increases with net investment. The correlation between long-run economic growth and the rate of investment is high.
- Technological advances are discoveries that make it possible to produce more or higher quality goods and services with the same resources or inputs. Technology is a major factor determining TFP. TFP is the main factor affecting long-term, sustainable economic growth rates in developed countries and also includes the cumulative effects of scientific advances, applied research and development, improvements in management methods, and ways of organizing production that raise the productive capacity of factories and offices.
- Total factor productivity, estimated using a growth accounting equation, is the residual component of growth once the weighted contributions of all explicit factors (e.g., labor and capital) are accounted for.
- Labor productivity is defined as output per worker or per hour worked. Growth in labor productivity depends on capital deepening and technological progress.
- The academic growth literature is divided into three theories —the classical view, the neoclassical model, and the new endogenous growth view.
- In the classical model, growth in per capita income is only temporary because an exploding population with limited resources brings per capita income growth to an end.

- In the neoclassical model, a sustained increase in investment increases the economy's growth rate only in the short run. Capital is subject to diminishing marginal returns, so long-run growth depends solely on population growth, progress in TFP, and labor's share of income.
- The neoclassical model assumes that the production function exhibits diminishing marginal productivity with respect to any individual input.
- The point at which capital per worker and output per worker are growing at equal, sustainable rates is called the steady state or balanced growth path for the economy. In the steady state, total output grows at the rate of labor force growth plus the rate of growth of TFP divided by the elasticity of output with respect to labor input.
- The following parameters affect the steady state values for the capital-to-labor ratio and output per worker: saving rate, labor force growth, growth in TFP, depreciation rate, and elasticity of output with respect to capital.
- The main criticism of the neoclassical model is that it provides no quantifiable prediction of the rate or form of TFP change. TFP progress is regarded as exogenous to the model.
- Endogenous growth theory explains technological progress within the model rather than treating it as exogenous. As a result, self-sustaining growth emerges as a natural consequence of the model and the economy does not converge to a steady state rate of growth that is independent of saving/investment decisions.
- Unlike the neoclassical model, where increasing capital will result in diminishing marginal returns, the endogenous growth model allows for the possibility of constant or even increasing returns to capital in the aggregate economy.
- In the endogenous growth model, expenditures made on R&D and for human capital may have large positive externalities or spillover effects. Private spending by companies on knowledge capital generates benefits to the economy as a whole that exceed the private benefit to the company.
- The convergence hypothesis predicts that the rates of growth of productivity and GDP should be higher in the developing countries. Those higher growth rates imply that the per capita GDP gap between developing and developed economies should narrow over time. The evidence on convergence is mixed.
- Countries fail to converge because of low rates of investment and savings, lack of property rights, political instability, poor education and health, restrictions on trade, and tax and regulatory policies that discourage work and investing.
- Opening an economy to financial and trade flows has a major impact on economic growth. The evidence suggests that more open and trade-oriented economies will grow at a faster rate.

REFERENCES

Ashauer, David. 1990. "Why Is Infrastructure Important?" In *Is There a Shortfall in Public Capital Investment?* Edited by Alicia Munnell. Federal Reserve Bank of Boston Conference Series No. 34.

Denison, Edward. 1985. *Trends in American Growth*. Washington, DC: Brookings Institution.

Heston, Alan, Robert Summers, and Bettina Aten. 2009. Penn World Table Version 6.3. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania (August).

IMF. 2011. "Statistical Appendix." In *World Economic Outlook* (September). Washington, DC: International Monetary Fund.

Jorgenson, Dale. 1966. "Technology in Growth Theory." Technology and Growth, Federal Reserve Bank of Boston Conference Series.

Jorgenson, Dale. 2000. "Raising the Speed Limit: US Economic Growth in the Information Age." *Brooking Papers on Economic Activity*.

Levine, R. 2005. "Finance and Growth: Theory and Evidence." In *Handbook of Economic Growth*. Edited by Philippe Aghion and Steven Durlauf. Amsterdam: Elsevier, B.V.

Mankiw, Gregory. 1995. "The Growth of Nations." *Brookings Papers on Economic Activity*.

OECD. 2003. *The Sources of Economic Growth in the OECD Countries*. Paris: Organisation for Economic Co-Operation and Development.

Romer, Paul. 1986. "Increasing Returns and Long-Run Growth." *Journal of Political Economy*, vol. 94, no. 5 (October):1002–1037.

Solow, Robert. 1957. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics*, vol. 39, no. 3 (August):312–320.

Stewart, Scott, Christopher Piros, and Jeffrey Heisler. 2011. *Running Money: Professional Portfolio Management*. McGraw-Hill/Irwin.

PRACTICE PROBLEMS

The following information refers to Questions 1–6

Hans Schmidt, CFA, is a portfolio manager with a boutique investment firm that specializes in sovereign credit analysis. Schmidt's supervisor asks him to develop estimates for GDP growth for three countries. Information on the three countries is provided in Exhibit 1.

Exhibit 1 Select Economic Data for Countries A, B, and C

Country	Economy	Capital per Worker
A	Developed	High
B	Developed	High
C	Developing	Low

After gathering additional data on the three countries, Schmidt shares his findings with colleague, Sean O'Leary. After reviewing the data, O'Leary notes the following observations:

Observation 1 The stock market of Country A has appreciated considerably over the past several years. Also, the ratio of corporate profits to GDP for Country A has been trending upward over the past several years and is now well above its historical average.

Observation 2 The government of Country C is working hard to bridge the gap between its standard of living and that of developed countries. Currently, the rate of potential GDP growth in Country C is high.

Schmidt knows that a large part of the analysis of sovereign credit is to develop a thorough understanding of what the potential GDP growth rate is for a particular country and the region in which the country is located. Schmidt is also doing research on Country D for a client of the firm. Selected economic facts on Country D are provided in Exhibit 2.

Exhibit 2 Select Economic Facts for Country D

- Slow GDP Growth
- Abundant Natural Resources
- Developed Economic Institutions

Prior to wrapping up his research, Schmidt schedules a final meeting with O'Leary to see if he can provide any other pertinent information. O'Leary makes the following statements to Schmidt:

Statement 1 Many countries that have the same population growth rate, savings rate, and production function will have growth rates that converge over time.

Statement 2 Convergence between countries can occur more quickly if economies are open and there is free trade and international borrowing and lending; however, there is no permanent increase in the rate of growth in an economy from a more open trade policy.

1 Based upon Exhibit 1, the factor that would *most likely* have the greatest positive impact on the per capita GDP growth of Country A is:

A free trade.

B technology.

C saving and investment.

2 Based upon Observation 1, in the long run the ratio of profits to GDP in Country A is *most likely* to:

A remain near its current level.

B increase from its current level.

C decrease from its current level.

3 Based upon Observation 2, Country C is *most likely* to have:

A relatively low real asset returns.

B a relatively low real interest rate.

C a relatively high real interest rate.

4 Based upon Exhibit 2, the *least likely* reason for the current pace of GDP growth in Country D is:

A a persistently strong currency.

B strong manufacturing exports.

C strong natural resource exports.

5 The type of convergence described by O'Leary in Statement 1 is *best* described as:

A club convergence.

B absolute convergence.

C conditional convergence.

6 Which of the following growth models is *most* consistent with O'Leary's Statement 2?

A Classical

B Endogenous

C Neoclassical

The following information relates to Questions 7–15

Victor Klymchuk, the chief economist at ECONO Consulting (EC), is reviewing the long-term GDP growth of three countries. Klymchuk is interested in forecasting the long-term change in stock market value for each country. Exhibit 1 presents current country characteristics and historical information on selected economic variables for the three countries.

Exhibit 1 Select Country Factors and Historical Economic Data

		2000–2010			
	Country Factors	**Growth in Hours Worked (%)**	**Growth in Labor Productivity (%)**	**Growth in TFP (%)**	**Growth in GDP (%)**
Country A	■ High level of savings and investment ■ Highly educated workforce ■ Low tariffs on foreign imports ■ Limited natural resources	0.9	2.4	0.6	3.3
Country B	■ Developed financial markets ■ Moderate levels of disposable income ■ Significant foreign direct and indirect investments ■ Significant natural resources	–0.3	1.6	0.8	1.3
Country C	■ Politically unstable ■ Limited property rights ■ Poor public education and health ■ Significant natural resources	1.8	0.8	–0.3	2.6

Klymchuk instructs an associate economist at EC to assist him in forecasting the change in stock market value for each country. Klymchuk reminds the associate:

Statement 1 "Over short time horizons, percentage changes in GDP, the ratio of earnings to GDP, and the price-to-earnings ratio are important factors for describing the relationship between economic growth and stock prices. However, I am interested in a long-term stock market forecast."

A client is considering investing in the sovereign debt of Country A and Country B and asks Klymchuk his opinion of each country's credit risk. Klymchuk tells the client:

Statement 2 "Over the next 10 years, I forecast higher potential GDP growth for Country A and lower potential GDP growth for Country B. The capital per worker is similar and very high for both countries, but per capita output is greater for Country A."

The client tells Klymchuk that Country A will offer 50-year bonds and that he believes the bonds could be a good long-term investment given the higher potential GDP growth. Klymchuk responds to the client by saying:

Statement 3 After the next 10 years, I think the sustainable rate of economic growth for Country A will be affected by a growing share of its population over the age of 65, a declining percentage under age 16, and minimal immigration."

The client is surprised to learn that Country C, a wealthy, oil-rich country with significant reserves, is experiencing sluggish economic growth and asks Klymchuk for an explanation. Klymchuk responds by stating:

Statement 4 "While countries with access to natural resources are often wealthier, the relationship between resource abundance and economic growth is not clear. My analysis shows that the presence of a dominant natural resource (oil) in Country C is constraining growth. Interestingly, Country A has few natural resources, but is experiencing a strong rate of increase in per capita GDP growth."

Klymchuk knows that growth in per capita income cannot be sustained by pure capital deepening. He asks the associate economist to determine how important capital deepening is as a source of economic growth for each country. Klymchuk instructs the associate to use the data provided in Exhibit 1.

Klymchuk and his associate debate the concept of convergence. The associate economist believes that developing countries, irrespective of their particular characteristics, will eventually equal developed countries in per capita output. Klymchuk responds as follows:

Statement 5 "Poor countries will only converge to the income levels of the richest countries if they make appropriate institutional changes."

7 Based upon the country factors provided in Exhibit 1, the country *most likely* to be considered a developing country is:

A Country A.

B Country B.

C Country C.

8 Based upon Exhibit 1, capital deepening as a source of growth was *most* important for:

A Country A.

B Country B.

C Country C.

9 Based upon Statement 1, over the requested forecast horizon, the factor that will *most likely* drive stock market performance is the percentage change in:

A GDP.

B the earnings to GDP ratio.

C the price-to-earnings ratio.

10 Based solely on the predictions in Statement 2, over the next decade Country B's sovereign credit risk will *most likely*:
   A increase.
   B decrease.
   C not change.

11 Based upon Statement 2, the difference in per capita output between Country A and Country B is *most likely* due to differences in:
   A capital deepening.
   B capital per worker.
   C total factor productivity.

12 Based upon Statement 3, after the next 10 years the growth rate of potential GDP for Country A will *most likely* be:
   A lower.
   B higher.
   C unchanged.

13 Based upon Statement 4 and Exhibit 1, the sluggish economic growth in Country C is *least likely* to be explained by:
   A limited labor force growth.
   B export driven currency appreciation.
   C poorly developed economic institutions.

14 Based upon Statement 4, the higher rate of per capita income growth in Country A is *least likely* explained by the:
   A rate of investment.
   B growth of its population.
   C application of information technology.

15 The type of convergence described by Klymchuk in Statement 5 is *best* described as:
   A club convergence.
   B absolute convergence.
   C conditional convergence.

The following information relates to Questions 16–21

At a recent international finance and economics conference in Bamako, Mali, Jose Amaral of Brazil and Lucinda Mantri of India are discussing how to spur their countries' economic growth. Amaral believes that growth can be bolstered by removing institutional impediments and suggests several possibilities for Brazil: launching a rural literacy program, clarifying property rights laws, and implementing a new dividend tax on foreign investors.

Mantri responds that for India capital deepening will be more effective and has proposed the following ideas: building a group of auto and textile factories in the southern states, developing a north–south and east–west highway network, and sponsoring a patent initiative.

In response, Amaral says to Mantri:

> "Based on endogenous growth theory, one of those proposals is more likely to raise total factor productivity than result in pure capital deepening."

While Mantri recognizes that India lacks the significant natural resources of Brazil, she states that India can overcome this challenge by bolstering long-term growth through three channels:

Channel 1	Deepening the capital base
Channel 2	Making investments in technology
Channel 3	Maintaining a low Rupee exchange rate

Each country's basic economic statistics were presented at the conference. Selected data for Brazil and India are presented in Exhibit 1. Adama Kanté, a fund manager based in Mali, is planning to increase the fund's allocation to international equities, and after some preliminary analysis, has determined the new allocation will be to Brazilian or Indian equities. After reviewing the data in Exhibit 1, Kanté decides that the allocation will be to Indian equities.

Exhibit 1 Economic Statistics, Brazil and India

Economic Statistic	Brazil	India
GDP/capita, 2010	$9,589	$3,575
GDP/capita Growth, 1990–2010	1.62%	4.84%
GDP Growth, 2005–2008	4.9%	8.2%
- Growth due to Labor Productivity Component	2.9%	6.0%
- Growth due to Capital Deepening Component	3.4%	3.6%

Kanté is concerned about the low standard of living in Mali and its large informal sector. To improve per capita GDP, Kanté is considering five specific strategies:

Strategy 1	Lower the country's tax rate.
Strategy 2	Introduce policies that encourage the return of highly-educated Malian emigrants.
Strategy 3	Build day care centers to permit greater participation of women in the workforce.
Strategy 4	Impose high tariffs on imports to protect the country's nascent industries.
Strategy 5	Use economic development bank loans to improve the country's transport and manufacturing infrastructure.

16 Which of Amaral's initiatives is *least likely* to achieve his stated growth objective?

A Dividend tax

B Rural literacy

C Property rights

17 Which proposal for India is Amaral *most likely* referring to in his response to Mantri?

A Patent initiative

B Highway network

C Auto and textile factories

18 The channel that is *least likely* to help India overcome its challenge of lacking significant natural resources is:

A Channel 1.

B Channel 2.

C Channel 3.

19 Based upon Exhibit 1, which Indian economic statistic *least likely* supports Kanté's international equity allocation preference?

A GDP per capita

B Growth due to labor productivity

C Growth due to capital deepening

20 The strategy that is *least likely* to improve per capita GDP in Mali is:

A Strategy 1.

B Strategy 2.

C Strategy 3.

21 Which of the following strategies being considered by Kanté is *most likely* to undermine or delay convergence with developed economies?

A Strategy 2

B Strategy 4

C Strategy 5

SOLUTIONS

1 B is correct. Country A is a developed country with a high level of capital per worker. Technological progress and/or more intensive use of existing technology can help developed countries increase productivity and thereby increase per capita GDP. Most developed countries have reasonably low trade barriers; thus, somewhat freer trade is likely to have only an incremental, and probably transitory, impact on per capita GDP growth. Also, since the country already has a high capital-to-labor ratio, increased saving/investment is unlikely to increase the growth rate substantially unless it embodies improved technology.

2 C is correct. The ratio of profits to GDP for Country A has been trending upward over the past several years, and is now well above its historical average. The ratio of profits to GDP cannot rise forever. At some point stagnant labor income would make workers unwilling to work without an increase in wages and would also undermine demand, making further expansion of profit margins unsustainable. Thus, it is likely that the ratio of profits to GDP will decline in the long run toward its historical average.

3 C is correct. A high growth rate of potential GDP would cause real incomes to rise more rapidly and also translate into higher real interest rates and higher expected/required real asset returns. The real interest rate is essentially the real return that consumers/savers demand in exchange for postponing consumption. Faster growth in potential GDP means that consumers expect their real income to rise more rapidly. This implies that an extra unit of future income/consumption is less valuable than it would be if income were expected to grow more slowly. All else the same, the real interest rate will have to be relatively high in order to induce the savings required to fund required/desired capital accumulation.

4 B is correct. Country D is a country with abundant resources and has developed the economic institutions necessary for growth, yet the country is experiencing slow economic growth. It is likely that Country D is experiencing the Dutch Disease, where currency appreciation driven by strong export demand for natural resources makes other segments of the economy, in particular manufacturing, globally uncompetitive. Strong manufacturing exports would indicate that Country D is globally competitive and likely to have adopted leading edge technology. Thus, it is unlikely that the slow growth reflects inability to maintain productivity growth. Similarly, strong exports would suggest adequate demand for its products. Thus, strong exports are unlikely to be the cause of slow growth.

5 C is correct. Conditional convergence means that convergence is conditional on the countries having the same savings rate, population growth rate, and production function. If these conditions hold, the neoclassical model implies convergence to the same *level* of per capita output as well as the same steady state growth rate.

6 C is correct. According to the neoclassical model, convergence should occur more quickly if economies are open and there is free trade and international borrowing and lending. Opening up the economy should increase the rate at which the capital-to-labor ratio converges among countries. However, in the neoclassical Solow model, after the reallocation of world savings, there is no permanent increase in the rate of growth in an economy. Both the developed and developing countries eventually grow at the same steady-state rate.

7 C is correct. Country C is the most likely to be a developing economy. Political instability, limited property rights, and poor public education and health are all factors that limit economic growth and thereby contribute to a relatively low standard of living.

8 A is correct. The associate economist can measure the effect of pure capital deepening by measuring the difference of the growth rates of labor productivity and total factor productivity (TFP). The larger the difference, the more important capital deepening is as a source of economic growth. From 2000–2010, Country A's labor productivity grew by 2.4% per year, of which 0.6% came from TFP growth and 1.8% from capital deepening (2.4% – 0.6% = 1.8%).

9 A is correct. In the long run, the growth rate of GDP is the most important driver of stock market performance. Therefore, the associate economist should focus on the drivers of long-run potential GDP growth. The ratio of earnings to GDP cannot increase indefinitely since that would imply that profit would eventually absorb all of GDP. This ratio cannot shrink forever either since unprofitable companies will go out of business. Thus, the annualized growth rate of the earnings to GDP ratio must be approximately zero over long time horizons, and this ratio should not be a dominant factor in forecasting long-term stock market performance. Similarly, the price-to-earnings ratio cannot grow or contract at a finite rate forever because investors will not pay an excessive price for each dollar of earnings, nor will they give away earnings for free. Therefore the rate of change in the price-to-earnings ratio must be approximately zero over long time horizons and should not be a dominant factor in the forecast of long-term stock market performance.

10 A is correct. Credit rating agencies consider the growth rate of potential GDP when evaluating the credit risk of sovereign debt. The chief economist's expectation for lower potential GDP growth for Country B over the next decade increases the perceived credit risk of its sovereign bonds.

11 C is correct. The higher per capita output for Country A is most likely due to differences in the cumulative impact of technological progress embodied in total factor productivity. Technological progress raises the productive capacity of a country. Technological progress causes an upward shift in the entire production function, resulting in higher output per worker for a given level of capital per worker.

12 A is correct. Demographic factors can positively or negatively contribute to a country's sustainable rate of economic growth. After the next 10 years, Country A is expected to experience a growing share of the population over the age of 65 and a declining percentage of the population under the age of 16. All else the same, this implies slower growth of the labor force and hence slower growth of potential GDP. Immigration could offset these demographic challenges. However, Statement 3 indicates that Country A is expected to experience minimal immigration.

13 A is correct. Country C is an example of a country endowed with an abundant natural resource yet experiencing slow economic growth. While labor force growth is an important source of economic growth, it is the least likely explanation of the sluggish economic growth in Country C. As shown in Exhibit 1, growth in total hours worked has accounted for most of Country C's growth. Furthermore, export driven currency appreciation and poorly developed economic institutions are both likely causes of sluggish growth in countries with abundant natural resources.

14 B is correct. Population growth can increase the growth rate of the overall economy, but does not impact the rate of increase in *per capita* GDP. Therefore, population growth does not explain Country A's higher rate of per capita income growth. An increase in labor force participation could, however, raise the growth of per capita GDP.

15 A is correct. Klymchuk is referring to the concept of club convergence. The basic premise is that lower income members of the club are converging to the income levels of the richest countries. This implies that the countries with the lowest per capita income in the club grow at the fastest rate. Countries outside the club, however, continue to fall behind.

16 A is correct. Amaral's initiative to implement a new dividend tax is likely to impede inflows of equity capital by making equity investment in Brazil less attractive for foreign investors. Capital flows, or lack thereof, have a major impact on economic growth because, in an open economy, world savings can finance domestic investment. As a potential source of funds, foreign investment breaks the vicious cycle of low income, low domestic savings, and low investment.

17 A is correct. Mantri's proposal to sponsor a patent initiative, which is likely to result in technology investment and improvement, is likely to cause a proportional upward shift in the entire production function, allowing the economy to produce higher output per worker for a given level of capital per worker. Technological progress also increases the marginal product of capital relative to its marginal cost.

18 C is correct. Maintaining a low currency exchange rate is a policy aimed at maintaining demand for the country's exports. It would have little direct impact on the potential growth rate of aggregate supply. It might boost long-term capacity growth indirectly, however, by encouraging adoption of leading edge technology. Nonetheless, it would not be expected to be as powerful as capital deepening and/or investment in technology.

19 A is correct. Kanté's decision to invest in equities in India is supported by the country's strong economic growth. For global investors, economic growth is important since equity composite valuations depend to a great extent on both the level of economic output (GDP per capita and GDP overall) and on the rate of economic growth. Relative to Brazil, the growth rate in per capita GDP has been much higher, and furthermore, the growth rate in GDP has also been much higher than that of Brazil. In contrast to the growth rate, the relatively low *level* of GDP per capita in India is less likely to indicate attractive equity investment opportunities. Low per capita GDP suggests that India may lack sufficient industrial and financial infrastructure to support some types of industries. It also indicates that domestic purchasing power is relatively limited, decreasing the potential for higher-margin, domestically-oriented businesses.

20 A is correct. With Mali's low standard of living, i.e., GDP per capita and large informal workforce, the tax rate is unlikely to be an impediment to growth, so lowering the tax rate is not likely to be a major contributor to growth.

21 B is correct. The strategy for Mali to impose high tariffs (trade restrictions) on imports is likely to undermine rather than enhance growth and therefore is not supportive of convergence with developed economies. Freer trade (fewer trade restrictions) tends to enhance growth by, for example, inducing a shift of resources into industries in which the country has a comparative thereby increasing overall productivity; forcing less efficient domestic companies to exit and more efficient ones to innovate; allowing domestic producers to more fully exploit economies of scale by selling to a larger market; and enabling less advanced sectors of an economy to catch up with more advanced countries or sectors through knowledge spillovers.

READING

12

Economics of Regulation

by Chester S. Spatt, PhD

Chester S. Spatt, PhD, is at Tepper School of Business at Carnegie Mellon University (USA).

LEARNING OUTCOMES

Mastery	The candidate should be able to:
☐	**a.** describe the economic rationale for regulatory intervention;
☐	**b.** explain the purposes of regulating commerce and financial markets;
☐	**c.** describe anticompetitive behaviors targeted by antitrust laws globally and evaluate the antitrust risk associated with a given business strategy;
☐	**d.** describe classifications of regulations and regulators;
☐	**e.** describe uses of self-regulation in financial markets;
☐	**f.** describe regulatory interdependencies and their effects;
☐	**g.** describe tools of regulatory intervention in markets;
☐	**h.** describe benefits and costs of regulation;
☐	**i.** describe the considerations when evaluating the effects of regulation on an industry.

INTRODUCTION

1

Regulation can be described as a form of government intervention in markets that involves rules and their enforcement. It is an important topic because regulation not only has potential effects at the macro level on the economy but also has potential effects at the micro level on companies and individuals. Regulation may develop either proactively in anticipation of consequences of changes in the market environment or reactively in response to some occurrence(s). For example, changes that resulted from technological advances in the markets because of new means of communication and applications of computers have led to a variety of regulation, both proactive and reactive. Regulation has also developed in response to financial crises and undesirable behaviors or actions that have occurred in the past. Regulations are necessary because in some situations market solutions are not adequate. In other words, regulations exist to protect end users from market failings.

A significant challenge for financial regulators is how to deal with systemic risk (the risk of failure of the financial system) and the consequences of risk taking by financial institutions. Issues such as labor regulation, environmental regulation, and electronic privacy are also receiving increased attention.

How regulations are developed and applied can have significant impacts on businesses. Changes in regulatory framework and regulatory uncertainty can also have substantial effects on business decisions. So, one of the significant challenges facing professionals in the finance industry is to anticipate and understand the consequences of potential changes in the regulatory environment and of specific regulations.

Section 2 of this reading describes the economic rationale of regulation, including how regulation improves fairness in markets and addresses the danger to society of financial system failure. Section 3 provides an overview of regulators, the tools at their disposal, and how the work of regulators around the globe is interdependent. Section 4 describes the assessment of costs and benefits of regulation and highlights practical issues that arise from the implementation of regulation. A summary and practice problems conclude the reading.

2 ECONOMIC RATIONALE FOR REGULATION

Regulations are necessary because market solutions are not adequate for all market situations. Conceptually, this need can be understood best using ideas from economic theory. One of the basic principles in economics is the "fundamental theorem of welfare economics." Assuming constant returns to scale, no frictions,[1] and no externalities, competitive market (equilibrium) allocations[2] are efficient, or *Pareto optimal*. That is, there is no way to redistribute resources and make some agents better off without making others worse off.[3] Furthermore, any efficient allocation of resources can be sustained as a market equilibrium for an appropriate set of prices. Hence, absent frictions and externalities, the market solution will be economically efficient and regulation would be needed only to ensure consumer protection and privacy rights.

The case for regulatory intervention rests on the presence of **informational frictions**, externalities, weak competition, and social objectives. Informational frictions are market inefficiencies that lead to sub-optimal outcomes. They include lack of access to information and inadequate information. Such frictions result in a variety of issues that regulators attempt to address. These issues include "adverse selection" (private information in the hands of some, but not all, market participants that allows the holder of that information to gain at the expense of others) and "moral hazard" (incentive conflicts that arise from the delegation of decision making to agents or from contracts that will affect the behavior of one party to the detriment of the other party to the contract). Asymmetrical information, in general, may give one entity an inherent advantage over another entity with which interaction occurs. The resulting regulation focuses on establishing rights and responsibilities of entities and on seeking to establish a level playing field in the dissemination of information in the market.

1 Examples of frictions are costs for or restraints on trading and asymmetrical information.

2 Market (equilibrium) allocations are ones in which (1) agents maximize utility given relative prices and (2) markets clear.

3 If resources can be redistributed such that any one agent can be made better off without making any other agent worse off, then the original allocation would not have been *Pareto optimal*.

Externalities are spillover effects of production and consumption activities onto others who are not directly involved in a particular transaction, activity, or decision. A positive externality provides a spillover benefit, and a negative externality generates a spillover cost. Systemic risk posed by failures of financial institutions is an example of an externality, as is environmental pollution. Both can have far-reaching consequences for the public. An example of a positive externality are home improvements where those living nearby may benefit from increases in their home values even though they have expended no resources to improve their properties.

Weak competition can also give rise to regulatory intervention. Weak competition is considered to be detrimental to consumers owing to high prices, less choice, and lack of innovation. It is associated with scenarios in which a dominant firm has significant market power or where firms collude and agree to keep prices high.

Social objectives are typically resolved by the provision of public goods that would not be provided by the market. An important feature of public goods is that consumption by one individual does not reduce the availability of the good for others. Usually funded by the government, examples include defense, police protection, and education. Alternatively, social objectives may be achieved by placing regulatory obligations on firms—for example, by requiring energy companies to give discounts on energy bills to vulnerable customers or by requiring telecommunication companies to provide service to remote customers who would otherwise not be served because of the additional costs of providing such service.

It is difficult, if not impossible, to think of an area of life unaffected by regulation. Regulations address a broad range of issues and can be classified by their objectives. These include the following:

- Safety (for example, food and products)
- Privacy (for example, financial information)
- Protection (for example, intellectual property)
- Environmental (for example, pollution)
- Labor or employment (for example, workers' rights and employment practices)
- Commerce or trade (for example, consumers' rights and protection, investors' protection, and antitrust)
- Financial system (for example, prudential supervision of institutions, capital requirements, and insider trading)

2.1 Rationale for the Regulation of Financial Markets

The regulation of securities markets and financial institutions is essential because of the consequences to society of failures in the financial system. These consequences could be experienced at both micro and macro levels. Potential consequences include individual financial losses experienced by individuals, an overall loss of confidence, and disruption to commerce. These consequences were evident in the 2008 global financial crisis. Securities regulation focuses on such goals as protecting investors, creating confidence in markets, and encouraging capital formation. Although it is difficult to define precisely how regulation enhances confidence in the financial system, increasing confidence is cited as one of the motives for securities regulation. Many of the rules oriented toward transparency, equitable access to information (which, in turn, encourages capital formation), and protecting small investors implicitly serve to promote confidence in the markets.

Among the objectives of many financial regulators is the protection of consumers and investors, the safety and soundness of financial institutions, the smooth operation of payment systems, and access to credit. Other (macroeconomic) concerns of financial regulators, particularly central banks, include price stability, levels of employment/unemployment, and economic growth.

A key focus of regulators is maintaining the integrity of markets, ensuring that they operate efficiently and that consumers and investors are informed and not exploited. This role is distinct from financial stability regulation, which is focused on specific outcomes. In addition to securities registration requirements, disclosure requirements are important to facilitate and support the marketplace and the confidence of investors. Disclosures allow investors to use available information to assess the consequences for investing in and valuing financial instruments and to allow markets to operate. Securities market disclosures occur at various levels, in various forms, and with varied and sometimes unexpected consequences. For example, in the European Union, the Markets in Financial Instruments Directive II (MiFID II), implemented in 2018, focuses on improving transparency in financial markets, including in fixed income, derivatives, and other over-the-counter markets in which prices and volumes were not previously publicly disclosed. The opacity of these markets meant that the buildup of risks prior to the 2008 financial crisis went largely undetected by regulators and market participants.

Disclosures are wide ranging and have high importance. They include financial reporting requirements and accounting standards, prospectus disclosure requirements in conjunction with both securities offerings and annual reports, disclosure requirements in the context of proxy proposals and contests, mutual fund disclosure rules, and financial market price transparency rules. Disclosure requirements tend to be oriented toward the protection of investors and the provision of information to investors (either to investors directly or to their service providers).

Many of the regulations governing securities markets are designed to mitigate agency problems that arise through delegation to intermediaries. For many financial transactions, parties need to act through others (agents), leading to the potential for agency conflicts. Among examples of regulations addressing potential agency conflicts are those related to mutual fund fees and governance, the governance of listed companies, rules for proxy voting in companies, best execution requirements for broker/dealers, and treatment of inducements (commissions and other non-monetary benefits) that arise in the provision of investment advice and in portfolio management.[4]

Historically, securities regulators have tended to focus primarily on protecting retail investors (individual investors with modest resources and less investment expertise). This tendency has resulted in a lesser focus on financial regulation of hedge funds, private equity, and venture capital funds because of the type of investors (institutional and affluent individual investors) that invest in these funds. For these larger investors, regulators have taken more of a "buyer beware" stance. For larger investors, it is more difficult to define suitability standards. One approach is to require a more modest range of disclosure requirements related to offering memorandums for a variety of different types of transactions, alongside basic antifraud rules.

Regulations related to prudential supervision of financial institutions and financial stability are critical because of the cost that failure of a financial institution can impose on the economy, capital markets and society. Prudential supervision is regulation and monitoring of the safety and soundness of financial institutions in order to promote financial stability, reduce system-wide risks, and protect customers of

4 MiFID II requires advisers to disclose all costs and charges, including all one-off and ongoing charges, transaction costs associated with the financial instrument, all third-party payments received, and the total combined costs of these three categories. These disclosures must also be accompanied by an illustration that shows the cumulative effect of the overall costs and charges on the return to investors.

financial institutions.[5] The failure of a bank can result in loss of savings and access to credit. The failure of an insurance company can result in unanticipated losses to those insured. If government-sponsored entities provide protection against these losses or the government chooses to cover all or a portion of these losses, the losses can be spread across a broader section of society than simply those directly affected. Additionally, the resulting loss of confidence in the financial system can have far-reaching consequences.

Types of prudential supervision include those that focus on diversifying assets, managing and monitoring risk taking, and ensuring adequate capitalization. In addition, regulators may set up funds to provide insurance against losses and mandate premiums or fees to be paid into these funds. Some regulators, such as those in the European Union, may also require that designated investment firms have in place appropriate recovery plans and resolution plans to be applied if they encounter financial distress.

The benefits of regulation, however, generally come with associated costs. For example, regulations that require certain entities or individuals to use insurance when undertaking certain activities may create moral hazard and result in greater risk-taking incentives. Similarly, regulations that increase capital-holding requirements can reduce the amount of capital available to be distributed in the market.

EXAMPLE 1

Rationale for Regulation

1 Which of the following is least likely to be a reason for the use of regulation?
   - A Systemic risk posed by the financial services industry
   - B Informational frictions in the form of private information
   - C Extensive disclosure of operating and financial information by companies seeking to attract investors' attention

2 Prudential supervision is primarily concerned with:
   - A treatment of inducements in the provision of investment advice.
   - B provision of information about financial products to retail investors.
   - C safety and soundness of the financial system.

Solution to 1:

C is correct. Extensive disclosures are not a reason for the use of regulation. They may be the result of regulation or a reflection of good business practices. Presence of systemic risk and informational frictions give rise to the need for regulation.

Solution to 2:

C is correct. The primary objective of prudential regulation is to ensure safety and soundness of the financial system.

5 Prudential supervision is part of prudential regulation, which includes setting capital adequacy and liquidity standards for such financial institutions as banks and insurers.

2.2 Regulation of Commerce

Government regulation in certain areas of commerce, such as consumer protection, commercial law, and antitrust, is critical to setting out an underlying framework for the operation of private markets and facilitating business decisions that involve a considerable degree of coordination.

Issues pertaining to externalities and public goods (for example, national defense and transportation infrastructure) are critical to the operation of national and global economies and are essential considerations for the work of investment analysts. The relevant decisions arise at a number of levels. Many of these decisions would be within the domain of national governments, but some of the relevant externalities are global.

Although common examples involve local environmental issues, such as pollution, global externalities—such as nuclear waste storage and global warming—occur across countries. So it is important to have international mechanisms to facilitate the coordination and acceptance of responsibilities among national governments (typically, national governments are best able to coordinate decisions within their respective countries). Some of these externalities have long-term consequences (costs) and implications. In most cases, these long-run consequences may be difficult to fully quantify and assess.

Government policy can be important for promoting commerce locally, nationally, regionally, and globally. Trade agreements are important to global commerce. At the national, regional, or local level, governments can facilitate basic features of the business environment, such as establishing the legal framework for contracting and setting standards. Regulation is also central to fundamental aspects of labor markets, such as workers' and employers' rights and responsibilities, as well as workplace safety. Immigration issues are also handled through regulation. Fundamental safety regulations with respect to drugs, food products, medical devices, and pollution are significant too.

Several issues have emerged as particularly relevant in the context of globalization and the internet. One issue is the recognition and protection of intellectual property. Government policies regulate intellectual property, prescribing standards and processes that define and govern patents, trademarks, and copyrights. The legal standards are country specific, and although most countries recognize the importance of protecting intellectual property, lack of enforcement and protection of intellectual property at a global level has emerged as an issue. Setting common technical standards is another global issue, given the focus on ensuring higher levels of interoperability between the technology and electronic tools used in commerce. Even something as basic as establishing domain names and the related standard setting requires some appropriate delegation of authority.

Technological change, including a shift toward digitization, is leading to an increasing amount of data being collected, processed, shared, and used in digital form at lower cost and on a larger scale. "Big Data" gives rise to potential market opportunities, such as the development of data-based business models that rely on the sharing of data and the extraction of commercial value from data. However, it is also giving rise to concerns about privacy and data protection. Privacy is particularly important with respect to medical, financial, academic, and employment records. New regulations, such as the General Data Protection Regulation (GDPR) and the e-Privacy Directive in the European Union, require entities, including businesses and governments, to apply certain protections and safeguards to personal data in their possession and maintain appropriate security procedures. The internet raises a broad set of issues involving privacy because of the breadth of information potentially available about a person's situation (financial and personal), activities, interactions, and purchases. How internet software navigates these privacy concerns will influence both the perceptions and actions of regulators, as well as the acceptance of software innovations and business models in the marketplace.

An effective legal environment is also crucial for the successful operation of commerce. Clearly defined rules governing contracts, their interpretation, and each party's legal rights under a contract are necessary. A framework for financial liability and dealing with bankruptcy is also necessary as an incentive to enter into economic contracts, particularly those that require long-term commitments. Such activities as construction projects, energy exploration, and extraction projects—and even mundane commercial activities, such as relocation decisions—involve significant long-term, dynamic commitments. Pre-commitment by society to a well-defined set of rules and standards is crucial to facilitating the willingness of market participants to engage in long-term commercial activities.

For example, consider the situation in which a company needs to incur significant costs to start a project. These costs are unrecoverable if the project does not progress; in other words, these are sunk costs. Without a strong legal framework that guarantees that the party will recover these initial costs, the party paying the sunk costs would be reluctant to incur them because of the potential of a "holdout" problem in which the other side exploits the fact that the sunk costs have been incurred to force a renegotiation of the deal. Such contractual difficulties would destabilize the operation of businesses and weaken the economy.

2.3 Antitrust Regulation

In a global context, an implicit regulatory goal of government may be to restrict competition from other countries. In a domestic context, a regulatory goal often pursued is to promote competition in most economic sectors (this goal can alternatively be viewed as monitoring and preventing activities that restrict or distort competition). There are several dimensions to this goal. Regulatory approval or notification is typically required for mergers and acquisition of major companies in a specific market. When a merger or acquisition is expected to substantially reduce competition, regulators can block the merger or acquisition or suggest remedies to resolve a perceived issue (for example, divestiture of particular segments of the businesses to resolve an antitrust issue). When there are competing bids, the regulator can effectively decide the outcome on the basis of its assessment of the effects of each bid. Considering the potential response of competition or antitrust agencies is a central aspect to the evaluation of mergers and acquisitions.

Competition and antitrust laws also typically prohibit anticompetitive arrangements or practices, such as price collusion or exchanging certain information, and anticompetitive behavior by companies that dominate a market. Types of behavior that are problematic when undertaken by a dominant company (beyond mergers that substantially lessen competition) include exclusive dealings and refusals to deal, price discrimination, and engaging in predatory pricing. In response to antitrust issues, regulators not only may impose monetary sanctions but also may require companies to change their business (for example, divest portions or change operating/marketing practices). In some jurisdictions, such as the United States, the United Kingdom, Germany, Denmark, Ireland, France, and Australia, individuals can also face imprisonment for engaging in a cartel.

There has been an increasing focus on applying antitrust laws to the technology sector, which includes investigations of Google,[6] Apple, Intel, and Microsoft. In Europe, for example, Google was found to have abused its dominant position in

6 Google is owned by a holding company, Alphabet Inc.

the internet search market by favoring its own comparison shopping service over those of its rivals. Similarly, Intel was found to have abused its dominant position by engaging in exclusive dealing with certain computer equipment manufacturers and retailers. Both companies challenged the claims on the basis that they have brought considerable benefits to consumers in terms of low prices and ever-improving quality. Using competition laws to challenge rivals can also represent a business strategy. An example of such a challenge is Microsoft's challenge in Europe that Google is unfairly impeding competition in the search engine market.

A significant issue that companies need to face in addressing antitrust (lack of competition) issues is that in many cases they need to satisfy simultaneously a range of regulators across multiple jurisdictions. For example, a company may have to satisfy both the US Department of Justice and the European Union if it plans to use a common product and market strategy across jurisdictions. Despite language and cultural differences, it often is advantageous to adopt a unified strategy around the globe because of business imperatives and likely overlapping views among regulators of competition.

EXAMPLE 2

Antitrust Regulation

Which of the following issues is least likely to be the subject of antitrust rules?

A Privacy and data protection

B Anticompetitive behavior by dominant companies in a market

C Mergers and acquisitions by major companies

Solution:

A is correct. Privacy and data protection issues are regulated, but only as part of regulations besides antitrust rules.

3 REGULATORS AND REGULATORY TOOLS

Regulations are sometimes enacted by legislative bodies (often these regulations are laws) but more typically arise from the determination of regulatory bodies.

3.1 Classification of Regulations and Regulators

Broadly speaking, regulators can be either sanctioned by the government or created by an industry on a voluntary basis.

Government-backed regulatory bodies can be either governmental departments and agencies or independent regulators, which derive their power and authority from the state. Government-backed regulatory bodies have legal authority to enact and enforce regulations within the parameters of the mandate given to them. In many instances, a legislative body enacts a statute at a broad level, leaving it to regulatory bodies to implement and apply the detail of the regulation.

Illustration of Regulatory Process

This description by the US Securities and Exchange Commission (SEC) is illustrative of how the process works: "Rulemaking is the process by which federal agencies implement legislation passed by Congress and signed into law by the President. Major pieces of legislation, such as the Securities Act of 1933, the Securities Exchange Act of 1934, and the Investment Company and Investment Adviser Acts of 1940, provide the framework for the SEC's oversight of the securities markets. These statutes are broadly drafted, establishing basic principles and objectives. To ensure that the intent of Congress is carried out in specific circumstances—and as the securities markets evolve technologically, expand in size, and offer new products and services—the SEC engages in rulemaking" (www.sec.gov/about/whatwedo.shtml).

Courts play an important role in regulation as well—helping interpret regulations and laws, defining permitted and proscribed regulatory practices, and, in some instances, imposing sanctions for regulatory violations. State-backed regulations can, therefore, be classified as comprising

- laws enacted by legislative bodies (**statutes**),
- rules issued by government agencies or other regulators (**administrative regulations or administrative law**), and
- interpretations of courts (**judicial law**).

Although government departments and agencies make many regulations, **independent regulators** can also make regulations in accordance with their powers and objectives. The authority of independent regulators, such as the Financial Conduct Authority in the United Kingdom, comes from their recognition, autonomy, and powers given to them by a statute, government department, or government agency, but they are not government agencies per se. One distinction between government agencies and independent regulators is that the latter typically do not rely on government funding and are often given a degree of autonomy in terms of decision making. Some argue that an advantage of independent regulators is that they are to some extent immune from political influence and pressure and can, therefore, take a more technical and long-term view of policies, which would achieve the objectives they have been created to pursue.

In contrast to state-backed government agencies or independent regulators, industry **self-regulatory bodies** are private organizations that both represent and regulate their members. Although these organizations are independent of the government and to an extent are isolated from political pressure, they may be subject to pressure from their members. Industry self-regulatory bodies derive authority from their members, who agree to comply with the organization's rules and standards and their enforcement. This authority does not have the force of law, but industry self-regulatory bodies do have the power to exclude or expel parties from being members. In order to ensure minimum standards are maintained, certain entry requirements (such as training or ethical standards) may be imposed.

Some industry self-regulatory bodies, particularly in the securities industry, are known as **self-regulating organizations (SROs)**. SROs differ from standard industry self-regulatory bodies in that they are given recognition and authority, including enforcement power, by a government body or agency. SROs are funded independently, rather than by the government. For example, the US SEC, the government agency that regulates the securities markets in the United States, allocates some regulatory responsibilities to the Financial Industry Regulatory Authority (FINRA), which is an SRO.[7] It has the authority to enforce industry rules and federal securities laws.

The role of SROs varies among countries. In some countries, such as the United States, SROs have specific regulatory authority, and in other countries, self-regulating organizations are rarely or never recognized as independent regulators. For example, the Australian financial regulator, ASIC, has stated, "One of the many significant recent legislative amendments that was introduced in Australia with the Financial Services Reform Act 2001 was the removal of the official regulatory standing of SROs. SROs, whether they are exchanges, industry associations, or some other form of peer group, have traditionally set standards of behavior or codes of conduct for market participants."[8] According to the World Bank,[9] the role of self-regulation in Europe, with the exception of the United Kingdom, was limited because of civil law systems and the resulting reliance on government supervision. In the United Kingdom and other countries with common law systems, reliance on self-regulation has been more extensive. The roles of SROs in regulation in these countries range from non-existent to having some regulatory authority. Regulators are concerned with the corporate governance of SROs and the management of their conflicts of interest. The extent of the concern is a factor in deciding the regulatory role, if any, of the SRO in question.

The relatively simple classification of regulators (legislative bodies, government agencies, independent regulators, and courts) and regulations (statutes, administrative regulations, and judicial law) is useful but does not reflect the complexities and nuances that exist with respect to regulators and regulation. In some cases, the classification of a regulator is clear, and in other cases, the classification is ambiguous. For example, the Public Company Accounting Oversight Board (PCAOB) is a non-profit corporation, established by the US Congress to oversee the audits of public companies. Previously, the audit profession was self-regulated. The PCAOB is funded primarily through annual fees paid by public companies, brokers, and dealers. The SEC oversees the PCAOB. The PCAOB is an independent regulator rather than a government agency, but it is not an SRO.

In Singapore, statutory boards are entities separate from the government, with specific legislation governing their operations. Most, if not all, statutory boards impose charges for some or all of their services. Those statutory boards that do not generate sufficient revenue to meet their expenses receive grants from the government to finance their operations. The grants are funded from the government's annual budget. The statutory boards are described as separate from the government, yet they are subject to specific legislation governing their operations and they may receive government funding. Whether Singapore's statutory boards are government agencies or independent regulators is ambiguous. The Singapore Economic Development Board (EDB), one such statutory board, describes itself as "a government agency under the Ministry of Trade and Industry…responsible for strategies that enhance Singapore's position as a global centre for business, innovation, and talent."[10] Another statutory board, the Accounting and Corporate Regulatory Authority (ACRA), describes itself as "the national regulator of business entities, public accountants and corporate service providers in Singapore."[11] Although the EDB clearly identifies itself as a government

7 On its website the FINRA states: "FINRA is dedicated to investor protection and market integrity through effective and efficient regulation of broker-dealers. FINRA is not part of the government. We're a not-for-profit organization authorized by Congress to protect America's investors by making sure the broker-dealer industry operates fairly and honestly." (http://www.finra.org/about)

8 https://asic.gov.au/media/1339352/integration-financial-regulatory-authorities.pdf

9 Carson (2011).

10 www.edb.gov.sg/en/about-edb/who-we-are.html

11 www.acra.gov.sg/about_Acra.

agency, it is less clear whether the ACRA, given the description of a statutory board and the description of itself, should be classified as a government agency or an independent regulator.

Classifying regulatory bodies that exist in unions, such as the Union of South American Nations and the European Union (EU), can also present challenges. For example, the European Commission, which has a mission to promote the general interest of the EU, can initiate legislation in the form of directives and regulations, which are subject to debate and approval by the European Parliament and the European Council (the co-legislators). The directives and regulations passed by the European Parliament and the European Council are jointly referred to as "EU law." Regulations have binding legal force in every EU member state on a par with national laws. Directives identify desired results and require national authorities to put laws in place to achieve them. Decisions are binding laws addressed to specific parties and are the result of specific cases.[12] Regulations appear to have the characteristics of administrative regulations. Directives appear to have the characteristics of statutes; they are at a broad level, and another body needs to fill in the implementation details. Decisions appear similar to judicial law. Regardless of how a regulatory body is classified, it is important to identify the regulators and regulations that might affect the industry or company being analyzed.

EXAMPLE 3

Classification of Regulators

1. The media devotes considerable coverage to a regulatory body that has been given autonomy by the government and is empowered by statute. The regulatory body has recently raised the fees charged to the companies it regulates. The regulatory body in question is most likely to be a(n):
   - **A** self-regulatory organization.
   - **B** government agency.
   - **C** independent regulator.
2. Which of the following is least likely to be a characteristic of a self-regulatory body?
   - **A** It represents and regulates its members.
   - **B** It carries out government policy.
   - **C** It can discipline members that violate its rules and principles.

Solution to 1:

C is correct. Independent regulators are given authority by the government and are empowered by statute. Unlike government agencies, they are funded by fees that they collect from the firms they regulate.

Solution to 2:

B is correct. Self-regulating bodies do not carry out government policy. They are meant to be independent from government and immune to its influence.

Regulatory authorities may reference the work of outside bodies in their regulations. Examples of these outside bodies are accounting standard–setting bodies, such as the International Accounting Standards Board (IASB) and the Financial Accounting

12 https://ec.europa.eu/info/law/law-making-process/types-eu-law_en

Standards Board (FASB), and credit-rating agencies. Regulatory authorities have the legal authority to enforce any regulation that references the work of these bodies. In the case of accounting standard–setting bodies—which are typically private sector, non-profit, self-regulated organizations—the requirement to prepare financial reports in accordance with specified accounting standards is the responsibility of regulatory authorities. The standard-setting bodies may set the standards, but the regulatory authorities recognize and enforce the standards. Ratings by credit-rating agencies—which are typically private sector, profit-oriented entities—were often referenced in regulations related to acceptable holdings by certain entities. Issues with conflicts of interest when the agencies were paid by the firms they rated, however, have resulted in efforts to reduce references to credit-rating agencies in regulations.

Although much of the focus of this reading is on the rules themselves and their development, impact, and implementation, regulatory enforcement and sanctions also play an important role. This division between development and enforcement of regulation also represents a possible way to classify laws or regulation. **Substantive law** focuses on the rights and responsibilities of entities and relationships among entities, and **procedural law** focuses on the protection and enforcement of the substantive laws. Regulators typically have responsibility for both substantive and procedural aspects of their regulations.

3.2 Regulatory Interdependencies

An interesting facet of regulation is how regulated entities view the regulation, which is often context specific. Although there are many examples in which regulated companies fight against new proposed regulations, an outright opposition is relatively rare. Regulated company efforts to fight particular regulations tend to attract more public attention than when the companies are sympathetic to the proposed regulations. Even more fundamentally, academics have argued that regulation can sometimes enhance and work to the benefit of the interests of the regulated. This argument is often called the "**regulatory capture**" theory (see Stigler 1971). For example, regulatory actions and determinations can restrict potential competition (for example, by limiting entry) or effectively coordinate the choices of rivals (by imposing certain quality standards or price controls). In the interactions between regulated entities and their regulators, the regulated entities may possess considerable expertise and knowledge and some of the individual regulators may have worked in the industry or aspire to be in the industry in which the regulated entities operate. These interactions may reinforce regulatory capture.

Regulatory differences across jurisdictions can lead to shifts in location and behavior of entities because of **regulatory competition** and **regulatory arbitrage**. Regulators may compete to provide a regulatory environment designed to attract certain entities (regulatory competition). As a result, companies may engage in regulatory arbitrage; for example, they may identify and use some aspect of regulations that allows them to exploit differences in economic substance and regulatory interpretation in foreign and domestic regulatory regimes to the companies' benefit.

Interdependence in the actions of regulators dealing with the same issues and activities is important in the international arena. Many regulatory issues are relatively similar around the globe. This commonality reflects both similarities in the challenges confronting different countries and the diffusion of the underlying problems around the globe. Such issues as financial systemic risk, terrorism financing, money laundering, and climate change reflect global concerns and, therefore, are well suited to an approach based on regulatory cooperation and coordination. For other issues, however, domestic regulators in specific jurisdictions often adopt different perspectives or face different trade-offs when developing and applying regulations in their jurisdiction. These varying perspectives can lead to differences in regulatory treatments of the same issue across countries. For example, some jurisdictions have significantly greater disclosure requirements to protect investors than other jurisdictions have. Although such differences are often justified, "regulatory competition" can reduce the effectiveness of regulation in particular countries. Regulatory competition can lead to what is sometimes referred as a "race to the bottom," where countries continually reduce their regulatory standards to attract as many companies as possible to their jurisdiction.

Consider issues related to global warming and pollution. How should governments manage and coordinate efforts around the globe? The relevant externality is not simply within countries but, rather, extends beyond country borders. One of the challenging aspects of this issue is that countries differ in how much they contribute to climate change and in terms of how they are affected by it. Put simply, the countries most affected by climate change may not be the ones that are contributing the most to it. What are the institutional and governance mechanisms that would be appropriate to address this issue on a global basis? Although an economist's solution to the problem of pollution externalities might be to tax pollution or to introduce an emission trading system (or a cap and trade system) in order to allocate the pollution to the parties that can absorb the cost, the practical application of the solution may be complicated. How should one allocate "permits" to pollute among countries? Should countries have the "right" to pollute related to their past pollution? If not, how would one accommodate differences in living standards? How should one address the equity issues associated with low wealth and developing countries' having a potential comparative advantage in absorbing pollution?[13]

The point of this overall discussion of interdependencies among jurisdictions is not to suggest the existence of global governance or a global regulator but, rather, is to recognize the reality and implications of diverse trade-offs and preferences among regional, national, and local regulators. To a degree, the presence of diverse and competing jurisdictions influences the stances of national and regional regulators. Evidence that governments recognize the necessity for global regulatory cooperation and coordination on some issues exists. For example, the Basel Accords established and promote internationally consistent capital requirements and risk management practices for larger international banks. The Basel Committee on Banking Supervision has evolved into a standard setter for bank supervision, among other functions. Another example is the International Organization of Securities Commissions (IOSCO), a self-regulating organization but not a regulatory authority. Its members (national regulators) regulate a significant portion of the world's capital markets. This organization has established objectives and principles to guide securities and capital market regulation, which its members agree to adhere to.

How IOSCO Enhances Regulatory Cooperation

The member agencies of IOSCO have resolved, through its permanent structures,

- "to cooperate in developing, implementing and promoting adherence to internationally recognized and consistent standards of regulation, oversight and enforcement in order to protect investors, maintain fair, efficient and transparent markets, and seek to address systemic risks;

13 A memo that Larry Summers wrote in 1991, while he was chief economist at the World Bank, suggested that poor countries should bear much of the pollution (with compensation) and resulted in controversy.

- to enhance investor protection and promote investor confidence in the integrity of securities markets, through strengthened information exchange and cooperation in enforcement against misconduct and in supervision of markets and market intermediaries; and
- to exchange information at both global and regional levels on their respective experiences in order to assist the development of markets, strengthen market infrastructure and implement appropriate regulation" (www.iosco.org/about/?subsection=about_iosco).

IOSCO is a standard setter and an establisher of best practices for securities regulators, and it has developed a framework of matters to be addressed in the domestic laws of a jurisdiction to facilitate effective securities legislation.[14] This framework is shown in Exhibit 1.

The framework also serves as a useful, but by no means exhaustive, list of areas of regulation relevant to an analyst. Labor, consumer protection, and environmental, health, and safety laws, which are not included in the list, may also significantly affect a business or industry.

Awareness of the basic types of laws and regulations that affect economies, financial systems, industries, and businesses is useful to an analyst. This knowledge will help the analyst to identify areas of concern and to consider proactively potential effects of regulations, existing and anticipated.

Exhibit 1 IOSCO's Objectives and Principles of Securities Regulation

Effective securities regulation depends on an appropriate legal framework. The matters to be addressed in the domestic laws of a jurisdiction include the following:

1 **Company Law**
  - **1.1** company formation
  - **1.2** duties of directors and officers
  - **1.3** regulation of takeover bids and other transactions intended to effect a change in control
  - **1.4** laws governing the issue and offer for sale of securities
  - **1.5** disclosure of information to security holders to enable informed voting decisions
  - **1.6** disclosure of material shareholdings

2 **Commercial Code/Contract Law**
  - **2.1** private right of contract
  - **2.2** facilitation of securities lending and hypothecation
  - **2.3** property rights, including rights attaching to securities, and the rules governing the transfer of those rights

3 **Taxation Laws**
  - **3.1** clarity and consistency, including, but not limited to, the treatment of investments and investment products

4 **Bankruptcy and Insolvency Laws**
  - **4.1** rights of security holders on winding up

14 International Organization of Securities Commissions, "Objectives and Principles of Securities Regulation" (May 2003).

Exhibit 1 (Continued)

- **4.2** rights of clients on insolvency of intermediary
- **4.3** netting

5 Competition Law

- **5.1** prevention of anticompetitive practices
- **5.2** prevention of unfair barriers to entry
- **5.3** prevention of abuse of a market dominant position

6 Banking Law

7 Dispute Resolution System

- **7.1** a fair and efficient judicial system (including the alternative of arbitration or other alternative dispute resolution mechanisms)
- **7.2** enforceability of court orders and arbitration awards, including foreign orders and awards

Source: International Organization of Securities Commissions, "Methodology For Assessing Implementation of the IOSCO Objectives and Principles of Securities Regulation" (May 2017): Appendix 1, "The Legal Framework." www.iosco.org/library/pubdocs/pdf/IOSCOPD562.pdf.

Even within countries, the objectives of diverse government regulators can differ and potentially lead to regulations that seem inconsistent. Bank supervisors (whether as a function of the central bank, another entity, or a combination of entities) generally focus on **prudential supervision**—regulation and monitoring of the safety and soundness of financial institutions in order to promote financial stability, reduce system-wide risks, and protect customers of financial institutions. The objectives of securities commissions or regulators are typically to protect investors; ensure that markets are fair, efficient, and transparent; and reduce systemic risk. In some situations, the goals of the bank supervisor and securities regulator can be in tension, resulting in conflicting objectives. For example, on the one hand, the bank supervisor may be reluctant or even unwilling to release the results of stress tests of financial institutions in order to promote financial stability and avoid systemic risk because of the potential loss of confidence. On the other hand, a securities regulator might advocate for the release of information that might be relevant to investor decision making and act to protect investors (see Spatt 2009).

A general conclusion is that regulation by different regulators, even with seemingly similar objectives, can lead to very different regulatory outcomes.

EXAMPLE 4

Regulatory Interdependencies

1 A country's securities regulator is looking to attract higher number of smaller companies to its capital markets. It proposes to ease hurdles that companies face when preparing to list shares on the country's stock exchange. The proposals include the lowering of the frequency of financial reporting, reducing the extent of disclosures required, and reducing the minimum size of company that can be accepted on the market.

This is an example of:

A regulatory competition.

B regulatory coordination.

C regulatory capture.

2 Regulatory capture is most likely to be a concern where there is reliance on:

A SROs.

B government agencies.

C government departments.

Solution to 1:

A is correct. Regulatory competition occurs when regulators compete to provide a regulatory environment designed to attract certain entities.

Solution to 2:

A is correct. Regulatory capture has been a concern when SROs are used.

3.3 Regulatory Tools

Regulatory and government policies should be predictable as well as effective in achieving objectives. It is very difficult for any entity to function with confidence and success in an environment where the rules are unclear or in a state of flux (in other words, where there is considerable regulatory uncertainty). Regulatory choices or government policies that will be consistent over time are desirable. If these choices occur, the regulatory environment is likely to be stable despite the fact that, in many countries, governmental decision makers (with diverse political preferences) change on a regular basis. It is helpful to use regulatory tools that are consistent with maintaining a stable regulatory environment. Regulatory tools and government interventions in markets include the use of price mechanisms, such as taxes and subsidies; regulatory mandates and restrictions on behaviors, including establishing rights and responsibilities; provision of public goods; and public financing of private projects.

The issue of how to address pollution is a classic example in regulation. By taxing polluters (or subsidizing those who do not pollute, by using a suitable baseline), one can create a system in which marginal incentives are equated across economic agents. The advantage of such an arrangement is that, theoretically, the rights to pollute are redistributed in an "efficient" manner relative to a fixed allocation. In particular, the structure of the regulation allows market incentives to redistribute the pollution rights to those for whom they are the most valuable at the margin. There are important issues, however, about how to initially establish and distribute the amount of acceptable total pollution. In some situations, historical usage (amount of pollution produced) is used to allocate pollution rights One problem is that marginal incentives may be altered in anticipation of this allocation. In other situations, the allocation is the outcome of a political process, which can lead to considerable lobbying. At the heart of this example is the use of a price mechanism to create the appropriate marginal incentives and an efficient allocation of resources. The Coase theorem states that if an externality can be traded and there are no transaction costs, then the allocation of property rights will be efficient and the resource allocation will not depend on the initial assignment of property rights.

Governments can intervene in markets in ways other than through the price mechanism. These include restricting some activities (e.g., insider trading and short selling), mandating some activities (e.g., capital requirements for banks and registration with a securities commission for certain activities), providing public goods (e.g., national defense and transportation infrastructure), and financing private projects (e.g., loans to individuals or companies for specified activities that the government deems desirable to encourage). The extent of government provision of public goods and government financing of private projects depends on a number of factors, including the political philosophy of the country and/or government in power, the structure of the government, and the country's gross domestic product. The problem of **systemic risk** (the risk of failure of the financial system) as a result of the failure of a major financial institution has emerged as an issue in many countries around the world in the aftermath of the 2008 global financial crisis. Systemic risk and **financial contagion** (a situation in which financial shocks spread from their place of origin to other regions; in essence, a faltering economy infects other, healthier economies) are examples of negative externalities. In the EU, the European Systemic Risk Board, formed in December 2010, is an advisory EU body within the European Central Bank tasked with advising national macroprudential bodies to take steps to address risks.

Exhibit 2 focuses on how "bail-in" tools can mitigate systemic risk.

Exhibit 2 Bail-In Tools

Among the regulatory tools introduced following the 2008 financial crisis are so-called bail-in powers, which were endorsed by the Financial Stability Board and the G–20. A bail-in tool is seen as improving the toolkit for dealing with the failure of large, globally systemic banks. Bail-in involves shareholders of a failing institution being divested of their shares and creditors of the institution having their claims canceled or reduced to the extent necessary to restore the institution to financial viability. Bail-in policies are intended to ensure that shareholders and creditors of the failed institution, rather than taxpayers, pay the costs of the failure. This situation contrasts with how many governments dealt with the financial crisis of 2008, when banks and insurers were rescued by taxpayers in a number of countries. Such policies also allow the failed institution to continue to operate, so that it can introduce restructuring measures to address the cause of the failure. These policies can limit disruption to the customers of the institution and help maintain confidence in the banking system.

It is difficult to assess the extent to which the new approaches and tools, such as bail-in policies, will reduce systemic risk. There are a number of reasons for this difficulty. The types and sources of future crises are likely to be different from those of the past, so regulations designed with a prior crisis in mind may not prevent a future crisis. It can be difficult to assess the potential effectiveness of regulatory actions before an event and even after the fact. The mere fact that a crisis does not occur is not necessarily evidence that regulations prevented one. It is also plausible that some regulatory responses have the unintended consequence of mitigating one source of risk while increasing another source of risk. All these issues make effective regulation challenging to design.

Generally, more than one regulatory approach or policy is feasible and worthy of consideration in a specific situation. Two examples that illustrate a range of possible regulatory responses are (1) conflict-of-interest policies and (2) trading restrictions on insiders, which are explored in Exhibits 3 and 4, respectively.

Exhibit 3 Regulating Conflicts of Interest

Consider the hypothetical scenario in which a potential employee of a regulator has some degree of financial exposure to a regulated company. Such exposure could come about in many ways (for example, spousal employment, a marketable position in an investment portfolio, or an illiquid position resulting from past employment) and at a variety of financial levels.

(continued)

Exhibit 3 (Continued)

What types of regulatory policies might be appropriate to mitigate these risks?

Among the potential regulatory responses are the following: The individual could be barred from employment at the regulatory agency or from working on specific (or all) projects involving the regulated company in question. The individual could sell the position; the sale could be voluntary or mandated. The individual could be required to disclose the nature of his potential conflict to higher-level decision makers to whom he will be providing recommendations. Other potential policies include a bar on involvement, resolution of the conflict, or disclosure of it.

Exhibit 4 Regulating Corporate Insiders

Turning to the case of corporate insiders, there are potential regulatory and corporate restrictions. Examples of regulatory responses are a ban from trading on non-public information and a requirement that when they do trade, the insiders disclose the trades. The company may impose a blackout period during which insiders are banned from trading on the company's stock (these periods often precede earnings announcements and continue for a short while afterward). The appropriate remedy depends on the underlying facts and circumstances, and arguably the appropriate standards would reflect the specific context.

An important aspect of effective regulation is the potential ability to impose sanctions on violators of the regulations; in other words, it is important to be able to enforce the regulations. IOSCO clearly identifies this aspect as one of the agreed-on principles of securities regulation: "The regulator should have comprehensive enforcement powers."[15] Enforcement of securities regulations and regulations on businesses may include sanctions for the violating corporation (business or company), the individual violator(s), or both. Corporate sanctions may be appropriate if the company caused harm to others. The sanctions often involve monetary fines/fees/settlement, and in the case of individuals, the sanctions may involve prison terms. In some situations, such as in cases of accounting fraud, shareholders may actually be the victims. In such instances, if the stockholders were harmed by the wrongdoing, the case for sanctions, such as fines, against the company is often far from compelling. The sanctions may simply redistribute funds from current shareholders to the stockholders who were the specific victims, and the company incurs real resource costs.

For various reasons, it can be difficult to prosecute or achieve settlements with individual violators. First, it often is difficult to detect violations and to identify exactly which individuals were at fault. Second, the individuals possess strong incentives to fight in order to protect their reputation and livelihood. Indeed, individuals are often able to fight using corporate resources because of indemnification provisions in their employment contract. The intent of these provisions may be to protect risk-averse executives against inadvertent liability and to potentially align their interests with those of the stockholders, but they may result in protecting executives to the detriment of the stockholders. The incentive to fight individual sanctions may be especially strong because of not only financial costs but also other costs, such as reputational costs.

15 International Organization of Securities Commissions, "Objectives and Principles of Securities Regulation" (June 2010).

EXAMPLE 5

Regulatory Tools

1 Globalization of capital markets is *most likely* to result in increased concerns about:

 A financial contagion.

 B regulatory competition.

 C both contagion and regulatory competition.

2 The regulatory tools *least likely* to be used by self-regulating organizations are:

 A price mechanisms.

 B restrictions on behaviors.

 C provision of public goods.

Solution to 1:

C is correct. Globalization is likely to result in increased concerns about contagion and regulatory competition. It is easier for a financial shock to spread. Governments may use their regulatory environment to attract entities from around the world.

Solution to 2:

A is correct. SROs are least likely to use price mechanisms. They typically regulate behaviors and often provide public goods in the form of standards.

ANALYSIS OF REGULATION

4

The effects of regulation can range from macro effects that impact large parts of the economy to micro effects on an individual business. Section 4.1 introduces the concept of cost–benefit analysis carried out by regulators. Section 4.2 illustrates how an analyst could approach the task of assessing the effect of regulation on a particular industry. Because regulations can evolve in response to market, technological, and societal changes, it is important to monitor issues of concern to regulators and ongoing developments to evaluate the implications of potential changes in regulation. Understanding the regulatory process will help an analyst recognize the types of challenges that regulators and policymakers face and formulate expectations of regulatory outcomes.

4.1 Basic Concepts of Cost–Benefit Analysis

In assessing regulation and regulatory outcomes, it is common practice for regulators to assess the overall benefits and costs of regulatory proposals to assess the trade-offs associated with a particular regulatory action and to assess alternative solutions. Regulators, guided by economic principles, strive to develop techniques to enhance the measurement of the costs and benefits of regulations. The general benefits of regulation as discussed in previous sections may be clear, but the measurement of the full impact of the regulation (both benefits and costs) can be challenging. In conducting cost–benefit analysis of regulation, it often is easier to assess the costs of regulation, although doing so can also be challenging.

Regulatory burden refers to the costs of regulation for the regulated entity; these costs are sometimes viewed as the private costs of regulation or government burden. **Net regulatory burden** is the private costs of regulation less the private benefits of regulation. Many regulators focus narrowly on the implementation costs of regulation (for example, how many compliance lawyers will need to be hired—and at what cost), but in many instances, the most significant costs are the indirect ones that relate to the ways in which economic decisions and behavior are altered and market allocations are changed.

Regulators view some of the costs associated with regulations as "unintended," but it is important to distinguish between two types of such costs. There may be implementation costs that were unanticipated (for example, if it turns out more compliance lawyers need to be hired than originally thought) and indirect costs because of unintended consequences. It is important for regulators to recognize that their evaluation of potential regulations should reflect indirect costs as well as the consequences that were the direct objective of the rule making. Furthermore, in some cases, regulatory filings and consultations in response to proposed regulations identify at least some of the "unintended consequences" prior to the implementation of the regulations. In these circumstances, it is difficult to argue that such consequences were unanticipated and unintended if they were identified prior to the implementation of the regulation. Unintended consequences are reflective of underlying policy risk and may result in high unanticipated costs.

Regulatory costs and benefits are especially difficult to assess on a prospective basis relative to a retrospective basis. An after-the-fact analysis allows a comparison of the item(s) of interest before and after the regulation occurs. This comparison allows for a more informed assessment of a regulation because the actual costs and benefits may be identifiable. In some instances, a trial or pilot analysis may be appropriate and helpful to more fully understand the potential impacts in advance of a proposed regulation. A potentially feasible and relevant approach in the context of an environment with frequent trading is to use natural experiments and trial phase-ins to generate data suitable for careful cost–benefit analysis.[16] This approach facilitates the assessment of statistical evidence to evaluate the effects prior to the full implementation of the proposed regulation. Such approaches are more feasible for a trading rule in a market with high trading frequency that will generate considerable data and run little risk of disrupting the real economy. Similar approaches, sometimes known as "regulatory sandboxes," are being introduced by regulators in such countries as the United Kingdom, Singapore, Australia, the United Arab Emirates, and Malaysia.

In the United States, administrative law requires that federal regulatory agencies conduct a cost–benefit analysis to assess the consequences of their actions. Court rulings have struck down regulatory actions because cost–benefit analyses performed were deemed inadequate. For example, the US Circuit Court of Appeals overturned the 2004 SEC rule requiring that mutual funds have independent chairs and at least 75% independent directors on such grounds.[17] In 2011, as reported in the *Wall Street Journal*, "Striking a blow to the shareholder rights movement, a federal appeals court threw out a controversial new Securities and Exchange Commission regulation that would give investors more power to oust corporate directors…The court issued a harsh rebuke to the SEC, saying it didn't adequately analyze the costs to U.S. companies of fighting in contested board elections."[18] Requirements to undertake cost–benefit analyses (also known as "impact assessments") prior to the introduction of new regulations exist in a number of other jurisdictions, including in the EU and Australia.

16 Among the contexts in which such techniques have been used by US securities regulators have been rules involving short sales, post-trade price reporting, and the tick size increment for trading.
17 See *Chamber of Commerce v. SEC*, 412 F.3d 133 (D.C. Cir. 2005) and 443 F.3d 890 (D.C. Cir. 2006).
18 *The Wall Street Journal, July 23, 2011*, "Court Deals Blow to SEC, Activists", Jessica Holzer

Ideally, regulatory judgments should reflect economic principles and full consideration of the economic costs and benefits, rather than the preferences of current decision makers. Although the potential failure of the fundamental theorem of welfare economics suggests the potential relevance of regulation, it is important to use economic principles to identify and assess alternative remedies and specific actions.

Illustration of Cost–Benefit Analysis

To illustrate the issues that may be relevant to a regulator when considering a cost–benefit analysis, we focus here on a proposal for the introduction of new regulation in Europe to remove the red tape and compliance burdens faced by small and medium-sized enterprises (SMEs) when seeking an initial public offering (IPO) of their shares or for issuing bonds on public markets. The proposed regulation is intended to address three problems. First, the "one-size-fits-all" approach to some areas of financial regulation has led to a perception that the costs of listing outweigh the benefits. Second, there has been a progressive decline in the number of smaller brokers and investment firms that specialize in trading shares of smaller companies. A small, local brokerage ecosystem is necessary to support smaller companies with the listing process. The third problem is the lack of investment in shares and bonds of small companies.

The objective of the proposed regulation would, therefore, be to revive IPOs and bond offerings of small companies. The cost–benefit assessment would examine two proposals: first, whether some of the existing regulations, such as MiFID II and the Market Abuse Regulation (MAR), could be adapted to accommodate smaller issuers. The costs and benefits of (1) targeted changes and clarifications to the existing regulations or (2) an overhaul of the provisions would be examined relative to a "baseline scenario" where the existing rules under MiFID II and MAR do not change. The second proposal would examine the cost and benefits of the introduction of new provisions. These could include simpler delisting rules, rules enabling easy transfer of a listing from less regulated "small company growth markets" to regulated markets, and less stringent free float requirements to make it more attractive for issuers, investors, and market operators.

A regulator carrying out cost–benefit analysis would need to assess the positive impact on the companies' investment and growth rates as a result of easier access to capital. The costs the regulator would have to consider include the additional regulatory oversight over the small companies in question and the consequences of any impact of corporate failures among the listed smaller firms on investor confidence and, consequently, the remaining firms' cost of capital.

EXAMPLE 6

Cost–Benefit Analysis

An investment adviser is discussing a client's portfolio exposure to the electric utilities sector. The sector's regulator has outlined series of proposals for new regulation, on which it is carrying out cost–benefit analysis. The adviser makes two statements to the client about the regulator's cost–benefit analysis.

Statement 1 "The regulator will assess and take into account as part of its cost–benefit analysis only the indirect costs of new regulation arising from changed economic decisions and behaviors."

Statement 2 "Regulatory costs and benefits are easier to assess on a retrospective, after-the-fact basis."

Which statement is correct?

A Only Statement 1 is correct.

B Only Statement 2 is correct.

C Both statements are correct.

Solution:

B is correct. Statement 2 is correct because actual costs and benefits may be available during retrospective analysis, allowing a more informed assessment of regulation. Statement 1 is incorrect because both indirect and implementation costs will be taken into account.

4.2 Analysis of Regulation[19]

In the previous section, we considered cost–benefit analysis of new regulations from the point of view of the regulator that is in the process of developing new regulations or is analyzing the impact of regulatory intervention retrospectively, after the event. In this section, we focus on the considerations that an analyst or investor could take into account when evaluating the effects of a specific regulation on a particular industry or company for the purpose of making an investment recommendation. In-depth coverage of industry and company analysis is featured elsewhere in the CFA Program curriculum.

The fact that rules and regulations can take different forms and can affect industries and individual companies in different ways adds to the complexity of the task. Analysts need to understand not just how regulation affects companies and industries at present; they should also be able to understand and anticipate the impact of proposed new or changing regulations on the future prospects for companies and industries. Having assessed the impact of regulations on the company and its prospects, analysts can then use suitable valuation tools to establish fair values for the business and make investment recommendations. Although no framework or template is adequate for all the possible scenarios, there are certain steps an analyst can take that are common to most circumstances.

Assessment of the likelihood of regulatory change

The analyst will need to assess the likelihood of the proposed regulation actually being implemented. Understanding the regulator's intentions, the cost–benefit analysis framework used by the regulator, and the extent of engagement with the regulated companies will help the analyst draw conclusions about the likelihood of the implementation of the proposed regulation. Where relevant, public and political pressure may also play a role in determining the likelihood that regulatory intervention will materialize.

Assessment of the impact of regulatory change on a sector

Industry and company analysis performed by an analyst (explored in depth elsewhere) will incorporate the analyst's or investor's opinion on the impact of regulations. The following text describes some, but not all, of the effects that regulations may have.

Impact on revenues Regulatory bodies sometimes introduce limits on prices, tariffs, rents, or fees that companies may charge, usually to protect consumers. Alternatively, certain products or services may be banned by the regulators, or companies may be

19 Christopher Decker (University of Oxford) contributed content for this discussion.

required to provide product descriptions that discourage their consumption (food or tobacco product labeling). The analyst would need to estimate the impact of such regulatory interventions on the companies' turnover, noting that not all entities in the sector would be affected equally. For example, telecommunication and utilities companies in Europe have been subject to caps on prices and tariffs in the last few decades. Another example is the former bi-annual pricing revision process in the pharmaceutical sector in Japan. The analyst would also need to be aware that if certain charges or fees are no longer allowed by the regulator, the companies may find alternative ways of generating revenues, helping to offset the negative impact of the regulation. For instance, the ban on commissions from financial product providers that financial advisory firms used to receive in parts of Europe led firms instead to charge their clients fees for giving financial advice. In this way, they could recoup some of the revenues they lost through the ban on commissions.

In some scenarios, pricing or charging may not be regulated or limited by the regulator in any way, but companies may be required to increase their pricing transparency and provide a detailed break-down of their fees. For example, power utilities in the United Kingdom must provide transparent monthly bills and, if applicable, suggest that their customers switch to a different tariff if the customer could benefit from a lower monthly payment by switching.

Another form of regulation that relates to the revenue line is the arrangement in which utility companies, often natural monopolies, are allowed to earn only a certain return on their assets.

Cost impact Compliance with some regulations results in additional costs for companies. These costs could take the form of higher operating expenses if, for example, manufactured products need to incorporate new safety features. Alternatively, these costs could take the form of higher capital expenditure if additional or new equipment is required. Analysts may also need to consider additional costs related to minimum wages, increased information disclosure and data protection requirements. Companies may also incur additional costs due to the need to use (or not to use) certain raw materials, introduce product features or subject themselves to regular inspections that make products or services safer for consumers. There may be costs related to more costly waste water treatment for certain industries. The analyst will try to estimate such costs and incorporate their impact into forecasts of a company's future profitability. When quantification of specific additional costs is not possible or relevant, the analyst will need to take into account the potentially reduced flexibility of the company's operations. It may also be possible for some companies to pass on some additional costs to their customers. The analyst will need to understand the competitive position of the industry.

Examples of regulations that companies may be subjected to include labor laws that may impose limits on hours of work, such as the maximum working week in parts of Europe or, more globally, limits on the number of work hours for pilots and cabin crew in the aviation industry. Of course, some companies that adhere to high environmental, social, and governance standards would incur some of those costs regardless of government regulation. In the financial services industry, the increasing requirements in relation to record keeping, data protection, risk control, and prevention of money laundering often result in significant additional personnel, training, and information technology infrastructure expenses.

Business risk Many industries have seen greater regulator involvement in the form of fines, requirements to pay compensation to customers, or bans on certain activities. Such events are difficult to forecast, and their impact may be hard to quantify and incorporate into future cash flow or growth forecasts. Analysts should take these types of events into account either by attempting to assign probabilities to them or by reflecting the risk in the discount rate used to value the company. For example, companies prone to a particular regulatory risk may deserve to trade on lower valuation multiples relative to peers or other industries. Or when future earnings or cash flows are discounted to present values, the discount rate should reflect an additional risk premium.

Example of Regulatory Analysis

The scenario outlined earlier (in "Illustration of Cost–Benefit Analysis") considered proposed changes to the regulations concerning the IPO market for smaller companies. An analyst may wish to evaluate the impact of such changes to the rules in order to analyze prospects for a particular company (issuer) or to assess the portfolio fund flows into or out of the smaller companies segment. The analyst may also want to understand the impact on companies in the brokerage business involved in that segment of the market.

The analyst might consider a series of questions. One set of questions could concern the design of the proposal, such as the following: Are there any exemptions from the new regulation, or will it apply to all small companies? What thresholds, if any, are used to determine whether a company is designated as small? Will the regulation be applied identically across all EU member states, or will each jurisdiction have an ability to tailor it to their own conditions? Will the regulation be subject to review or withdrawal later?

Another set of questions that an analyst might consider relates to the potential market and participant impacts: Who will benefit most from the regulation? Will it benefit high-growth companies in particular sectors or countries? What is the scale of the potential benefit associated with the regulation? What costs are associated with the regulation, and what is their scale? Will all participants face the same costs, or will they differ by market segment? Are the costs likely to be one-off in nature or recurring? Is the regulation likely to lead to market entry, expansion, or innovation by certain SMEs? Could the regulation lead to the potential exit of some existing providers of brokerage services or potentially crowd out other means of supplying capital and finance? How might the regulation change the behavior of SMEs? For example, will it reduce their reliance on bank loans?

Finally, the analyst might consider any wider impacts of the regulation on the market, industry, and society: Will the regulation affect financial stability or resilience? Could the regulation widen the opportunities for other investors, allowing them to better diversify their portfolios? Could there be potential spillover effects, allowing firms, for example, to shift between "junior markets" and more regulated markets? Could the impacts on financial markets be greater in some EU member states than in others? What are the possible macroeconomic impacts of the regulation? For example, could it improve capital inflows?

EXAMPLE 7

Analysis of Regulation

1 Jessica Wong, CFA, is an equity analyst responsible for the materials and industrial sectors in Europe. The regulatory authorities are preparing new rules on transportation, further limiting the age and exhaust emissions of the trucks used by industrial companies. What is likely to be of greatest concern to the analyst when evaluating the impact of the new rules on companies?

A Changes to costs related to the acquisition, operation, and maintenance of trucks

B Positive impact of reduced pollution on public health and subsequent health care cost savings in the wider society

   C The methods and techniques used by the regulator during its cost–benefit analysis

2 Which of the following questions would be the least relevant for the analyst to ask?

   A Will the costs be one-off in nature, or will they be recurring?

   B Will all companies in the sector be affected equally?

   C What is the legal status of the regulator? Is it a government department, a government agency, or a self-regulating organization?

Solution to 1:

A is correct. The analyst will want to understand and analyze the impact of the proposed regulations on the performance of companies she covers. Answers B and C relate to cost–benefit analysis carried out by the regulatory authority.

Solution to 2:

C is correct. The status of the regulator is the least important question. Answers A and B represent items that the analyst will find relevant when evaluating the impact of new regulations on the companies under coverage.

Regulators and the Regulated

How may regulation affect the economics of businesses?

One example is the effect of the SEC's Regulation National Market System (NMS) on competition among equity trading platforms in the United States. Regulation NMS, adopted in 2005, was intended to reflect technological advances and achieve the objectives of efficient, competitive, fair, and orderly markets. Since the 2005 adoption of Regulation NMS, the market share of the trading floor of the New York Stock Exchange (NYSE) has fallen substantially to account for a fraction of the overall NYSE volumes. Prior to Regulation NMS, NYSE "specialists" or market makers could take up to 30 seconds to react to orders sent by other platforms. The other platforms were checking whether the NYSE would execute at a more favorable price than the original platform had quoted. This process provided considerable opportunity for an NYSE specialist to observe subsequent pricing and to exploit the implicit optionality in the process. This process also made it hard for the rival platform to compete. Consequently, the NYSE could position itself to attract and concentrate much of the market liquidity, and so it came to resemble a natural monopoly. After Regulation NMS, which the NYSE had endorsed, the advantage to the NYSE diminished.[20] Because of the change in regulation, many new trading platforms developed and trading execution fragmented. Clearly, the structure of regulation plays a crucial role with respect to the viability of different order tactics and even the viability of the business models underlying different trading platforms.

The history of the money market mutual fund industry is another example of how regulation can affect business models. Money market mutual funds in the United States first arose in the early 1970s in response to Regulation Q, which imposed a ceiling on the interest rates paid by banks for various types of bank deposits. When market interest rates rose above the ceiling, there was considerable migration from bank deposits toward marketed fixed-income instruments, such as Treasury bills and notes. Money market mutual funds developed in response to the binding Regulation Q rate ceilings. During the 2008 global financial crisis, the collapse of a major US money market mutual fund

20 See the discussion of the impact of Regulation NMS in Angel, Harris, and Spatt (2011).

(the Reserve Fund) led to a run until the government launched a short-term insurance program to protect money market mutual fund balances. Government policy (motivated by an attempt to stabilize the financial system) helped protect this product. In response to resulting pressures from banks and the new advantage that the money market fund industry obtained, however, the Federal Deposit Insurance Corporation subsequently raised its insurance limit from \$100,000 to \$250,000. As this example illustrates, regulatory constraints have played a major role in the organization of short-term deposits in the United States. Changes in the effective regulatory structure have led to dramatic changes in the competitive landscape.

Government regulation can affect the structure of the industry. The issues can be seen in the pricing of joint products in the utility industries. For example, it can be difficult to separate fully the underlying economics associated with the production, transmission, and distribution of such services as electricity, telecommunications, and water. Suppose, for example, that there is a natural monopoly with respect to transportation (transmission and distribution) of a utility service but that there is competition in complementary activities (such as gas or electricity production or retail competition in telecommunications). How much should the provider of the natural monopoly services be able to obtain from the consumer or other companies providing upstream services, such as an energy product or access to a communication network? Although for some products there is increased and vigorous competition, these issues are still important with respect to the returns available from building various types of infrastructure. Although the market can sort out the allocation of profits and pricing across stages when there is vigorous competition at each stage, these issues are challenging in the case of a natural monopoly. Monopoly power is at the root of one of the most important traditional uses of regulation—to set pricing and returns at utility providers. In many jurisdictions, a government regulator sets or approves public utility prices because a utility provider has a monopolistic position.

SUMMARY

Knowledge of regulation is important because regulation has potentially far-reaching and significant effects. These effects can range from macro-level effects on the economy to micro-level effects on individual entities and securities.

Regulation originates from a variety of sources and in a variety of areas. A framework that includes types of regulators and regulation as well as areas of regulation that may affect the entity of interest (including the economy as an entity) is useful. The framework will help in assessing possible effects of new regulation. It can also help in assessing the effects of regulation on various entities.

More than one regulator may develop regulations in response to a particular issue. Each of the relevant regulators may have different objectives and choose to address the issue using different regulatory tools.

In developing regulations, the regulator should consider costs and benefits. In the analysis, the net regulatory burden (private costs less private benefits of regulation) may also be relevant. Potential costs and benefits, regardless of the perspective, may be difficult to assess. A critical aspect of regulatory analysis, however, is assessing the costs and benefits of regulation.

The following are some key points of the reading.

- The existence of informational frictions and externalities creates a need for regulation. Regulation is expected to have societal benefits and should be assessed using cost–benefit analysis.

- The regulation of securities markets and financial institutions is extensive and complex because of the consequences of failures in the financial system. These consequences include financial losses, loss of confidence, and disruption of commerce.
- The focus of regulators in financial markets includes prudential supervision, financial stability, market integrity, and economic growth.
- Regulatory competition is competition among different regulatory bodies to use regulation in order to attract certain entities.
- The breadth of regulation of commerce necessitates the use of a framework that identifies potential areas of regulation. This framework can be referenced to identify specific areas of regulation, both existing and anticipated, that may affect the entity of interest.
- Legislative bodies, regulatory bodies, and courts typically enact regulation.
- Regulatory bodies include government agencies and independent regulators granted authority by a government or governmental agency. Some independent regulators are self-regulating organizations.
- Typically, legislative bodies enact broad laws or statutes. Regulatory bodies issue administrative regulations, often implementing statutes. Courts interpret statutes and administrative regulations; these interpretations may result in judicial law.
- Interdependence in the actions and potentially conflicting objectives of regulators is an important consideration for regulators, regulated entities, and those assessing the effects of regulation.
- Regulation that arises to enhance the interests of regulated entities reflects regulatory capture.
- Regulators have responsibility for both substantive and procedural laws. The former focuses on rights and responsibilities of entities and relationships among entities. The latter focuses on the protection and enforcement of the former.
- Regulatory arbitrage is the use of regulation by an entity to exploit differences in economic substance and regulatory interpretation or in regulatory regimes to the entity's benefit.
- There are many regulatory tools available to regulators, including regulatory mandates and restrictions on behaviors, provision of public goods, and public financing of private projects.
- The choice of regulatory tool should be consistent with maintaining a stable regulatory environment. "Stable" does not mean unchanging but, rather, refers to desirable attributes of regulation, including predictability, effectiveness in achieving objectives, time consistency, and enforceability.
- In assessing regulation and regulatory outcomes, regulators should conduct ongoing cost–benefit analyses, develop techniques to enhance the measurement of these outcomes, and use economic principles to guide them.
- Net regulatory burden to the entity of interest is an important consideration for analysts.

REFERENCES

Angel, James, Lawrence Harris, and Chester Spatt. 2011. "Equity Trading in the 21st Century." *Quarterly Journal of Finance* 1 (1): 1–53.

Carson, John W. 2011. "*Self-Regulation in Securities Markets.*" World Bank Financial Sector Policy Group Working Paper 5542 (1 January).

Spatt, Chester. 2009. "Regulatory Conflict: Market Integrity vs. Financial Stability." *University of Pittsburgh Law Review. University of Pittsburgh. School of Law* 71 (3): 625–39.

Stigler, George J. 1971. "The Economic Theory of Regulation." *Bell Journal of Economics* 2 (1): 3–21.

PRACTICE PROBLEMS

The following information relates to Questions 1–4

Tiu Asset Management (TAM), a hypothetical financial services firm, recently hired Jonna Yun. Yun is a member of TAM's Global Equity portfolio team and is assigned the task of analyzing the effects of regulation on the financial services sector of a particular country. In her first report to the team, Yun makes the following statements:

Statement 1 The country's regulator, a government agency, concerned about systemic risk, is calling for an accelerated adoption of centralized derivatives settlement (as opposed to bilateral settlement between two counterparties)—a more stringent rule—ahead of other major countries that are considering a similar move.

Statement 2 Regulators use various tools to intervene in the financial services sector.

Statement 3 Regulations may bring benefits to the economy, but they may also have unanticipated costs.

Statement 4 The country's regulatory authorities are considering a regulation that is similar to Regulation Q in the United States, which imposed a ceiling on interest rates paid by banks for certain bank deposits.

1 What is the *most likely* basis for the concerns noted in Statement 1?

A Externalities

B Regulatory arbitrage

C Informational friction

2 The tools *least likely* to be used by regulators to intervene in financial markets owing to informational frictions are:

A blackout periods.

B capital requirements.

C insider-trading restrictions.

3 Which of the following is *most likely* an unanticipated effect of regulation?

A Hiring compliance lawyers

B Setting legal standards for contracts

C Establishing employers' rights and responsibilities

4 After Regulation Q was imposed, the demand for money market funds *most likely*:

A increased.

B decreased.

C remained unchanged.

The following information relates to Questions 5–11

Cate Stephenson is an analyst in the economics research division of an international securities firm. She is conducting research on the regulatory environment in certain European countries. Stephenson begins with an analysis of a hypothetical country, Genovia.

Genovia has recently introduced a new accounting statute. In Genovia, there is an independent regulator—"Le régulateur." Le régulateur is not a self-regulating organization (SRO). There is also an SRO—"L'organisation." L'organisation is not an independent regulator.

In her research report, Stephenson makes the following statements:

Statement 1	Le régulateur has been given legal authority by the government to enforce the new statute.
Statement 2	L'organisation issues administrative regulations related to the new statute using government funding.
Statement 3	L'organisation has member companies that accept the authorization of L'organisation to set and enforce rules and standards.

Stephenson and her supervisor discuss the intended and unintended effects of implementing the new statute, and Stephenson makes two comments.

Comment 1	It is likely that some unintended consequences will be identified in regulatory filings prior to implementation of the new legislation.
Comment 2	Indirect costs arise because of unintended consequences and may result in high unanticipated costs.

Stephenson reads a report titled "International Trade," which has three sections about Genovia's policies and regulations.

- The first section of the report discusses policies that legislators may implement to accomplish Genovia's objective of promoting free trade on industrial goods.
- The second section of the report covers corporate domicile. Stephenson learns that regulators in Genovia recently amended regulations to encourage foreign businesses to move their corporate domicile to Genovia.
- The third section of the report reviews the regulation of commerce. Genovia's goal is to establish an environment that encourages foreign businesses to increase trade with domestic businesses. Stephenson considers two features of Genovia's regulation of commerce.

Feature 1	Recent court decisions have upheld financial liability and bankruptcy laws.
Feature 2	A legal structure that governs contracts and each party's rights is in place.

Stephenson then reviews two initiatives by Genovia to improve domestic policies and regulations.

- The first initiative by Genovia is its passage of conflict-of-interest regulations. Regulators implement regulatory restrictions and regulatory mandates that apply to employees of securities firms. One of Stephenson's research colleagues writes reports on a company in which he owns shares.

- The second initiative by Genovia is to reduce pollution and promote renewable electricity generation. Two years ago, the government implemented taxes on fossil fuels and subsidies on hydropower and other renewables. Stephenson reviews the changes in sources of electricity production since the policies were introduced, shown in Exhibit 1.

Exhibit 1 Genovia's Domestic Electricity Generation Production

Sector	Year 0	Year 1	Year 2
Fossil fuels	462	446	426
Hydropower	186	231	273
Other renewables	97	120	154
Total	745	797	853

Note: Amounts are in terawatt hours (TWh).

5 Which of Stephenson's statements regarding Le régulateur and L'organisation is correct?
   A Only Statement 1 is correct.
   B Only Statement 2 is correct.
   C Both Statement 1 and Statement 2.

6 Is Stephenson's Statement 3 correct?
   A Yes
   B No, because L'organisation is given the authority to enforce regulations by a government agency
   C No, because pressure from its member companies prevents L'organisation from enforcing its rules and standards

7 Which of Stephenson's comments to her supervisor is most likely correct?
   A Only Comment 1 is correct.
   B Only Comment 2 is correct.
   C Both Comment 1 and Comment 2.

8 Which of the following policies would *best* address Genovia's objective of promoting free trade on industrial goods?
   A Imposing tariffs on foreign-produced goods
   B Allowing a floating currency
   C Providing subsidies to domestic companies

9 By amending regulations to encourage foreign businesses to change their corporate domicile, regulators are encouraging regulatory:
   A capture.
   B arbitrage.
   C competition.

10 Which feature discussed in the third section of "International Trade" will *most likely* help Genovia achieve its goal of encouraging foreign businesses to increase trade with domestic businesses?
   A Only Feature 1
   B Only Feature 2

C Both Feature 1 and Feature 2

11 Based on Exhibit 1, which government policy has been *most effective* in helping Genovia achieve its second initiative?

A Tax on fossil fuels

B Subsidy on hydropower

C Subsidy on other renewables

SOLUTIONS

1 B is correct. Firms based in the country are likely to be concerned because of the earlier timing of the application of new (more stringent) regulations in the country than in other large countries. With more stringent regulations, some business may flow to less stringent regulatory environments or jurisdictions.

2 A is correct. Blackout periods are established by *companies* in response to concerns about insider trading. Thus, blackout periods are not a tool used by regulators to intervene in the financial services sector. Capital requirements are used by government regulators to reduce systemic risk and financial contagion. Insider-trading restrictions are used by regulators concerned about insiders using their greater knowledge to the disadvantage of others; insider-trading restrictions respond to informational frictions.

3 A is correct. The hiring of more lawyers to deal with compliance is an example of an "unintended" implementation cost. Establishing legal standards for contracts and employers' rights and responsibilities are objectives (intended consequences) of some regulation.

4 A is correct. Regulation Q set a ceiling on the interest rates paid by banks for various types of deposits, which resulted in investors' shifting funds to money market funds.

5 A is correct. Le régulateur, as an independent regulator but not an SRO, has legal authority from the Genovia government to regulate. Therefore, Le régulateur both enacts and enforces regulations related to the new accounting statute in Genovia.

6 A is correct. L'organisation is an SRO but not an independent regulator, so it is a private entity that is not affiliated with Genovia's government. SROs that are not independent regulators receive authority from their members, who agree to comply with the organization's rules and standards and its enforcement thereof.

7 C is correct. Comment 1 is correct because regulatory filings, in response to proposed regulations, often identify at least some of the unintended consequences prior to the implementation of the regulation. Comment 2 is correct because the cost of unintended consequences, including both indirect costs and unanticipated implementation costs, can be high.

8 B is correct. A floating currency allows international trade in Genovia to be market based. International disputes about whether a country is manipulating or fixing its currency price often center on issues related to competitiveness.

9 C is correct. Regulatory competition describes actions by regulators to encourage behaviors. Regulators may compete to provide a regulatory environment designed to attract certain entities (regulatory competition). By amending regulations, Genovia's regulators seek to encourage foreign companies to change their corporate domicile.

10 C is correct. Genovia needs unambiguous laws concerning financial liability and bankruptcy to encourage foreign businesses to enter into contracts, particularly those that are long term and may involve sunk costs. The court decisions help Genovia achieve its goal. Also, clearly defined rules governing contracts, their interpretation, and each party's legal rights under a contract are necessary. Thus, both features help Genovia achieve its goal.

11 C is correct. At the end of Year 2, the compound annual growth rate (CAGR) for each sector is calculated as follows: (Year 2/Year 0)$^{0.5}$ – 1.

Fossil fuels: $(426/462)^{0.5} - 1 = -4\%$
Hydropower: $(273/186)^{0.5} - 1 = 21\%$
Other renewables: $(154/97)^{0.5} - 1 = 26\%$

The CAGR indicates that the 26% increase in production from the subsidy on other renewables has been more effective than the 4% decrease in production from the tax on fossil fuels or the 21% increase in production from the subsidy on hydropower. Thus, the subsidy on other renewables of 26% is the highest, indicating that this policy has been the most effective in helping Genovia achieve its second initiative.

APPENDICES

Appendix A
Cumulative Probabilities for a Standard Normal Distribution
$P(Z \leq x) = N(x)$ for $x \geq 0$ or $P(Z \leq z) = N(z)$ for $z \geq 0$

x or z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.50	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.60	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.70	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.80	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.90	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

For example, to find the *z*-value leaving 2.5 percent of the area/probability in the upper tail, find the element 0.9750 in the body of the table. Read 1.90 at the left end of the element's row and 0.06 at the top of the element's column, to give 1.90 + 0.06 = 1.96. *Table generated with Excel.*

Appendix A (continued)
Cumulative Probabilities for a Standard Normal Distribution
P(*Z* ≤ x) = *N*(x) for x ≤ 0 or *P*(*Z* ≤ z) = *N*(z) for z ≤ 0

x or z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
−0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.60	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−1.00	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.20	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−2.00	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−3.10	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.20	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.30	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.40	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.50	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.60	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.70	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.80	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.90	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
−4.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

For example, to find the *z*-value leaving 2.5 percent of the area/probability in the lower tail, find the element 0.0250 in the body of the table. Read −1.90 at the left end of the element's row and 0.06 at the top of the element's column, to give −1.90 − 0.06 = −1.96. *Table generated with Excel.*

Appendix B
Table of the Student's *t*-Distribution (One-Tailed Probabilities)

df	p = 0.10	p = 0.05	p = 0.025	p = 0.01	p = 0.005	df	p = 0.10	p = 0.05	p = 0.025	p = 0.01	p = 0.005
1	3.078	6.314	12.706	31.821	63.657	**31**	1.309	1.696	2.040	2.453	2.744
2	1.886	2.920	4.303	6.965	9.925	**32**	1.309	1.694	2.037	2.449	2.738
3	1.638	2.353	3.182	4.541	5.841	**33**	1.308	1.692	2.035	2.445	2.733
4	1.533	2.132	2.776	3.747	4.604	**34**	1.307	1.691	2.032	2.441	2.728
5	1.476	2.015	2.571	3.365	4.032	**35**	1.306	1.690	2.030	2.438	2.724
6	1.440	1.943	2.447	3.143	3.707	**36**	1.306	1.688	2.028	2.434	2.719
7	1.415	1.895	2.365	2.998	3.499	**37**	1.305	1.687	2.026	2.431	2.715
8	1.397	1.860	2.306	2.896	3.355	**38**	1.304	1.686	2.024	2.429	2.712
9	1.383	1.833	2.262	2.821	3.250	**39**	1.304	1.685	2.023	2.426	2.708
10	1.372	1.812	2.228	2.764	3.169	**40**	1.303	1.684	2.021	2.423	2.704
11	1.363	1.796	2.201	2.718	3.106	**41**	1.303	1.683	2.020	2.421	2.701
12	1.356	1.782	2.179	2.681	3.055	**42**	1.302	1.682	2.018	2.418	2.698
13	1.350	1.771	2.160	2.650	3.012	**43**	1.302	1.681	2.017	2.416	2.695
14	1.345	1.761	2.145	2.624	2.977	**44**	1.301	1.680	2.015	2.414	2.692
15	1.341	1.753	2.131	2.602	2.947	**45**	1.301	1.679	2.014	2.412	2.690
16	1.337	1.746	2.120	2.583	2.921	**46**	1.300	1.679	2.013	2.410	2.687
17	1.333	1.740	2.110	2.567	2.898	**47**	1.300	1.678	2.012	2.408	2.685
18	1.330	1.734	2.101	2.552	2.878	**48**	1.299	1.677	2.011	2.407	2.682
19	1.328	1.729	2.093	2.539	2.861	**49**	1.299	1.677	2.010	2.405	2.680
20	1.325	1.725	2.086	2.528	2.845	**50**	1.299	1.676	2.009	2.403	2.678
21	1.323	1.721	2.080	2.518	2.831	**60**	1.296	1.671	2.000	2.390	2.660
22	1.321	1.717	2.074	2.508	2.819	**70**	1.294	1.667	1.994	2.381	2.648
23	1.319	1.714	2.069	2.500	2.807	**80**	1.292	1.664	1.990	2.374	2.639
24	1.318	1.711	2.064	2.492	2.797	**90**	1.291	1.662	1.987	2.368	2.632
25	1.316	1.708	2.060	2.485	2.787	**100**	1.290	1.660	1.984	2.364	2.626
26	1.315	1.706	2.056	2.479	2.779	**110**	1.289	1.659	1.982	2.361	2.621
27	1.314	1.703	2.052	2.473	2.771	**120**	1.289	1.658	1.980	2.358	2.617
28	1.313	1.701	2.048	2.467	2.763	**200**	1.286	1.653	1.972	2.345	2.601
29	1.311	1.699	2.045	2.462	2.756	∞	1.282	1.645	1.960	2.326	2.576
30	1.310	1.697	2.042	2.457	2.750						

To find a critical t-value, enter the table with df and a specified value for α, the significance level. For example, with 5 df, α = 0.05 and a one-tailed test, the desired probability in the tail would be p = 0.05 and the critical t-value would be $t(5, 0.05) = 2.015$. With α =0.05 and a two-tailed test, the desired probability in each tail would be $p = 0.025 = \alpha/2$, giving $t(5, 0.025) = 2.571$.
Table generated using Excel.

Quantitative Methods for Investment Analysis, Second Edition, by Richard A. DeFusco, CFA, Dennis W. McLeavey, CFA, Jerald E. Pinto, CFA, and David E. Runkle, CFA. Copyright © 2004 by CFA Institute.

Appendix C
Values of χ^2 (Degrees of Freedom, Level of Significance)

Degrees of Freedom	Probability in Right Tail 0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.000157	0.000982	0.003932	0.0158	2.706	3.841	5.024	6.635	7.879
2	0.020100	0.050636	0.102586	0.2107	4.605	5.991	7.378	9.210	10.597
3	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.170

To have a probability of 0.05 in the right tail when df = 5, the tabled value is $\chi^2(5, 0.05) = 11.070$.

Quantitative Methods for Investment Analysis, Second Edition, by Richard A. DeFusco, CFA, Dennis W. McLeavey, CFA, Jerald E. Pinto, CFA, and David E. Runkle, CFA. Copyright © 2004 by CFA Institute.

Appendix D
Table of the *F*-Distribution

Panel A. Critical values for right-hand tail area equal to 0.05 — **Numerator: df_1 and Denominator: df_2**

	df1:1	2	3	4	5	6	7	8	9	10	11	12	15	20	21	22	23	24	25	30	40	60	120	˜
df2: 1	161	200	216	225	230	234	237	239	241	242	243	244	246	248	248	249	249	249	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.70	8.66	8.65	8.65	8.64	8.64	8.63	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.86	5.80	5.79	5.79	5.78	5.77	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.62	4.56	4.55	4.54	4.53	4.53	4.52	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.94	3.87	3.86	3.86	3.85	3.84	3.83	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.51	3.44	3.43	3.43	3.42	3.41	3.40	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.22	3.15	3.14	3.13	3.12	3.12	3.11	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.01	2.94	2.93	2.92	2.91	2.90	2.89	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.85	2.77	2.76	2.75	2.75	2.74	2.73	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.72	2.65	2.64	2.63	2.62	2.61	2.60	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.62	2.54	2.53	2.52	2.51	2.51	2.50	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.53	2.46	2.45	2.44	2.43	2.42	2.41	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.46	2.39	2.38	2.37	2.36	2.35	2.34	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.40	2.33	2.32	2.31	2.30	2.29	2.28	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.35	2.28	2.26	2.25	2.24	2.24	2.23	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.31	2.23	2.22	2.21	2.20	2.19	2.18	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.27	2.19	2.18	2.17	2.16	2.15	2.14	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.23	2.16	2.14	2.13	2.12	2.11	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.20	2.12	2.11	2.10	2.09	2.08	2.07	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.18	2.10	2.08	2.07	2.06	2.05	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.15	2.07	2.06	2.05	2.04	2.03	2.02	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.13	2.05	2.04	2.02	2.01	2.01	2.00	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.11	2.03	2.01	2.00	1.99	1.98	1.97	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.09	2.01	2.00	1.98	1.97	1.96	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.01	1.93	1.92	1.91	1.90	1.89	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.92	1.84	1.83	1.81	1.80	1.79	1.78	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.84	1.75	1.73	1.72	1.71	1.70	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.83	1.75	1.66	1.64	1.63	1.62	1.61	1.60	1.55	1.50	1.43	1.35	1.25
Infinity	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.67	1.57	1.56	1.54	1.53	1.52	1.51	1.46	1.39	1.32	1.22	1.00

Appendix D (continued)
Table of the *F*-Distribution

Panel B. Critical values for right-hand tail area equal to 0.025 — Numerator: df_1 and Denominator: df_2

	df1: 1	2	3	4	5	6	7	8	9	10	11	12	15	20	21	22	23	24	25	30	40	60	120	˜
df2: 1	648	799	864	900	922	937	948	957	963	969	973	977	985	993	994	995	996	997	998	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.41	39.43	39.45	39.45	39.45	39.45	39.46	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.37	14.34	14.25	14.17	14.16	14.14	14.13	14.12	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.66	8.56	8.55	8.53	8.52	8.51	8.50	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.43	6.33	6.31	6.30	6.29	6.28	6.27	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.27	5.17	5.15	5.14	5.13	5.12	5.11	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.57	4.47	4.45	4.44	4.43	4.41	4.40	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.10	4.00	3.98	3.97	3.96	3.95	3.94	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.77	3.67	3.65	3.64	3.63	3.61	3.60	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.52	3.42	3.40	3.39	3.38	3.37	3.35	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.33	3.23	3.21	3.20	3.18	3.17	3.16	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.18	3.07	3.06	3.04	3.03	3.02	3.01	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.05	2.95	2.93	2.92	2.91	2.89	2.88	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	2.95	2.84	2.83	2.81	2.80	2.79	2.78	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.86	2.76	2.74	2.73	2.71	2.70	2.69	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.79	2.68	2.67	2.65	2.64	2.63	2.61	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.72	2.62	2.60	2.59	2.57	2.56	2.55	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.67	2.56	2.54	2.53	2.52	2.50	2.49	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.62	2.51	2.49	2.48	2.46	2.45	2.44	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.57	2.46	2.45	2.43	2.42	2.41	2.40	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68	2.64	2.53	2.42	2.41	2.39	2.38	2.37	2.36	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.50	2.39	2.37	2.36	2.34	2.33	2.32	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62	2.57	2.47	2.36	2.34	2.33	2.31	2.30	2.29	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.44	2.33	2.31	2.30	2.28	2.27	2.26	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.56	2.51	2.41	2.30	2.28	2.27	2.26	2.24	2.23	2.18	2.12	2.05	1.98	1.91
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.31	2.20	2.18	2.16	2.15	2.14	2.12	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.18	2.07	2.05	2.03	2.02	2.01	1.99	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.06	1.94	1.93	1.91	1.90	1.88	1.87	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.10	2.05	1.94	1.82	1.81	1.79	1.77	1.76	1.75	1.69	1.61	1.53	1.43	1.31
Infinity	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.99	1.94	1.83	1.71	1.69	1.67	1.66	1.64	1.63	1.57	1.48	1.39	1.27	1.00

Appendix D (continued)
Table of the *F*-Distribution

Panel C. Critical values for right-hand tail area equal to 0.01 — Numerator: df_1 and Denominator: df_2

	df1: 1	2	3	4	5	6	7	8	9	10	11	12	15	20	21	22	23	24	25	30	40	60	120	˝
df2: 1	4052	5000	5403	5625	5764	5859	5928	5982	6023	6056	6083	6106	6157	6209	6216	6223	6229	6235	6240	6261	6287	6313	6339	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1	26.9	26.7	26.7	26.6	26.6	26.6	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.5	14.4	14.2	14.0	14.0	14.0	13.9	13.9	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	10.0	9.89	9.72	9.55	9.53	9.51	9.49	9.47	9.45	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.56	7.40	7.37	7.35	7.33	7.31	7.30	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.31	6.16	6.13	6.11	6.09	6.07	6.06	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.52	5.36	5.34	5.32	5.30	5.28	5.26	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	4.96	4.81	4.79	4.77	4.75	4.73	4.71	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.56	4.41	4.38	4.36	4.34	4.33	4.31	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.25	4.10	4.08	4.06	4.04	4.02	4.01	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.01	3.86	3.84	3.82	3.80	3.78	3.76	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.82	3.66	3.64	3.62	3.60	3.59	3.57	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.66	3.51	3.48	3.46	3.44	3.43	3.41	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.52	3.37	3.35	3.33	3.31	3.29	3.28	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.41	3.26	3.24	3.22	3.20	3.18	3.16	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.31	3.16	3.14	3.12	3.10	3.08	3.07	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.23	3.08	3.05	3.03	3.02	3.00	2.98	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.15	3.00	2.98	2.96	2.94	2.92	2.91	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.09	2.94	2.92	2.90	2.88	2.86	2.84	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.03	2.88	2.86	2.84	2.82	2.80	2.79	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	2.98	2.83	2.81	2.78	2.77	2.75	2.73	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.93	2.78	2.76	2.74	2.72	2.70	2.69	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.89	2.74	2.72	2.70	2.68	2.66	2.64	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.85	2.70	2.68	2.66	2.64	2.62	2.60	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.70	2.55	2.53	2.51	2.49	2.47	2.45	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.52	2.37	2.35	2.33	2.31	2.29	2.27	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.35	2.20	2.17	2.15	2.13	2.12	2.10	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.19	2.03	2.01	1.99	1.97	1.95	1.93	1.86	1.76	1.66	1.53	1.38
Infinity	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18	2.04	1.88	1.85	1.83	1.81	1.79	1.77	1.70	1.59	1.47	1.32	1.00

Appendix D (continued)
Table of the *F*-Distribution

Panel D. Critical values for right-hand tail area equal to 0.005 — **Numerator: df_1 and Denominator: df_2**

	df1: 1	2	3	4	5	6	7	8	9	10	11	12	15	20	21	22	23	24	25	30	40	60	120	˝
df2: 1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24222	24334	24426	24630	24836	24863	24892	24915	24940	24959	25044	25146	25253	25359	25464
2	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	200
3	55.55	49.80	47.47	46.20	45.39	44.84	44.43	44.13	43.88	43.68	43.52	43.39	43.08	42.78	42.73	42.69	42.66	42.62	42.59	42.47	42.31	42.15	41.99	41.83
4	31.33	26.28	24.26	23.15	22.46	21.98	21.62	21.35	21.14	20.97	20.82	20.70	20.44	20.17	20.13	20.09	20.06	20.03	20.00	19.89	19.75	19.61	19.47	19.32
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.49	13.38	13.15	12.90	12.87	12.84	12.81	12.78	12.76	12.66	12.53	12.40	12.27	12.14
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.13	10.03	9.81	9.59	9.56	9.53	9.50	9.47	9.45	9.36	9.24	9.12	9.00	8.88
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.27	8.18	7.97	7.75	7.72	7.69	7.67	7.64	7.62	7.53	7.42	7.31	7.19	7.08
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.10	7.01	6.81	6.61	6.58	6.55	6.53	6.50	6.48	6.40	6.29	6.18	6.06	5.95
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.31	6.23	6.03	5.83	5.80	5.78	5.75	5.73	5.71	5.62	5.52	5.41	5.30	5.19
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.75	5.66	5.47	5.27	5.25	5.22	5.20	5.17	5.15	5.07	4.97	4.86	4.75	4.64
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.32	5.24	5.05	4.86	4.83	4.80	4.78	4.76	4.74	4.65	4.55	4.45	4.34	4.23
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.99	4.91	4.72	4.53	4.50	4.48	4.45	4.43	4.41	4.33	4.23	4.12	4.01	3.90
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.72	4.64	4.46	4.27	4.24	4.22	4.19	4.17	4.15	4.07	3.97	3.87	3.76	3.65
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.51	4.43	4.25	4.06	4.03	4.01	3.98	3.96	3.94	3.86	3.76	3.66	3.55	3.44
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.33	4.25	4.07	3.88	3.86	3.83	3.81	3.79	3.77	3.69	3.59	3.48	3.37	3.26
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.18	4.10	3.92	3.73	3.71	3.68	3.66	3.64	3.62	3.54	3.44	3.33	3.22	3.11
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	4.05	3.97	3.79	3.61	3.58	3.56	3.53	3.51	3.49	3.41	3.31	3.21	3.10	2.98
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.94	3.86	3.68	3.50	3.47	3.45	3.42	3.40	3.38	3.30	3.20	3.10	2.99	2.87
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.84	3.76	3.59	3.40	3.37	3.35	3.33	3.31	3.29	3.21	3.11	3.00	2.89	2.78
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.76	3.68	3.50	3.32	3.29	3.27	3.24	3.22	3.20	3.12	3.02	2.92	2.81	2.69
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.68	3.60	3.43	3.24	3.22	3.19	3.17	3.15	3.13	3.05	2.95	2.84	2.73	2.61
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.61	3.54	3.36	3.18	3.15	3.12	3.10	3.08	3.06	2.98	2.88	2.77	2.66	2.55
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.55	3.47	3.30	3.12	3.09	3.06	3.04	3.02	3.00	2.92	2.82	2.71	2.60	2.48
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.50	3.42	3.25	3.06	3.04	3.01	2.99	2.97	2.95	2.87	2.77	2.66	2.55	2.43
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.45	3.37	3.20	3.01	2.99	2.96	2.94	2.92	2.90	2.82	2.72	2.61	2.50	2.38
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.25	3.18	3.01	2.82	2.80	2.77	2.75	2.73	2.71	2.63	2.52	2.42	2.30	2.18
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	3.03	2.95	2.78	2.60	2.57	2.55	2.52	2.50	2.48	2.40	2.30	2.18	2.06	1.93
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.82	2.74	2.57	2.39	2.36	2.33	2.31	2.29	2.27	2.19	2.08	1.96	1.83	1.69
120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.62	2.54	2.37	2.19	2.16	2.13	2.11	2.09	2.07	1.98	1.87	1.75	1.61	1.43
Infinity	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.43	2.36	2.19	2.00	1.97	1.95	1.92	1.90	1.88	1.79	1.67	1.53	1.36	1.00

With 1 degree of freedom (df) in the numerator and 3 df in the denominator, the critical *F*-value is 10.1 for a right-hand tail area equal to 0.05.

Appendix E
Critical Values for the Durbin-Watson Statistic (α = .05)

	K = 1		K = 2		K = 3		K = 4		K = 5	
n	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Note: K = the number of slope parameters in the model.

Source: From J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression, II." *Biometrika* 38 (1951): 159–178.