UNIVERSITY OF LONDON | INTERNATIONAL PROGRAMMES

Quantitative finance

A. Patton

FN3142

2015

Undergraduate study in
Economics, Management, Finance and the Social Sciences

This subject guide is for a 300 course offered as part of the University of London International Programmes in Economics, Management, Finance and the Social Sciences. This is equivalent to Level 6 within the Framework for Higher Education Qualifications in England, Wales and Northern Ireland (FHEQ).

For more information about the University of London International Programmes undergraduate study in Economics, Management, Finance and the Social Sciences, see: www.londoninternational.ac.uk

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

This guide was prepared for the University of London International Programmes by:

A. Patton, Department of Economics, Duke University.

This is one of a series of subject guides published by the University. We regret that due to pressure of work the author is unable to enter into any correspondence relating to, or arising from, the guide. If you have any comments on this subject guide, favourable or unfavourable, please use the form at the back of this guide.

University of London International Programmes
Publications Office
Stewart House
32 Russell Square
London WC1B 5DN
United Kingdom
www.londoninternational.ac.uk

Published by: University of London

Contents

1 Introduction 1

1.1 Route map to the guide . . . 1

1.2 Why study quantitative finance? . . . 1

1.3 Syllabus . . . 2

1.4 Aims of the course . . . 2

1.5 Learning outcomes for the course . . . 2

1.6 Overview of learning resources . . . 3

1.6.1 The subject guide . . . 3

1.6.2 Essential reading . . . 3

1.6.3 Further reading . . . 3

1.6.4 Online study resources . . . 4

1.7 The structure of the subject guide . . . 5

1.8 Examination advice . . . 7

2 Financial econometrics concepts and statistics review 9

2.1 Introduction . . . 9

2.1.1 Aims of the chapter . . . 9

2.1.2 Learning outcomes . . . 9

2.1.3 Essential reading . . . 10

2.1.4 Further reading . . . 10

2.1.5 References cited . . . 10

2.2 What is financial econometrics? . . . 10

2.3 Some important concepts . . . 10

2.4 Forecasting returns and prices . . . 12

2.5 Revision of basic statistics . . . 15

2.5.1 Random variables, CDFs, PMFs, PDFs . . . 15

2.5.2 Means, variances, and other moments . . . 18

2.5.3 Multiple random variables, bivariate CDFs, covariances . . . 22

2.5.4 Moments of random vectors . . . 24

2.6 Overview of chapter . . . 25

2.7 Reminder of learning outcomes . . . 25
2.8 Test your knowledge and understanding . . . 26
2.9 Solutions to activities . . . 26

3 Basic time series concepts 31
3.1 Introduction . . . 31
3.1.1 Aims of the chapter . . . 31
3.1.2 Learning outcomes . . . 31
3.1.3 Essential reading . . . 31
3.1.4 Further reading . . . 32
3.2 Covariance stationary time series . . . 32
3.3 The Law of Iterated Expectations . . . 33
3.3.1 Refresher: Rules for expectations, variances and covariances . . . 34
3.4 White noise and other innovation series . . . 34
3.5 Application to an AR(1) process . . . 35
3.6 Overview of chapter . . . 37
3.7 Reminder of learning outcomes . . . 37
3.8 Test your knowledge and understanding . . . 38
3.9 Solutions to activities . . . 38

4 ARMA processes 41
4.1 Introduction . . . 41
4.1.1 Aims of the chapter . . . 41
4.1.2 Learning outcomes . . . 41
4.1.3 Essential reading . . . 41
4.1.4 Further reading . . . 42
4.2 Autoregressive-moving average (ARMA) processes . . . 42
4.2.1 Autoregressive (AR) processes . . . 42
4.2.2 The MA(1) process . . . 42
4.2.3 Moving average (MA) processes . . . 43
4.2.4 ARMA processes . . . 43
4.3 Autocovariance functions . . . 43
4.4 Predictability, R^2 and ARMA processes . . . 47
4.5 Choosing the best ARMA model . . . 48
4.6 Overview of chapter . . . 50
4.7 Reminder of learning outcomes . . . 50

4.8 Test your knowledge and understanding 50
4.9 Solutions to activities 51

5 Empirical features of financial asset returns 55
5.1 Introduction 55
5.1.1 Aims of the chapter 55
5.1.2 Learning outcomes 55
5.1.3 Essential reading 55
5.1.4 Further reading 55
5.1.5 References cited 56
5.2 Summary statistics 56
5.2.1 Sample moments 56
5.2.2 Sample quantiles 58
5.3 Jarque-Bera test for normality 59
5.4 A detailed look at some financial data 59
5.5 Overview of chapter 62
5.6 Reminder of learning outcomes 62
5.7 Test your knowledge and understanding 62
5.8 Solutions to activities 63

6 Testing for predictability in financial time series 65
6.1 Introduction 65
6.1.1 Aims of the chapter 65
6.1.2 Learning outcomes 65
6.1.3 Essential reading 65
6.1.4 Further reading 65
6.1.5 References cited 66
6.2 Sample autocorrelations 66
6.3 Tests on individual autocorrelation coefficients 67
6.4 Joint tests on many autocorrelations 69
6.5 Testing for predictability in our financial data 71
6.6 Overview of chapter 73
6.7 Reminder of learning outcomes 73
6.8 Test your knowledge and understanding 73
6.9 Solutions to activities 76

7 The efficient markets hypothesis and market predictability 79

7.1 Introduction . . . 79

7.1.1 Aims of the chapter . . . 79

7.1.2 Learning outcomes . . . 79

7.1.3 Essential reading . . . 79

7.1.4 Further reading . . . 80

7.1.5 References cited . . . 81

7.2 The efficient markets hypothesis . . . 81

7.3 Standard definitions of market efficiency . . . 81

7.3.1 Roberts, Jensen and Malkiel definitions . . . 82

7.4 The EMH and ‘bubbles’ . . . 83

7.5 Transaction costs and other market frictions . . . 84

7.6 New forecasting models and model selection . . . 85

7.7 Ephemeral forecastability . . . 86

7.8 Data snooping . . . 87

7.8.1 Refresher: Type I and Type II errors . . . 87

7.8.2 Individual vs. collective data snooping . . . 88

7.9 Predictability of other properties of asset returns . . . 89

7.10 Random walks versus EMH . . . 90

7.11 Conclusion . . . 90

7.12 Overview of chapter . . . 91

7.13 A reminder of your learning outcomes . . . 91

7.14 Test your knowledge and understanding . . . 91

7.15 Solutions to activities . . . 92

8 Modelling asset return volatility: Introduction 95

8.1 Introduction . . . 95

8.1.1 Aims of the chapter . . . 95

8.1.2 Learning outcomes . . . 95

8.1.3 Essential reading . . . 95

8.1.4 Further reading . . . 96

8.1.5 References cited . . . 96

8.2 Implications of homoskedasticity . . . 96

8.3 Predictability of asset return volatilities . . . 97

8.4 Autoregressive conditional heteroscedasticity (ARCH) processes . . . 99

8.5 Stationarity, moments, and restrictions on parameters . . . 100

8.6 Maximum likelihood estimation of GARCH models . . . 101
8.7 Testing for volatility clustering . . . 103
8.8 Overview of chapter . . . 104
8.9 Reminder of learning outcomes . . . 104
8.10 Test your knowledge and understanding . . . 106
8.11 Solutions to activities . . . 106

9 Modelling asset return volatility: Extensions 109
9.1 Introduction . . . 109
9.1.1 Aims of the chapter . . . 109
9.1.2 Learning outcomes . . . 109
9.1.3 Essential reading . . . 109
9.1.4 Further reading . . . 110
9.1.5 References cited . . . 110
9.2 Extensions of the univariate ARCH model . . . 111
9.2.1 Models with a 'leverage effect' . . . 111
9.2.2 ARCH-in-mean model . . . 113
9.2.3 NARCH, PARCH, QARCH, STARCH... . . . 114
9.2.4 Does anything beat a GARCH(1,1)? . . . 115
9.3 Choosing a volatility model . . . 116
9.3.1 Comparing nested models . . . 116
9.3.2 Using information criteria . . . 116
9.3.3 Using statistical goodness-of-fit measures . . . 117
9.3.4 Using economic goodness-of-fit measures . . . 118
9.4 Overview of chapter . . . 120
9.5 Reminder of learning outcomes . . . 120
9.6 Test your knowledge and understanding . . . 121
9.7 Solutions to activities . . . 121

10 Multivariate volatility models 123
10.1 Introduction . . . 123
10.1.1 Aims of the chapter . . . 123
10.1.2 Learning outcomes . . . 123
10.1.3 Essential reading . . . 123
10.1.4 Further reading . . . 123
10.1.5 References cited . . . 124

10.2 Modelling covariance matrices 124

10.2.1 Recap: Means and variances of random vectors 124

10.2.2 Empirical example 125

10.2.3 The two main problems in multivariate volatility modelling 127

10.3 Two popular multivariate volatility models 128

10.3.1 The constant conditional correlation (CCC) model 128

10.3.2 The RiskMetrics exponential smoother 129

10.4 Economic evaluation of multivariate volatility models 129

10.5 Overview of chapter 133

10.6 Reminder of learning outcomes 133

10.7 Test your knowledge and understanding 133

10.8 Solutions to activities 136

11 Optimal forecasts and forecast evaluation 137

11.1 Introduction 137

11.1.1 Aims of the chapter 137

11.1.2 Learning outcomes 137

11.1.3 Essential reading 137

11.1.4 References cited 138

11.2 Optimal forecasts 138

11.2.1 Definition of an optimal forecast 138

11.2.2 Deriving some optimal forecasts 139

11.3 Testing forecast optimality 140

11.3.1 Testable implications of forecast optimality 140

11.3.2 Regression-based tests of forecast optimality 141

11.3.3 Forecasting unobservable variables 142

11.4 Numerical example 143

11.5 Overview of chapter 145

11.6 Reminder of learning outcomes 147

11.7 Test your knowledge and understanding 147

11.8 Solutions to activities 147

12 Forecast comparison and combination 149

12.1 Introduction 149

12.1.1 Aims of the chapter 149

12.1.2 Learning outcomes 149

12.1.3 Essential reading 149

12.1.4 References cited 150

12.2 Comparing forecasts 150

12.2.1 Estimating the variance of $\hat{E}\,[d_t]$: 'HAC' standard errors 151

12.2.2 Numerical example, continued 153

12.3 Forecast encompassing and combining 153

12.3.1 Numerical example, continued 154

12.3.2 Forecast combination 155

12.3.3 Numerical example, concluded 155

12.4 Overview of chapter 156

12.5 Reminder of learning outcomes 156

12.6 Test your knowledge and understanding 157

12.7 Solutions to activities 157

13 Risk management and Value-at-Risk: Models 159

13.1 Introduction 159

13.1.1 Aims of the chapter 159

13.1.2 Learning outcomes 159

13.1.3 Essential reading 159

13.1.4 Further reading 160

13.1.5 References cited 160

13.2 An alternative measure of risk: VaR 160

13.3 Historical simulation 162

13.3.1 Empirical example using IBM returns 163

13.4 Weighted historical simulation 165

13.5 Models based on the normal distribution 166

13.5.1 Empirical example, continued 168

13.6 Models based on flexible distributions 168

13.7 Semiparametric models 170

13.8 Overview of chapter 172

13.9 Reminder of learning outcomes 172

13.10 Test your knowledge and understanding 172

13.11 Solutions to activities 173

14 Risk management and Value-at-Risk: Backtesting 175

14.1 Introduction 175

14.1.1 Aims of the chapter 175
14.1.2 Learning outcomes 175
14.1.3 Essential reading 175
14.1.4 References cited 176
14.2 Evaluating VaR forecasts 176
14.2.1 Unconditional coverage tests 176
14.2.2 Conditional coverage tests 177
14.2.3 Empirical example, continued 179
14.3 Comparing VaR forecasts 179
14.3.1 Diebold-Mariano tests for VaR forecasts 180
14.3.2 Empirical example, continued 181
14.4 Overview of chapter 181
14.5 Reminder of learning outcomes 182
14.6 Test your knowledge and understanding 182
14.7 Solutions to activities 182

15 Modelling high frequency financial data: Diurnality 185

15.1 Introduction 185
15.1.1 Aims of the chapter 185
15.1.2 Learning outcomes 185
15.1.3 Essential reading 186
15.1.4 Further reading 186
15.2 Diurnality in asset returns 186
15.3 Dummy variable models 186
15.4 Polynomial trend models 190
15.5 Modelling conditional volatility in the presence of diurnality 192
15.6 Overview of chapter 198
15.7 Reminder of learning outcomes 198
15.8 Test your knowledge and understanding 198
15.9 Solutions to activities 199

16 Modelling high frequency financial data: Irregularly-spaced data 201

16.1 Introduction 201
16.1.1 Aims of the chapter 201
16.1.2 Learning outcomes 201
16.1.3 Essential reading 201

16.1.4 Further reading 202
16.1.5 References cited 202
16.2 The market microstructure of IBM stock prices 202
16.3 Modelling durations 204
16.3.1 Diurnality in durations 204
16.3.2 Autoregressive Conditional Duration (ACD) 209
16.3.3 Estimating the ACD model 211
16.3.4 Moments and parameter restrictions for the ACD model 212
16.3.5 Application to IBM durations 212
16.3.6 Evaluating the ACD model 215
16.4 Overview of chapter 215
16.5 Reminder of learning outcomes 215
16.6 Test your knowledge and understanding 216
16.7 Solutions to activities 216

17 Modelling high frequency financial data: Discreteness 219

17.1 Introduction 219
17.1.1 Aims of the chapter 219
17.1.2 Learning outcomes 219
17.1.3 Essential reading 219
17.1.4 Further reading 219
17.1.5 References cited 220
17.2 Modelling discrete high frequency prices 220
17.3 Discreteness of high frequency IBM prices 220
17.4 The 'ADS' model for discrete prices 222
17.5 Overview of chapter 227
17.6 Reminder of learning outcomes 227
17.7 Test your knowledge and understanding 227
17.8 Solutions to activities 228

18 Spurious regressions and persistent time series 233

18.1 Introduction 233
18.1.1 Aims of the chapter 233
18.1.2 Learning outcomes 233
18.1.3 Essential reading 233
18.1.4 Further reading 234

18.1.5 References cited . . . 234
18.2 Random walks and time trends . . . 234
18.3 Spurious regressions . . . 236
18.4 Testing for a unit root . . . 241
18.4.1 The Dickey-Fuller test . . . 243
18.4.2 Testing for higher-order integration . . . 245
18.5 Overview of chapter . . . 245
18.6 Reminder of learning outcomes . . . 245
18.7 Test your knowledge and understanding . . . 246
18.8 Solutions to activities . . . 248

A Sample examination paper **251**

B Guidance on answering the Sample examination questions **253**

Chapter 1
Introduction

1.1 Route map to the guide

This subject guide is designed to help you understand, and eventually master, the material to be covered in the final examination of **FN3142 Quantitative finance**. This material is generally technical in nature, and the best way to learn it is to work through all of the activities and derivations in this subject guide and the accompanying readings. This is not a passive course! Merely reading this subject guide is not enough – you need to be willing to devote time to solving the numerous practice questions and problems presented here. Be sure to check the VLE for additional practice questions and discussion. Solutions for the activities are presented at the end of each chapter, to help you learn some 'tricks' for answering these questions. The 'test your knowledge' questions at the end of each chapter have no solutions – you need to try to solve those questions for yourself, and then convince yourself that you have done it correctly (and then perhaps compare your answers with a friend or classmate).

1.2 Why study quantitative finance?

Modern financial analysis is quantitative by necessity. The mass of information, both financial and non-financial, available to investors, traders, risk managers, and regulators is best handled with a structured, quantitative approach. Indeed, it is hard to imagine a way of sifting through this information without using some sort of quantitative tool. This course will provide students with a framework for thinking about financial data, and with some quantitative tools for conducting analyses of financial data.

This subject guide provides an introduction to some of the most useful techniques in modern quantitative finance. Quantitative finance builds on both financial economics (covered in **FN3092 Corporate finance** and **FN2024 Principles of banking and finance**) and econometrics (covered in **EC2020 Elements of econometrics**). As such, we will draw on financial methods, such as utility maximisation and standard portfolio decision rules, and also on econometric methods, such as regressions and hypothesis testing. The methods and models we cover in this course have numerous possible applications – portfolio decisions, risk management, derivatives pricing, and macroeconomic policy, amongst many others. Students who complete this course will be well-prepared to take on quantitative jobs in the finance industry, or to continue their studies in finance at the postgraduate level.

1.3 Syllabus

Building on concepts introduced in **FN3092 Corporate finance** and **EC2020 Elements of econometrics**, this course introduces econometric tools related to time-series analysis and applies them to study issues in asset pricing, investment theory, risk analysis and management, market microstructure, and return forecasting.

Topics addressed by this course are:

- Concepts and measures of risk
- Time-series analysis
- Empirical features of financial asset returns
- Market risk models
- Models of financial market correlations
- Forecast evaluation methods
- Risk management
- Asset allocation decisions
- Market microstructure and high frequency data

1.4 Aims of the course

This course provides the econometric techniques, such as time-series analysis, required to analyse theoretical and empirical issues in finance. It provides applications in asset pricing, investments, risk analysis and management, market microstructure, and return forecasting.

1.5 Learning outcomes for the course

At the end of this course, and having completed the essential reading and activities, you should:

- have mastered the econometric techniques required in order to analyse issues in asset pricing and market finance
- be familiar with recent empirical findings based on financial econometric models
- have gained valuable insights into the functioning of financial markets
- understand some of the practical issues in the forecasting of key financial market variables, such as asset prices, risk and dependence.

1.6 Overview of learning resources

1.6.1 The subject guide

This subject guide is designed to complement, not replace, the listed readings for each chapter. Most of the essential and further reading materials come from the following text books. Each chapter of this guide builds on the earlier chapters, as is often the case with quantitative courses, and so I suggest that the chapters be studied in the order in which they are presented here.

1.6.2 Essential reading

Essential reading for this course comes from:

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487].
- Diebold, F.X. *Elements of Forecasting.* (Thomson South-Western, Canada, 2006) fourth edition [ISBN 9780324323597].

The book by Christoffersen is the closest to this guide, though it is missing some econometric methods that we need. The book by Diebold fills the econometrics requirement, even though his focus is more on macroeconomic applications than financial applications.

1.6.3 Further reading

In addition to the above books, further reading material is available in the following books:

- Campbell, J.Y., A.W. Lo and A.C. Mackinlay *The Econometrics of Financial Markets.* (Princeton University Press, Princeton, New Jersey, 1997) [ISBN 9780691043012].
- Taylor, Stephen J. *Asset Price Dynamics, Volatility and Prediction.* (Princeton University Press, Oxford, 2005) [ISBN 9780691134796].
- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354].

The book by Tsay is the closest to this guide, though it is pitched at the Masters rather than undergraduate level. He covers some of the material in more depth than is required for this course. If you are interested in postgraduate study in finance or econometrics, you may find the readings from Tsay helpful. Taylor's book is also aimed at Masters students, but covers several of the topics we cover in this guide. Campbell, Lo and Mackinlay is a classic, graduate-level, book covering topics in finance and financial econometrics.

For additional reading on finance and investments topics that arise in this subject guide see the following books:

- Bodie, Z., A. Kane and A.J. Marcus *Investments.* (McGraw-Hill, U.S.A., 2013) ninth edition [ISBN 9780077861674].
- Elton, E.J., M.J. Gruber, S.J. Brown and W.N. Goetzmann *Modern Portfolio Theory and Investment Analysis.* (John Wiley & Sons, New York, 2009) eighth edition [ISBN 978118038093].

For additional reading/revision of regression and hypothesis testing topics see the following books, both of which are aimed at undergraduate students:

- Stock, J.H. and M.W. Watson *Introduction to Econometrics.* (Pearson Education, Boston, 2010) third edition. [ISBN 9781408264331].
- Wooldridge, J.M. *Introductory Econometrics: A Modern Approach.* (South-Western, USA, 2012) fifth edition. [ISBN 9781111531041].

1.6.4 Online study resources

In addition to the subject guide and the Essential reading, it is crucial that you take advantage of the study resources that are available online for this course, including the VLE and the Online Library. You can access the VLE, the Online Library and your University of London email account via the Student Portal at:
`http://my.londoninternational.ac.uk`

You should have received your login details for the Student Portal with your official offer, which was emailed to the address that you gave on your application form. You have probably already logged in to the Student Portal in order to register. As soon as you registered, you will automatically have been granted access to the VLE, Online Library and your fully functional University of London email account.

If you have forgotten these login details, please click on the 'Forgotten your password' link on the login page.

The VLE

The VLE, which complements this subject guide, has been designed to enhance your learning experience, providing additional support and a sense of community. It forms an important part of your study experience with the University of London and you should access it regularly.

The VLE provides a range of resources for EMFSS courses:

- Self-testing activities: Doing these allows you to test your own understanding of subject material.
- Electronic study materials: The printed materials that you receive from the University of London are available to download, including updated reading lists and references.

- Past examination papers and Examiners' commentaries: These provide advice on how each examination question might best be answered.
- A student discussion forum: This is an open space for you to discuss interests and experiences, seek support from your peers, work collaboratively to solve problems and discuss subject material.
- Videos: There are recorded academic introductions to the subject, interviews and debates and, for some courses, audio-visual tutorials and conclusions.
- Recorded lectures: For some courses, where appropriate, the sessions from previous years' Study Weekends have been recorded and made available.
- Study skills: Expert advice on preparing for examinations and developing your digital literacy skills.
- Feedback forms.

Some of these resources are available for certain courses only, but we are expanding our provision all the time and you should check the VLE regularly for updates.

Making use of the Online Library

The Online Library contains a huge array of journal articles and other resources to help you read widely and extensively.

To access the majority of resources via the Online Library you will either need to use your University of London Student Portal login details, or you will be required to register and use an Athens login: `http://tinyurl.com/ollathens`

The easiest way to locate relevant content and journal articles in the Online Library is to use the Summon search engine. If you are having trouble finding an article listed in a reading list, try removing any punctuation from the title, such as single quotation marks, question marks and colons.

For further advice, please see the online help pages:
`www.external.shl.lon.ac.uk/summon/about.php`

1.7 The structure of the subject guide

The following is a brief outline of this subject guide.

- **Chapter 1:** Introduction
  - Aims and objectives for the course
  - Recommended reading
- **Chapter 2**: Financial econometrics concepts and statistics review
  - Key definitions in financial econometrics
  - Review of moments, distributions and densities

  - Vector random variables

- **Chapter 3**: Basic time series concepts
  - Autocovariances and autocorrelations
  - The Law of Iterated Expectations
  - White noise processes
- **Chapter 4**: ARMA processes
  - Autoregressive (AR) processes
  - Moving average (MA) processes
  - Choosing an ARMA model
- **Chapter 5**: Empirical features of financial asset returns
  - Common summary statistics
  - Tests of Normality of asset returns
- **Chapter 6**: Testing for predictability in financial time series
  - Sample autocorrelations
  - Individual and joint tests on sample autocorrelations
- **Chapter 7**: The efficient markets hypothesis and market predictability
  - Review of the basic definition of the efficient markets hypothesis
  - Extensions and refinements of the EMH
- **Chapter 8**: Modelling asset return volatility – Introduction
  - Evidence of volatility clustering in financial asset returns
  - Models of time-varying conditional variance: ARCH and GARCH
  - Testing for volatility clustering
- **Chapter 9**: Modelling asset return volatility – Extensions
  - Extensions of the ARCH model: asymmetric GARCH and ARCH-in-mean
  - Methods for choosing a volatility model
- **Chapter 10**: Multivariate volatility models
  - The two main problems in multivariate volatility modelling
  - Multivariate GARCH models: CCC and RiskMetrics
  - Portfolio decision-making with multivariate volatility models
- **Chapter 11**: Optimal forecasts and forecast evaluation
  - Defining an ‘optimal’ forecast
  - Mincer-Zarnowitz regressions for forecast evaluation

- **Chapter 12:** Forecast comparison and combination
  - Diebold-Mariano tests for forecast comparison
  - Forecast encompassing and forecast combination regressions
- **Chapter 13:** Risk management and Value-at-Risk: Models
  - Introduction to Value-at-Risk (VaR)
  - Common models for measuring and predicting VaR
- **Chapter 14**: Risk management and Value-at-Risk: Backtesting
  - Unconditional coverage tests
  - Conditional coverage tests
  - Comparing VaR forecasts
- **Chapter 15**: Modelling high frequency financial data: Diurnality
  - Introduction to high frequency financial data
  - Diurnality in volatility and liquidity
- **Chapter 16**: Modelling high frequency financial data: Irregularly-spaced time series
  - Predictability in the ‘durations’ of trades
  - The ACD model for durations
- **Chapter 17**: Modelling high frequency financial data: Discreteness
  - Price discreteness at high frequencies
  - Methods for modelling discrete time series
  - The ADS model for discrete prices
- **Chapter 18**: Spurious regressions and persistent time series
  - Random walks and time trends
  - Spurious regressions and the breakdown of standard econometric theory
  - Testing for a unit root: the Dickey-Fuller test

1.8 Examination advice

Important: the information and advice given in this section are based on the examination structure at the time that this guide was written. Please note that subject guides may be used for several years, and thus we strongly advise you to check both the current *Regulations* for relevant information about the examination, and the current *Examiners' commentaries* where you should be advised of any forthcoming changes.

The **FN3142 Quantitative finance** examination paper is three hours in duration. You will be asked to answer **three** questions out of **four**, giving you an hour to answer each question. You are strongly advised to divide your time in this manner. The examination for this course contains a mix of quantitative and qualitative questions. Examples of examination questions are provided at the end of each chapter, and a complete sample examination paper is provided at the end of this guide.

Chapter 2
Financial econometrics concepts and statistics review

2.1 Introduction

This chapter introduces some key concepts and definitions from financial econometrics that will be used throughout this subject guide: time series, sampling frequencies, return definitions. We then review some fundamental definitions and results from statistics: definitions and calculation of moments (means, variances, skewness, etc.) and distribution and density functions. We also review definitions of moments for vector random variables.

2.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce some terminology and concepts from financial econometrics for studying financial data
- Show the equivalence of forecasting prices and forecasting returns
- Review results and definitions for moments, distributions and densities, for both scalar and vector random variables.

2.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be:

- Able to compute arithmetic (simple) and logarithmic (continuously-compounded) returns
- Familiar with standard concepts from statistics: moments, distributions and densities
- Able to compute moments numerically for discrete random variables
- Able to derive means and variances for vector random variables.

2.1.3 Essential reading

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487], Chapter 3 Sections 1–3.

2.1.4 Further reading

- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354], Chapter 1.

2.1.5 References cited

- Hamilton, J. D. *Time Series Analysis.* (Princeton University Press, New Jersey, 1994) [ISBN 9780691042893].
- Student, 'Errors of routine analysis,' *Biometrika*, 1927, 19(1-2), pp.151–164.

2.2 What is financial econometrics?

The field of *econometrics* was first defined (back in 1933) as the application of mathematics and statistical methods to the analysis of economic data. *Financial econometrics* is a branch of econometrics focusing on the analysis of financial data and problems from financial economics. Financial econometrics has much in common with other branches of econometrics. Like all branches of econometrics, financial econometrics is concerned with estimation (e.g., least squares, maximum likelihood, methods of moments) and hypothesis testing. Like *macro*econometrics, much of the data used in financial econometrics comes in the form of *time series*, and so we need tools that can handle such data. Also like macroeconometrics, many of the problems in financial econometrics are problems of *prediction*, and so we spend time studying the construction and evaluation of forecasts. Like *micro*econometrics, many problems in financial econometrics involve vast amounts of data, which requires some care for analysis.

Financial econometrics has one key feature that distinguishes it from other branches of econometrics: the analysis of *risk* is central to financial economics, and the econometric analysis of risk is one of the defining features of financial econometrics. The fact that most measures of risk are *unobservable* requires new econometric methods for their study. Another novel aspect of financial econometrics follows from the fact that financial decisions are made across a wide range of time spans, from many years (planning for retirement, purchases of insurance) to seconds (high-frequency trading strategies, reacting to news announcements). This requires econometric methods that can be applied or adapted across a range of *sampling frequencies*.

2.3 Some important concepts

Much of the data used in finance are time series, which are sequences of observations on the same variable at different points in time. Some **standard financial time series** include:

- stock prices or stock returns
- exchange rates
- interest rates
- bond prices or bond returns
- inflation rates
- options prices or futures prices

The fact that a time series is comprised of data on the same variable (or set of variables) through time means that it may be possible to use the value of the variable today to predict something about the variable at some point in the future. For example, knowing that the FTSE 100 index was at 6778.56 today might tell us something about what it will be tomorrow.

A central question in finance relates to the risk/return trade off, and so forecasting just the price, or just the change in the price of an asset (i.e., the *return* on the asset) is only one half of the problem. As risk-averse investors, we will also care about forecasting *risk*, measured in some way, and so we may also be interested in forecasting **other properties of a time series**, such as:

- the volatility of the asset
- the probability of a crash in the market
- the correlation or dependence between asset returns (between individual stocks, between international stock markets, between exchange rates, etc.)
- the liquidity of the market for the asset (measured by trading volume, trade intensity, bid-ask spread, etc.)
- the entire conditional distribution of asset returns

There are many **applications of forecasts in financial markets**, including:

- risk management
- portfolio management
- option pricing
- government/monetary policy
- (statistical) arbitrage trading

As we will see, many of these forecasting problems are more general (and more complicated) than simply trying to predict the price of some financial asset.

Financial data is available at a **range of frequencies** (how often we see a data point):

- annually
- monthly
- weekly
- daily
- 30-minute, 10-minute, 5-minute
- random (e.g., high frequency 'tick' data)

Most analysis in finance is done using daily or monthly data, but increasingly attention is being paid to intra-daily data. We will consider a variety of data frequencies in this course.

Skills required to develop and evaluate forecasts in financial markets:

- **Statistics and econometrics**: need to know about random variables, probabilities, regression theory, amongst other things.
- **Economics**: need to be able to tell whether a model makes economic sense or not.
  - Most disciplines that use forecasts (or statistics in general) have specialist forecasters within their field: biology has biostatisticians, medicine has epidemiologists, atmospheric science has meteorologists, and economics has econometricians. Why aren't there just expert forecasters out there ready to work with any data set? Because knowing where the data comes from and how it is generated generally leads to better forecasts.
- **Common sense**: do you believe the forecast a particular model tells you? Should you?
  - For example, using an 'insanity filter' on a series of forecasts from a model.

2.4 Forecasting returns and prices

Returns can be constructed from prices in two ways. 'Simple' or 'arithmetic' returns are defined as:

$$\begin{aligned} R_{t+1} &\equiv \frac{P_{t+1}-P_t}{P_t} \equiv \frac{\Delta P_{t+1}}{P_t} = \frac{P_{t+1}}{P_t} - 1 \\ \text{so } P_{t+1} &= P_t\left(1+R_{t+1}\right) \end{aligned}$$

Here R_{t+1} is called the 'net return' (it will be a number like 0.03, -0.01, etc.) and $(1+R_{t+1})$ is called the 'gross return' (which will be something like 1.03, 0.99, etc.)

'Continuously compounded' or 'logarithmic' or 'log' returns are defined as[1]

$$\begin{aligned} R_{t+1} &\equiv \log\left(\frac{P_{t+1}}{P_t}\right) = \log P_{t+1} - \log P_t \equiv \Delta \log P_{t+1} \\ \text{so } P_{t+1} &= P_t \exp\left\{R_{t+1}\right\} \end{aligned}$$

The continuously compounded return is sometimes referred to as the 'log-difference' of the price series. Here R_{t+1} is again the 'net return' and $\exp\{R_{t+1}\}$ is the 'gross return'. We can convert arithmetic returns to log returns, and vice versa, using the fact that

$$\begin{aligned} P_t\left(1 + R^A_{t+1}\right) &= P_{t+1} = P_t \exp\left\{R^L_{t+1}\right\} \\ \text{so } R^L_{t+1} &= \log\left(1 + R^A_{t+1}\right) \\ R^A_{t+1} &= \exp\left\{R^L_{t+1}\right\} - 1 \end{aligned}$$

Both definitions give approximately the same answer when returns are not 'too large' (less than around 0.10, or 10%, in absolute value).

Activity 2.1 Compute the arithmetic returns and the continuously compounded returns for the following cases:

P_t	P_{t+1}	R^A_{t+1}	R^L_{t+1}
100	103		
100	92		
100	145		
100	30		

Throughout this course we will focus on continuously compounded returns. One reason for doing so is that it allows for simple time series aggregation of returns. For example, let $Y_{t+5} = \log P_{t+5} - \log P_t$ be the weekly return on an asset, and let $X_{t+1} = \log P_{t+1} - \log P_t$ be the daily return on the same asset. Then notice that:

$$\begin{aligned} Y_{t+5} &= \log P_{t+5} - \log P_t \\ &= \log P_{t+5} - \log P_{t+4} \\ &\quad + \log P_{t+4} - \log P_{t+3} \\ &\quad + \log P_{t+3} - \log P_{t+2} \\ &\quad + \log P_{t+2} - \log P_{t+1} \\ &\quad + \log P_{t+1} - \log P_t \\ &= X_{t+5} + X_{t+4} + X_{t+3} + X_{t+2} + X_{t+1} \end{aligned}$$

[1] All logs in these notes are *natural* logarithms unless otherwise noted.

and so the weekly continuously compounded return is simply the sum of the daily continuously compounded returns through the week. It might be noted that while continuously compounded returns allow for simple aggregation of returns through time, they do not allow for simple aggregation of returns across stocks, to get a portfolio return for example. However for reasonable values of returns the difference is small.

> **Activity 2.2** Let $Z_{t+2} = \left(P_{t+2} - P_t\right)/P_t$ be the two-day arithmetic return, and let $W_{t+1} = \left(P_{t+1} - P_t\right)/P_t$ be the one-day arithmetic return. Find an expression for Z_{t+2} as a function of W_{t+1} and W_{t+2} (and notice that it is not as nice as the expression for continuously compounded returns).

Here we will show that forecasting prices is equivalent to forecasting returns. This is true so long as we include today's price as part of the information set (which we always do).

$$\begin{aligned}
\text{Price forecast} &: \quad E\left[P_{t+1}|\mathcal{F}_t\right] \equiv E_t\left[P_{t+1}\right] \equiv \hat{P}_{t+1} \\
\text{Return forecast} &: \quad E\left[R_{t+1}|\mathcal{F}_t\right] \equiv E_t\left[R_{t+1}\right] \equiv \hat{R}_{t+1} \\
P_{t+1} &= P_t\left(1 + R_{t+1}\right) \\
E_t\left[P_{t+1}\right] &= E_t\left[P_t\left(1 + R_{t+1}\right)\right] \\
&= P_t \cdot \left(1 + E_t\left[R_{t+1}\right]\right) \\
\hat{P}_{t+1} &= P_t\left(1 + \hat{R}_{t+1}\right), \text{ or} \\
\hat{R}_{t+1} &= \frac{\hat{P}_{t+1}}{P_t} - 1
\end{aligned}$$

and so forecasting prices is equivalent to forecasting returns. If instead we use continuously compounded returns we get:

$$\begin{aligned}
P_{t+1} &= P_t \exp\left\{R_{t+1}\right\} \\
E_t\left[P_{t+1}\right] &= E_t\left[P_t \exp\left\{R_{t+1}\right\}\right] \\
&= P_t \cdot E_t\left[\exp\left\{R_{t+1}\right\}\right] \\
\hat{P}_{t+1} &= P_t \cdot \widehat{\exp\left\{R_{t+1}\right\}} \\
\widehat{\exp\left\{R_{t+1}\right\}} &= \frac{\hat{P}_{t+1}}{P_t}
\end{aligned}$$

Note that by Jensen's inequality[2]: $\widehat{\exp\left\{R_{t+1}\right\}} \neq \exp\left\{\hat{R}_{t+1}\right\}$, and so when using continuously compounded returns we should forecast the gross return, $\exp\left\{R_{t+1}\right\}$, and not the net return, R_{t+1}. For arithmetic returns we can forecast either gross returns or net returns, R_{t+1}. In practice, most people just forecast the net return and assume that $\widehat{\exp\left\{R_{t+1}\right\}} \approx \exp\left\{\hat{R}_{t+1}\right\}$, which is generally a pretty good approximation.

The reason for the emphasis on the equivalence between prices and returns is that while prices are often the economic object of interest, they have statistical properties that

[2] Jensen's inequality states that if g is a convex function (like the exponential) then $g\left(E\left[X\right]\right) \leq E\left[g\left(X\right)\right]$.

make them hard to deal with. Prices (usually) have a ‘unit root’, meaning, amongst other things, that the variance of prices diverges to infinity as time goes on. Dealing with variables that have a unit root requires more care than required for variables with no unit root. (We will look at this problem in a Chapter 18 on spurious regressions.) Returns, generally, do *not* have a unit root, which makes their analysis a lot easier econometrically.

For the remainder of the course we will discuss forecasting prices and returns interchangeably.

2.5 Revision of basic statistics

2.5.1 Random variables, CDFs, PMFs, PDFs

Definition 2.1 (Random variable) A random variable, X, is a function with domain S (the ‘sample space’, which is the set of all possible outcomes for the random variable) and range $\mathbb{R}$, the real line. If the random variable can only take on a finite number of values, then it is called a ‘discrete random variable’. In general, if it can take on any value in some interval of the real line then it is called a ‘continuous random variable’ (though there are exceptions, a formal definition is below). A random variable is usually denoted with an upper-case letter, while a realisation from a random variable is denoted with a lower-case letter.

> **Example 2.1** Coin-tossing: Let $X = 1$ if the coin comes up heads, and let $X = 0$ if the coin comes up tails. Then X is a discrete random variable, and $\{0, 1\}$ is the set of possible realisations of the random variable.

> **Example 2.2** Rainfall: Let X be the rainfall in London in milliliters over the past year. Then X is a continuous random variable with support on the non-negative real line. (It’s not the entire real line because we can’t see negative rainfall.) Example realisations of X in this example are 200, 34.3535, 0, etc.

Definition 2.2 (Cumulative distribution function, or *cdf*) The cumulative distribution function (or *cdf*) is the function describing the probability of the random variable taking a value less than or equal to a particular number. It completely describes a random variable, and it is the same for both discrete and continuous random variables. A cdf is usually denoted in upper case letters:

$$F(x) \equiv \Pr[X \leq x]$$

If X has *cdf* F, we write that ‘X is distributed according to F’, or ‘$X \sim F$’ in shorthand. A cdf has the following properties:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$
2. $F(x)$ is a non-decreasing function of x
3. $F(x)$ is right-continuous, i.e., $\lim_{\delta\to0^+} F(x+\delta) = F(x)$ for all x

If we do not know the complete distribution of a random variable, but we do know its first 2 moments (i.e., its mean and variance) then we write that $X \sim (\mu, \sigma^2)$. For some distributions, such as the normal, knowing the mean and variance is sufficient to completely describe the random variable. E.g.: if we know X is normally distributed with mean 2 and variance 5 we write $X \sim N(2, 5)$ .Other distributions are characterised by other properties: for example, if we know X is uniformly distributed between -3 and 10 we write $X \sim Unif\left(-3, 10\right)$.

Definition 2.3 (Probability mass function, or *pmf*) The probability mass function, f, of a **discrete** random variable X is given by:

$$f(x) \equiv \Pr[X = x]$$

A pmf satisfies the following properties:

1. $f(x) \geq 0$ for all x
2. $\sum_x f(x) = 1$

The points at which a discrete random variable has a positive *pmf* are known as the 'support' of this random variable.

A continuous random variable is formally defined by $F(x)$ being a continuous function of x. For continuous random variables $\Pr[X = x] = 0$ for all x, by the continuity of the cdf, and so instead of using *pmf*'s we use probability density functions:

Definition 2.4 (Probability density function, or *pdf*) The probability density function, f, of a **continuous** random variable X is the function that satisfies:

$$F(x) = \int_{-\infty}^{x} f(s)\, ds \text{ for all x}$$

If F is differentiable then the pdf may be obtained as

$$f(x) = \frac{\partial F(x)}{\partial x}$$

A pdf satisfies the following properties:

1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$

Figure 2.1 shows an illustration of CDFs for a discrete and a continuous random variable, and their corresponding PMF and PDF.

> **Activity 2.3** Unlike a *pmf*, a *pdf* can take values greater than one. (As the *pmf* is a probability, it will always lie between zero and one.) To prove this, consider a random variable uniformly distributed on the interval $[a, b]$, where $a < b$. A 'Unif(a, b)' random variable has the *cdf* $F(x) = (x - a)/(b - a)$, for $a < x < b$. Find the *pdf* of this random variable, and then find values of a and b such that the *pdf* takes values greater than one.

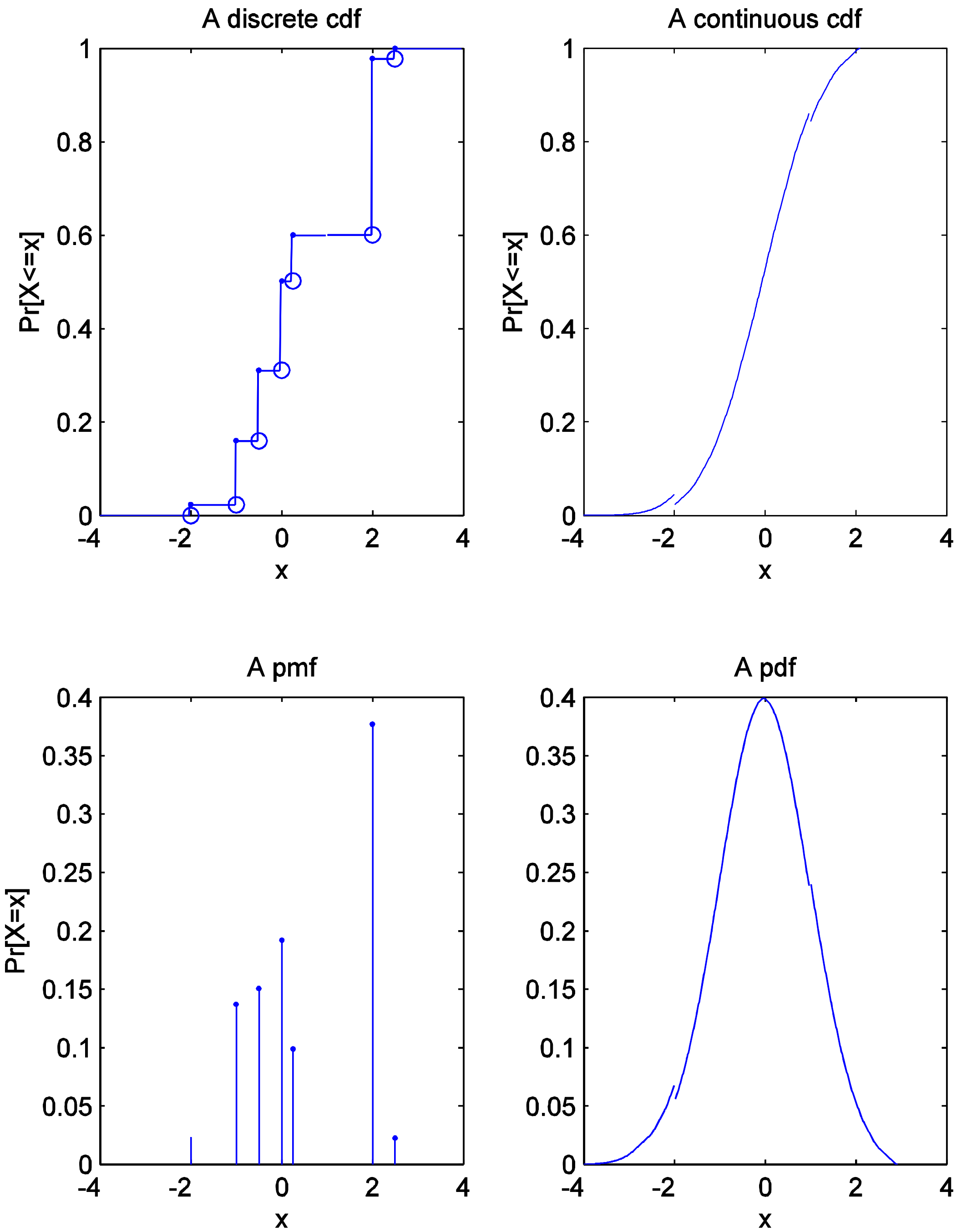


Figure 2.1: CDFs, PDF and PMF for a continuous and discrete random variable.

Definition 2.5 (Time series) A time series is an ordered set of realisations from some random variable. Usually the set is ordered according to time (thus the name 'time series').

The field of time series analysis is broad and complex. Some of the definitions given here are adequate for the purposes of this course, but may not be sufficient for higher-level study. The standard graduate text on time series analysis for econometricians is Hamilton (1994), and the interested student should look there for a more rigorous treatment.

Examples of time series can be found everywhere: daily temperatures in Paris, closing price of Google shares, sales of prawn sandwiches at Wright's Bar each Tuesday, etc. In Figure 2.2 I plot a time series of daily EUR/US dollar exchange rates and exchange rate returns over the period 1999-2009, and in Figure 2.3 I plot a time series of one-second prices on IBM on 31 December 2009.

2.5.2 Means, variances, and other moments

Definition 2.6 (Mean) The mean (or expected value) of the random variable X with *pdf* f is:

$$\mu \equiv E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Definition 2.7 (Variance) The variance of the random variable X with *pdf* f is:

$$\begin{aligned}\sigma^2 &\equiv V[X] \equiv E\left[(X-\mu)^2\right] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\, dx \\ &= E\left[X^2\right] - \mu^2\end{aligned}$$

The 'standard deviation' of a random variable is the square root of the variance:

$$\sigma = \sqrt{E[X^2] - \mu^2}$$

Definition 2.8 (Skewness) The skewness of the random variable X with *pdf* f is:

$$\begin{aligned}s &\equiv Skew[X] \equiv \frac{E\left[(X-\mu)^3\right]}{\sigma^3} \\ &= \frac{1}{\sigma^3}\int_{-\infty}^{\infty} (x-\mu)^3 \cdot f(x)\, dx\end{aligned}$$

Definition 2.9 (Kurtosis) The kurtosis of the random variable X with *pdf* f is:

$$\begin{aligned}\kappa &\equiv Kurt[X] \equiv \frac{E\left[(X-\mu)^4\right]}{\sigma^4} \\ &= \frac{1}{\sigma^4}\int_{-\infty}^{\infty} (x-\mu)^4 \cdot f(x)\, dx\end{aligned}$$

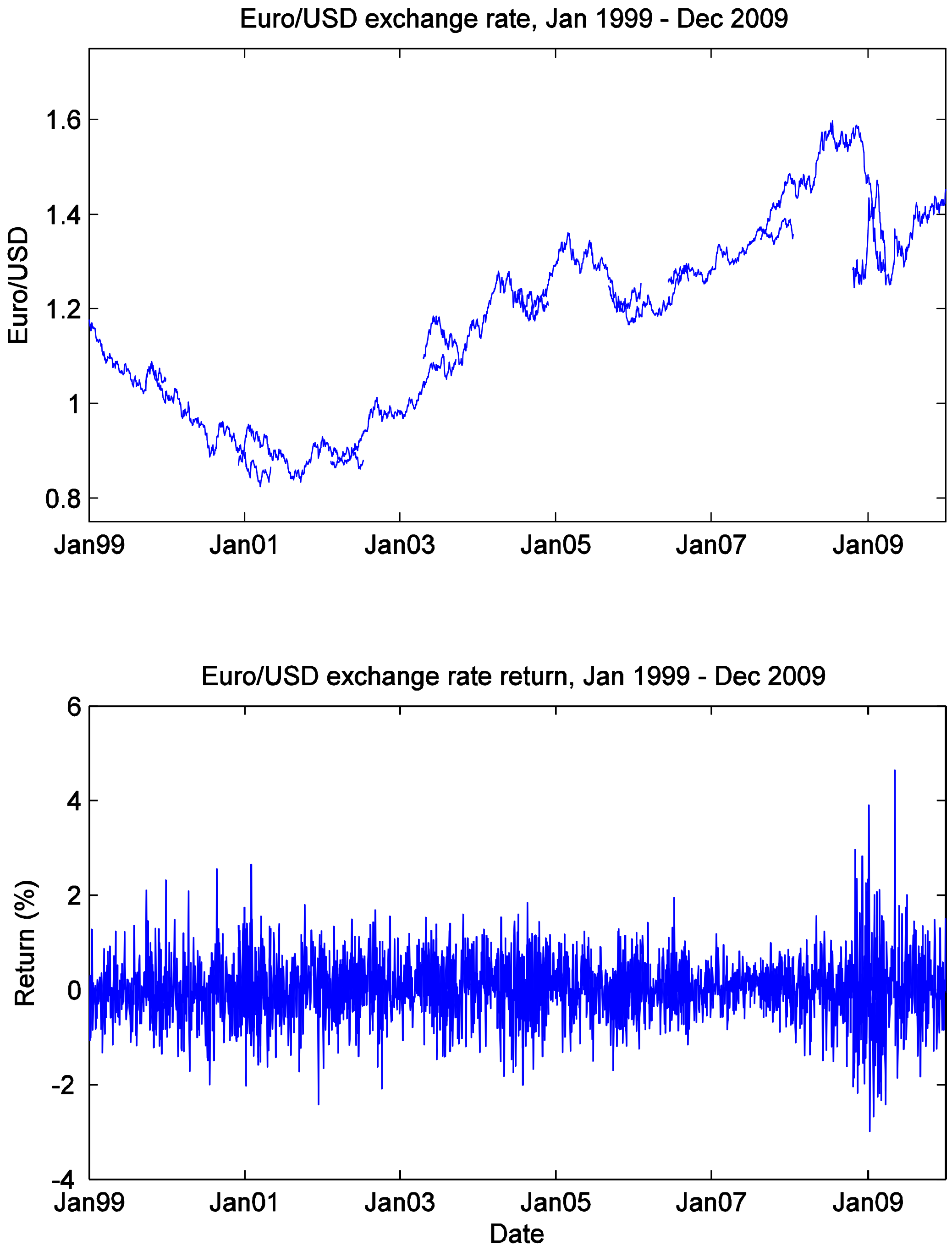


Figure 2.2: Euro/US dollar exchange rate, and daily exchange rate return, January 1999 to December 2009.

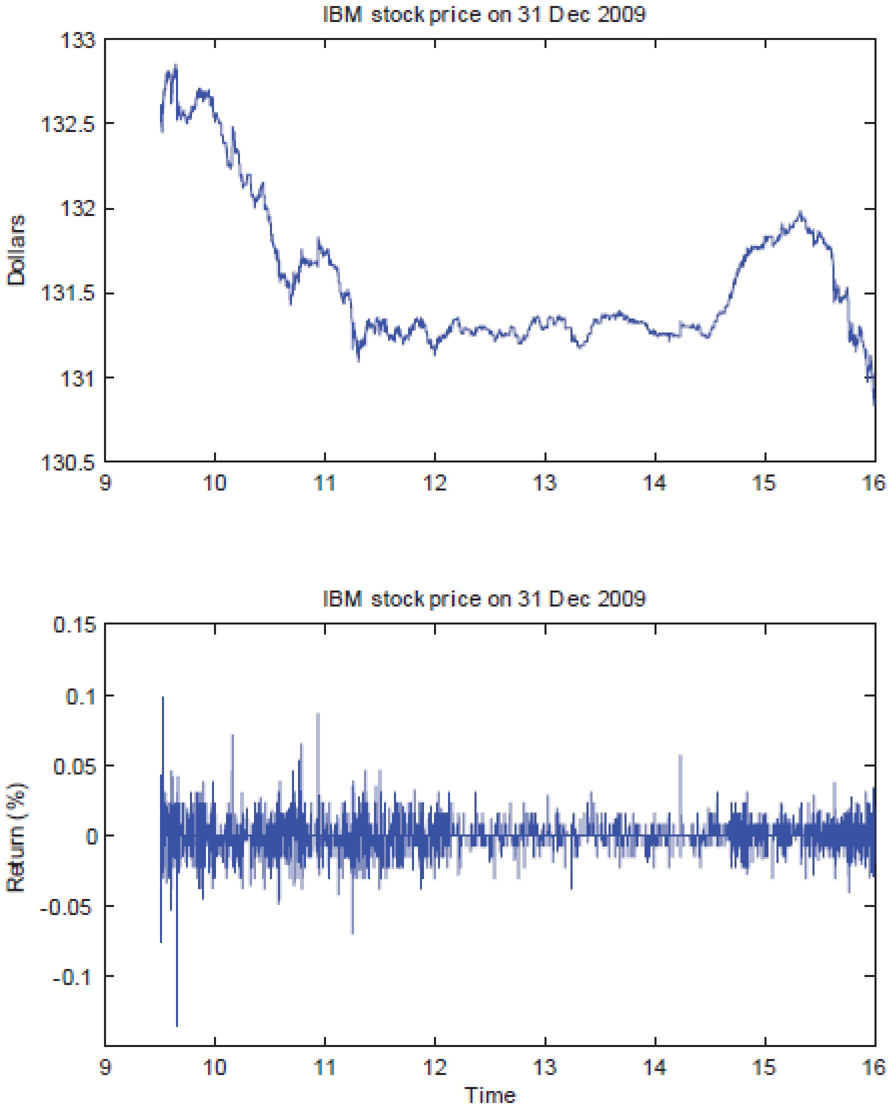
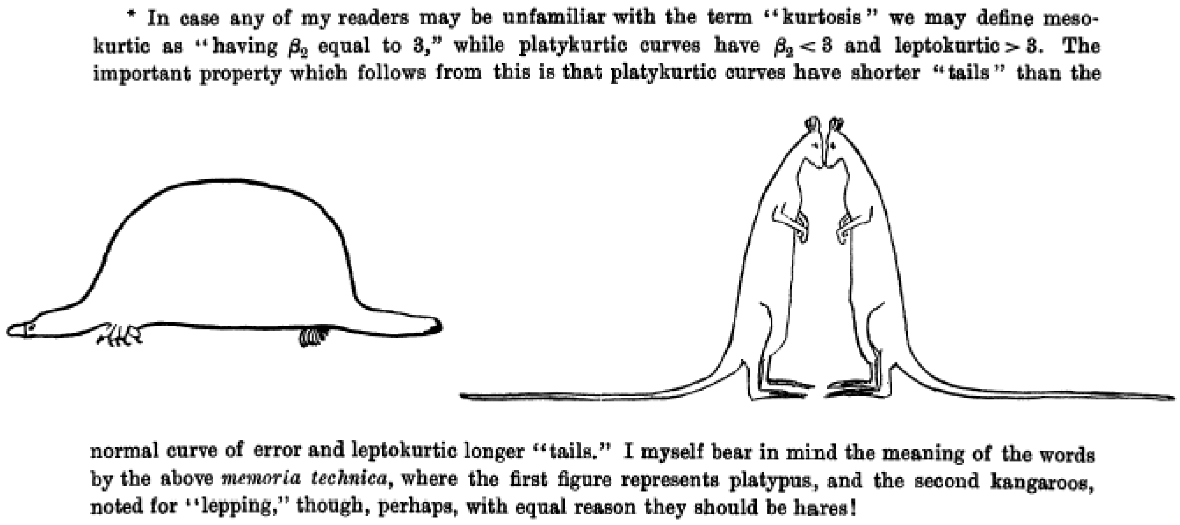


Figure 2.3: IBM stock price, and 1-second returns, on 31 December 2009.



* In case any of my readers may be unfamiliar with the term "kurtosis" we may define mesokurtic as "having β_2 equal to 3," while platykurtic curves have $\beta_2 < 3$ and leptokurtic > 3. The important property which follows from this is that platykurtic curves have shorter "tails" than the normal curve of error and leptokurtic longer "tails." I myself bear in mind the meaning of the words by the above *memoria technica*, where the first figure represents platypus, and the second kangaroos, noted for "lepping," though, perhaps, with equal reason they should be hares!

Figure 2.4: Student's (1927) memory aids for platykurtosis ($\kappa < 3$) and leptokurtosis ($\kappa > 3$).

Definition 2.10 (Moment) The pth 'moment' of the random variable X with *pdf* f is:

$$\tilde{m}_p \equiv E\left[X^p\right] = \int_{-\infty}^{\infty} x^p \cdot f(x)\, dx$$

Definition 2.11 (Central moment) The pth 'central moment' of the random variable X with *pdf* f is:

$$m_p \equiv E\left[(X-\mu)^p\right] = \int_{-\infty}^{\infty} (x-\mu)^p \cdot f(x)\, dx$$

> **Activity 2.4** A random variable is symmetric around zero if $f(x) = f(-x)$ for all x. Using the integral definition of skewness, show that all random variables that are symmetric around zero must have zero skewness.

The above definitions apply for continuous random variables, which is the most common case. For a discrete random variable with *pmf f,* the definition is modified slightly. For example, the pth moment is defined as

$$\tilde{m}_p \equiv E\left[X^p\right] = \sum_{i=1}^{n} x_i^p f(x_i)$$

where the sum is taken over all the possible values for X. These values are denoted $(x_1, x_2, ..., x_n)$. If we set $p = 1$ obtain the mean for a discrete random variable:

$$\mu = E\left[X\right] = \sum_{i=1}^{n} x_i f(x_i)$$

> **Activity 2.5** Consider a stock that generates the following returns:

Payoff	*Probability*
2	0.40
0	0.50
-5	0.10

- Find the mean and standard deviation of the return on this stock.

2.5.3 Multiple random variables, bivariate CDFs, covariances

The above definitions apply to individual random variables, but we are often interested (particularly in forecasting) with multiple random variables. The equivalent definitions of *cdf* and pmf or *pdf* for the *bivariate* case are given below.

Definition 2.12 (Bivariate cumulative distribution function) The joint cumulative distribution function (or *cdf*) of two random variables X and Y is defined as:

$$F_{XY}(x, y) \equiv \Pr[X \leq x \cap Y \leq y]$$

If (X, Y) have *cdf* F_{XY}, we write that '(X, Y) is distributed according to F_{XY}', or '$(X, Y) \sim F_{XY}$' in shorthand.

The symbol $\cap$ denotes intersection, and can be thought of as 'and' in this application. Union is denoted by $\cup$ and can be thought of as 'or' in this application.

Definition 2.13 (Bivariate probability mass function) The probability mass function, f_{XY}, of **discrete** random variables X and Y is given by:

$$f_{XY}(x) \equiv \Pr[X = x \cap Y = y]$$

Definition 2.14 (Bivariate probability density function) The probability density function, f$_{XY}$, of **continuous** random variables X and Y is the function that satisfies:

$$F_{XY}(x) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(s, t)\, dtds \text{ for all } (x, y)$$

If F_{XY} is differentiable then the pdf may be obtained as

$$f_{XY}(x) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}$$

The *marginal* distribution of X (i.e., the distribution just of X rather than of both X and Y) is obtained as

$$\begin{aligned} F_X(x) &= F_{XY}(x, \infty) \\ &= \Pr[X \leq x \cap Y \leq \infty] \\ &= \Pr[X \leq x] \end{aligned}$$

The marginal density of X is obtained from the joint density by 'integrating out' Y :

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy$$

Recall that if X and Y are **independent** then their joint *cdf* is just the product of the univariate *cdf*s:

$$F_{XY}(x, y) \equiv \Pr[X \leq x \cap Y \leq y] = \Pr[X \leq x] \Pr[Y \leq y] \equiv F_x(x) F_y(y),$$

and if they are discrete then their joint *pmf* is the product of their univariate *pmf*s:

$$f_{XY}(x) \equiv \Pr[X = x \cap Y = y] = \Pr[X = x]\Pr[Y = y] \equiv f_x(x)\, f_y(y)\,;$$

if they are continuous then their joint *pdf* is the product of their univariate *pdf*s:

$$f_{XY}(x) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} = \frac{\partial^2}{\partial x \partial y}\left(F_x(x)\, F_y(y)\right) = f_x(x)\, f_y(y)$$

The following two important quantities are derived from the joint distribution of two random variables:

Definition 2.15 (Covariance) The covariance between the random variables X and Y with joint *pdf* f_{XY} is:

$$\begin{aligned} Cov[X,Y] &\equiv E[(X-\mu_x)(Y-\mu_y)] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)\, f_{XY}(x,y)\, dxdy \\ &= E[XY] - \mu_x\mu_y \end{aligned}$$

where $\mu_x = E[X]$ and $\mu_y = E[Y]$.

Definition 2.16 (Correlation) The correlation between the random variables X and Y is:

$$\begin{aligned} Corr[X,Y] &\equiv \frac{Cov[X,Y]}{\sqrt{V[X]\cdot V[Y]}} \\ -1 &\leq Corr[X,Y] \leq 1 \end{aligned}$$

More generally than covariance or correlation, we might be interested in the *conditional* distribution of one variable given information on the other. For example, we may be interested in the distribution of Y given X, written as $Y|X$. The conditional distribution of Y given $X \leq x$ is obtained as:

$$\begin{aligned} F_{Y|X\leq x}(y|X \leq x) &\equiv \Pr[Y \leq y | X \leq x] \\ &= \frac{\Pr[X \leq x \cap Y \leq y]}{\Pr[X \leq x]} \\ &\equiv \frac{F_{XY}(x,y)}{F_X(x)} \end{aligned}$$

More often, we are interested in the conditional distribution or density of $Y|X = x$, that is, the conditional *cdf* or *pdf* of Y given that X takes a particular value.

Definition 2.17 (Conditional CDF and PDF) The conditional *cdf* and *pdf* of $Y|X = x$ are given by:

$$\begin{aligned} F_{Y|X}(y|x) &= \frac{\partial F_{XY}(x,y)/\partial x}{f_X(x)} \\ f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} \end{aligned}$$

Recall that if X and Y are independent then the conditional *cdf* and *pdf* of $Y|X = x$ is equal to the unconditional *cdf* and *pdf* of Y.

> **Activity 2.6** Covariance and correlation measure the *linear* relationship between two variables, and it is possible that two variables have zero correlation but are not independent. For example, let $X \sim N(0,1)$, then
>
> 1. show that $Corr\left[X, X^2\right] = 0$.
> 2. show that X and X^2 are *not* independent.

2.5.4 Moments of random vectors

The above definitions refer to individual, *scalar*, random variables. These definitions also apply to vectors of random variables ('random vectors'). Below we use the fact that the expectation of a matrix (or vector) of random variables is equal to the matrix of expectations of the individual elements. In this subject guide (and in most text books) vectors will always be *column vectors*, and so to write these in a (horizontal) equation, I will often use the transpose operation. For example:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = [X_1, X_2, ..., X_n]'$$

Definition 2.18 (Mean of a vector) The mean of a $n \times 1$ vector random variable $\mathbf{X} \equiv [X_1, X_2, ..., X_n]'$ is μ, a $n \times 1$ vector:

$$\begin{aligned} \underset{(n\times1)}{\mu} &\equiv E[\underset{(n\times1)}{\mathbf{X}}] \\ &\equiv E\left[[X_1, X_2, ..., X_n]'\right] \\ &= \left[E[X_1], E[X_2], ..., E[X_n]\right]' \\ &\equiv [\mu_1, \mu_2, ..., \mu_n]' \end{aligned}$$

Definition 2.19 (Covariance matrix) The covariance matrix of a $n \times 1$ vector random variable $\mathbf{X} \equiv [X_1, X_2, ..., X_n]'$ is $\mathbf{\Sigma}$, a $n \times n$ matrix:

$$\begin{aligned} \underset{(n\times n)}{\mathbf{\Sigma}} &\equiv E[\underset{(n\times1)}{(\mathbf{X}-\mu)}\underset{(1\times n)}{(\mathbf{X}-\mu)'}] \\ &\equiv E\left[\begin{bmatrix} (X_1-\mu_1) \\ (X_2-\mu_2) \\ \vdots \\ (X_n-\mu_n) \end{bmatrix} \begin{bmatrix} (X_1-\mu_1) & (X_2-\mu_2) & \cdots & (X_n-\mu_n) \end{bmatrix}\right] \\ &= E\left[\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_n-\mu_n) \\ (X_1-\mu_1)(X_2-\mu_2) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (X_1-\mu_1)(X_n-\mu_n) & (X_2-\mu_2)(X_n-\mu_n) & \cdots & (X_n-\mu_n)^2 \end{bmatrix}\right] \\ &\equiv \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix} \end{aligned}$$

Notice that $\mathbf{\Sigma}$ is a symmetric matrix, so element (i,j) is equal to element (j,i), for any i, j.

Definition 2.20 (Correlation matrix) Any covariance matrix can be decomposed into a matrix containing the standard deviations on the diagonal and the correlation matrix

$$\begin{aligned}
\underset{(n\times n)}{\mathbf{\Sigma}} &= \underset{(n\times n)}{\mathbf{D}}\underset{(n\times n)}{\mathbf{R}}\underset{(n\times n)}{\mathbf{D}} \\
\text{where } \mathbf{D} &= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \\
\mathbf{R} &= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{12} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n} & \rho_{2n} & \cdots & 1 \end{bmatrix}
\end{aligned}$$

If all variances are strictly positive, then the correlation matrix can be obtained from the covariance matrix be pre- and post-multiplying the covariance matrix by $\mathbf{D}^{-1}$:

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}$$

Higher moments, such as skewness and kurtosis, can also be defined for vector random variables, but it requires some cumbersome notation, and will not be needed in this course.

> **Activity 2.7** Let $X \sim N(2,5)$, $Y \sim N(0,3)$ and $W \sim N(1,6)$, and assume that all three variables are independent of each other. Define $\mathbf{Z} = [X,Y,W]'$. Find the mean vector, covariance matrix, and correlation matrix of $\mathbf{Z}$.

2.6 Overview of chapter

This chapter introduced some key concepts and definitions from financial econometrics, and reviewed some fundamental definitions and results from statistics: definitions and calculation of moments (means, variances, skewness, etc.) and distribution and density functions. We considered both continuous and discrete random variables, and scalar and vector random variables.

2.7 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be:

- Able to compute arithmetic (simple) and logarithmic (continuously-compounded) returns

- Familiar with standard concepts from statistics: moments, distributions and densities
- Able to compute moments numerically for discrete random variables
- Able to derive means and variances for vector random variables.

2.8 Test your knowledge and understanding

1. The pay-off on a (risky) corporate bond for company XYZ is as follows

Payoff	*Probability*
100	0.80
70	0.05
50	0.10
0	0.05

   (a) Plot the *cdf* and *pmf* of the pay-off on this bond.

   (b) Find the mean (expected) pay-off on this bond.

   (c) Find the standard deviation of the pay-off on this bond.

2. Let $Z_{t+3} = \left(P_{t+3} - P_t\right)/P_t$ be the three-day arithmetic return, and let $W_{t+1} = \left(P_{t+1} - P_t\right)/P_t$ be the one-day arithmetic return. Find an expression for Z_{t+3} as a function of W_{t+1}, W_{t+2} and W_{t+3}.

3. Don't forget to check the VLE for additional practice problems for this chapter.

2.9 Solutions to activities

Activity 2.1

P_t	P_{t+1}	R^A_{t+1}	R^L_{t+1}
100	103	0.0300	0.0296
100	92	-0.0800	-0.0834
100	145	0.4500	0.3716
100	30	-0.7000	-1.2040

So when returns (either arithmetic or logarithmic) are small, less than around 0.10, the two definitions of returns are very close. When returns are large the two definitions give different answers. (Also notice that while arithmetic returns can never go below -100%, logarithmic returns can be below -100%: in the fourth row the arithmetic return is -70% while the logarithmic return is -120.4%.)

Activity 2.2

We are given:

$$\begin{aligned} W_{t+1} &= \frac{P_{t+1} - P_t}{P_t} \\ \text{and} \quad Z_{t+2} &= \frac{P_{t+2} - P_t}{P_t} \end{aligned}$$

then we derive

$$\begin{aligned} Z_{t+2} &= \frac{P_{t+2} - P_t}{P_t} \\ &= \frac{P_{t+2} - P_{t+1}}{P_t} + \frac{P_{t+1} - P_t}{P_t} \\ &= \frac{P_{t+2} - P_{t+1}}{P_{t+1}} \frac{P_{t+1}}{P_t} + W_{t+1} \\ &= W_{t+2} \left(1 + W_{t+1}\right) + W_{t+1} \end{aligned}$$

Activity 2.3

If $X \sim Unif\,(a, b)$, then its *cdf* is

$$\begin{aligned} F(x) &\equiv \Pr[X \leq x] = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & x \geq b \end{cases} \\ f(x) &\equiv \frac{\partial F(x)}{\partial x} = \begin{cases} 0, & x \leq a \\ \frac{1}{b-a}, & a < x < b \\ 0, & x \geq b \end{cases} \end{aligned}$$

(Note that the *pdf* is not a function of x: this means it is a flat line over the interval $[a, b]$, which is why this distribution is called the 'uniform' distribution.) So now we just need to find values of a and b where $f(x)$ is greater than one. This will be true for any a and b where $b - a < 1$. For example, if $a = 0$ and $b = 1/2$, then $f(x) = 1/(1/2) = 2$.

Activity 2.4

Skewness will be zero if $E[X^3] = 0$, and so we can just focus on this term.

$$
\begin{aligned}
E\left[X^3\right] &= \int_{-\infty}^{\infty} x^3 f(x)\,dx \\
&= \int_{-\infty}^{0} x^3 f(x)\,dx + \int_{0}^{\infty} x^3 f(x)\,dx \\
&= \int_{-\infty}^{0} x^3 f(-x)\,dx + \int_{0}^{\infty} x^3 f(x)\,dx, \text{ since } f(x) = f(-x) \\
&= \int_{0}^{\infty} \left(-x^3\right) f(x)\,dx + \int_{0}^{\infty} x^3 f(x)\,dx, \text{ changing the sign of } x \text{ in first integral} \\
&= \int_{0}^{\infty} \left(-x^3 + x^3\right) f(x)\,dx, \text{ gathering terms} \\
&= 0
\end{aligned}
$$

It can also be shown that skewness is zero when the variable is symmetric around some general point a, using the same logic as above (though a little more notation is needed).

Activity 2.5

It is useful to expand the table and fill it with some other calculations:

	Payoff	*Probability*	*Payoff*^{2}	*Payoff*×*Prob*	*Payoff*^{2}×*Prob*
	2	0.4	4	0.8	1.6
	0	0.5	0	0.0	0.0
	-5	0.1	25	-0.5	2.5
Sum	-3	1.0	29	0.3	4.1

Then we can obtain the mean:

$$E[X] = \sum_{i=1}^{3} x_i f(x_i) = 2 \times 0.4 + 0 \times 0.5 + (-5) \times 0.1 = 0.3$$

Next we compute the uncentered second moment, the variance and the standard deviation:

$$
\begin{aligned}
E\left[X^2\right] &= \sum_{i=1}^{3} x_i^2 f(x_i) = 4 \times 0.4 + 0 \times 0.5 + 25 \times 0.1 = 4.1 \\
V[X] &= E\left[X^2\right] - E[X]^2 = 4.1 - 0.3^2 = 4.01 \\
\text{Standard deviation} &= \sqrt{V[X]} = \sqrt{4.01} = 2.0025
\end{aligned}
$$

Activity 2.6

Part (1): we first compute the covariance:

$$
\begin{aligned}
Cov\left[X, X^2\right] &= E\left[X^3\right], \text{ since } E[X] = 0 \\
&= 0, \text{ since } X \sim N(0,1) \text{ is symmetric.}
\end{aligned}
$$

This implies that the correlation of X and X^2 is also zero. Thus even though X and X^2 are clearly strongly related, they have zero correlation.

Part (2): There are many possible ways to show that these variables are not independent. The easiest is to show that the conditional density of X given X^2 is not equal to the unconditional density of X. For example, assume that X^2 takes the value of 2. Then there are only two possible values for X, namely $\sqrt{2}$ and $-\sqrt{2}$, so the distribution of $X|X^2 = 2$ is discrete:

$$f_{x|x^2}\left(x|x^2 = 2\right) = \begin{cases} -\sqrt{2}, & \text{prob } 1/2 \\ +\sqrt{2}, & \text{prob } 1/2 \end{cases},$$

whereas the unconditional distribution of X is $N(0,1)$, which is continuous. Thus we have easily shown that $f_{x|x^2} \neq f_x$ and so X is not independent of X^2.

Activity 2.7

$$\begin{aligned} \mathbf{Z} &= [X, Y, W]' \\ \text{so } E[\mathbf{Z}] &= E\left[[X,Y,W]'\right] = [E[X], E[Y], E[W]]' = [2, 0, 1]' \end{aligned}$$

The covariance matrix is

$$\begin{aligned} \mathbf{\Sigma} &\equiv V[\mathbf{Z}] = V\left[[X,Y,W]'\right] \\ &= \begin{bmatrix} V[X] & Cov[X,Y] & Cov[X,W] \\ Cov[X,Y] & V[Y] & Cov[Y,W] \\ Cov[X,W] & Cov[Y,W] & V[Z] \end{bmatrix} \\ &= \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{bmatrix} \end{aligned}$$

All of the covariances are zero since we are told that the variables are independent.

Finally, the correlation matrix is very simple (this was a bit of a trick question): since all the covariances are zero, all correlations are also zero, and so there is nothing to work out:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Chapter 3
Basic time series concepts

3.1 Introduction

Many problems in quantitative finance involve the study of financial data. Such data most often comes in the form of 'time series,' which is a sequence of random variables that are ordered through time. Before moving on to financial applications, we must first cover some fundamental topics in time series analysis, such as autocorrelation, white noise processes and ARMA procrresses (covered in the next chapter). These two chapters are the most theoretical in this guide, and it may not appear too related to finance, but they lay the foundations for the topics we will cover in later chapters.

3.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce standard measures for predictability of time series, autocovariances and autocorrelations, and a class of time series that are *not* predictable, 'white noise' processes.
- Present the 'law of iterated expectations' and illustrate its use in time series analysis.

3.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Describe the various forms of 'white noise' processes used in the analysis of financial data
- Use the 'law of iterated expectations' to derive unconditional means from conditional means
- Apply these tools to a simple AR(1) process

3.1.3 Essential reading

- Diebold, F.X. *Elements of Forecasting.* (Thomson South-Western, Canada, 2006) fourth edition [ISBN 9780324323597], Chapters 7 and 8 (only the parts that overlap with these notes).

3.1.4 Further reading

- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354], Chapter 2.

3.2 Covariance stationary time series

We denote the sample of observations on a time series as $(y_1, y_2, ..., y_T)$. Observations before the start of the sample are $(..., y_{-2}, y_{-1}, y_0)$ and observations beyond the end are denoted $(y_{T+1}, y_{T+2}, ...)$. The doubly infinite sequence of observations on this series is

$$\{y_t\}_{t=-\infty}^{\infty} = (..., y_{-1}, y_0, \underbrace{y_1, y_2, ..., y_T}_{\text{observed data}}, y_{T+1}, y_{T+2}, ...)$$

We will generally focus on the special case of 'covariance stationary' time series, where:

$$\begin{aligned} E[Y_t] &= \mu \;\forall\; t \\ V[Y_t] &= \sigma^2 \;\forall\; t \\ \text{and} \quad Cov[Y_t, Y_{t-j}] &= \gamma_j, \;\forall\; j, t \end{aligned}$$

(The notation '$\forall$ t' means 'for all t.') The first two conditions imply that the *unconditional* means and variances of each of the Y_t's are assumed to be the same through time. This does not imply that their *conditional* means and variances will be the same, and we will spend a lot of time looking at these. Many economic and financial time series can be treated as though this assumption holds. The third condition implies that that all autocovariances, denoted γ_j and defined below, are also constant through time, so when describing an autocovariance we need only denote it with a 'j,' not also with a t.

Definition 3.1 (Autocovariance) The j^{th}-order autocovariance of a time series Y_t is:

$$\begin{aligned} \gamma_j &= Cov[Y_t, Y_{t-j}] \\ &= E[(Y_t - \mu)(Y_{t-j} - \mu)] \\ &= E[Y_t \cdot Y_{t-j}] - \mu^2 \end{aligned}$$

Note that $\gamma_0 = Cov[Y_t, Y_t] = V[Y_t]$.

Definition 3.2 (Autocorrelation) The j^{th}-order autocorrelation of a time series Y_t is:

$$\begin{aligned} \rho_j &= Corr[Y_t, Y_{t-j}] \\ &= \frac{Cov[Y_t, Y_{t-j}]}{V[Y_t]} \\ &\equiv \frac{\gamma_j}{\gamma_0} \end{aligned}$$

Note that $\rho_0 = Corr[Y_t, Y_t] = 1$.

Autocovariance and autocorrelation are sometimes referred to as 'serial covariance' and 'serial correlation.' We will use 'autocorrelation' and 'serial correlation' interchangeably.

Example 3.1 Let Y_t take the value 1 or 0 depending on whether the t^{th} coin toss came up heads or tails. If the coin is fair (that is, the probability of seeing a 'head' is always equal to one-half) then all of the autocorrelations of Y_t are zero. That is

$$Corr\left[Y_t, Y_{t-j}\right] = 0 \ \forall \ j \neq 0$$

This is because a fair coin has no 'memory': the probability of seeing a tail at time t is unaffected by whether we saw a tail at time $t - j$ $(j \neq 0)$.

3.3 The Law of Iterated Expectations

A useful tool in time series analysis is a result known as the 'Law of Iterated Expectations.'

Definition 3.3 (Law of Iterated Expectations) Let I_1 and I_2 be two information sets, and assume that $I_1 \subseteq I_2$, i.e., I_2 is 'bigger' than I_1. Then

$$E\left[\ E\left[Y|I_2\right] \quad | \quad I_1\right] = E\left[Y|I_1\right]$$

The Law of Iterated Expectations is sometimes shortened to the 'LIE'.

Example 3.2 Let I_t be all the information available as at date t, so $I_t \subseteq I_{t+1}$. Then

$$E_t\left[\ E_{t+1}\left[Y_{t+2}\right]\ \right] = E_t\left[Y_{t+2}\right]$$

Example 3.3 Again, let I_t be all the information available as at date t, and notice that an unconditional expectation employs an 'empty' information set, which is smaller than any non-empty information set. Then

$$E\left[\ E_t\left[Y_{t+1}\right]\ \right] = E\left[Y_{t+1}\right]$$

Thus the unconditional expectation of the conditional expectation of some random variable is equal to the unconditional expectation of the random variable. This result is employed in many different places in time series analysis.

Activity 3.1 (1) If $X = -1$ with probability $1/2$ and $X = +1$ with probability $1/2$, and $E\left[Y|X\right] = X$, show that $E\left[Y\right] = 0$.

(2) If $E\left[Y|X\right] = X^2$, and $X \sim N\left(0, 3^2\right)$, find $E\left[Y\right]$.

Almost everything we derive in this chapter, and much of what we will derive in later chapers, is based on just three things:

1. Rules for expectations, variances and covarianes
2. Implications of covariance stationarity
3. The law of iterated expectations

3.3.1 Refresher: Rules for expectations, variances and covariances

Let X, Y and Z be three (scalar) random variables, and let a, b, c and d be constants. Then:

$$\begin{aligned}
E\left[a+bX\right] &= a+bE\left[X\right] \\
\\
V\left[a+bX\right] &= V\left[bX\right]=b^{2}V\left[X\right] \\
\\
Cov\left[a+bX,cY\right] &= Cov\left[bX,cY\right]=bc\cdot Cov\left[X,Y\right] \\
\\
E\left[a+bX+cY\right] &= a+bE\left[X\right]+cE\left[Y\right] \\
\\
V\left[a+bX+cY\right] &= V\left[bX+cY\right] \\
&= V\left[bX\right]+V\left[cY\right]+2Cov\left[bX,cY\right] \\
&= b^{2}V\left[X\right]+c^{2}V\left[Y\right]+2bc\cdot Cov\left[X,Y\right] \\
\\
Cov\left[a+bX,cY+dZ\right] &= Cov\left[bX,cY+dZ\right] \\
&= Cov\left[bX,cY\right]+Cov\left[bX,dZ\right] \\
&= bc\cdot Cov\left[X,Y\right]+bd\cdot Cov\left[X,Z\right]
\end{aligned}$$

> **Activity 3.2** Consider two stocks that generate returns as $X\sim N\left(1,2\right)$ and $Y\sim N\left(2,3\right)$ and assume that these returns are independent. Now consider two *portfolios* of these two stocks, where
>
> $$\begin{aligned}
W &= \frac{1}{2}X+\frac{1}{2}Y \\
Z &= \frac{3}{4}X+\frac{1}{4}Y
\end{aligned}$$
>
> Let $\mathbf{U}=\left[W,Z\right]'$. Find the mean vector, covariance matrix, and correlation matrix of $\mathbf{U}$.

3.4 White noise and other innovation series

One of the main building blocks in forecasting is a time series process called a 'white noise process.' This is a variable with zero serial correlation. We usually also assume that it has zero mean, just for simplicity, but its main characteristic is its lack of serial correlation. There are three main types of innovation series that people consider: white noise, *iid* white noise and Gaussian white noise.

Definition 3.4 (White noise) ε_t is a white noise process if

$$Corr\left[\varepsilon_t,\varepsilon_{t-j}\right]=0\ \forall\ j\neq 0$$

In this case we write that $\varepsilon_t \sim WN$. If in addition we know that $E[\varepsilon_t] = 0$ then we write $\varepsilon_t \sim WN(0)$, and say that ε_t is a zero-mean white noise process. If we know that $E[\varepsilon_t] = 0$ and $V[\varepsilon_t] = \sigma^2$ then we write $\varepsilon_t \sim WN(0, \sigma^2)$ and we say that ε_t is a zero-mean white noise process with variance σ^2.

Definition 3.5 (iid white noise) ε_t is independent and identically distributed (iid) white noise if

$$\begin{aligned} &\varepsilon_t \text{ is independent of } \varepsilon_{t-j}\ \forall\ j \neq 0, \text{ and} \\ &\varepsilon_t \sim\ F\ \forall \text{ t, where F is some distribution} \end{aligned}$$

In this case we write that $\varepsilon_t \sim iid\ WN$ (or $\varepsilon_t \sim iid\ F$). If in addition we know that $E[\varepsilon_t] = 0$ then we write $\varepsilon_t \sim iid\ WN(0)$, and say that ε_t is a zero-mean *iid* white noise process. If we know that $E[\varepsilon_t] = 0$ and $V[\varepsilon_t] = \sigma^2$ then we write $\varepsilon_t \sim iid\ WN(0, \sigma^2)$ and we say that ε_t is a zero-mean *iid* white noise process with variance σ^2.

Definition 3.6 (Gaussian white noise) ε_t is Gaussian white noise if

$$\varepsilon_t \sim iid\ N\left(0, \sigma^2\right)$$

Notice that these three definitions carry an increasing amount of information. Simple white noise only imposes that the process has zero serial correlation. *iid* white noise imposes zero serial *dependence*, which is stronger than zero serial correlation, and further imposes that the distribution is the same at all points in time. Gaussian white noise imposes both serial independence and a distributional assumption on the series. Most often, people work with Gaussian white noise (it simplifies many calculations) but it should be noted that this is the most restrictive form of white noise.

Activity 3.3 Any time series, Y_{t+1}, may be decomposed into its conditional mean, $E_t[Y_{t+1}]$, and a 'remainder' process, ε_{t+1}:

$$Y_{t+1} = E_t[Y_{t+1}] + \varepsilon_{t+1}$$

1. Show that ε_{t+1} is a zero-mean white noise process.
2. Show that ε_{t+1} has mean zero *conditional* on the information set available at time t.
3. The white noise term ε_{t+1} is uncorrelated with the conditional mean term, $E_t[Y_{t+1}]$. (Hint: it might make the problem easier if you define a new variable, $\mu_{t+1} = E_t[Y_{t+1}]$, and treat μ_{t+1} as a separate random variable which is observable at time t. Note that we denote it with a subscript '$t+1$' even though it is observable at time t.)

3.5 Application to an AR(1) process

Let us now consider a time series process Y_t, defined as follows:

$$Y_t = \phi Y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim WN\left(0, \sigma^2\right) \text{ and } |\phi| < 1$$

The above equation is a particular type of time series, namely an 'autoregressive process of order 1' or a 'first-order autoregression', or an 'AR(1)' process. It's called this because the variable Y_t is 'regressed' onto itself (the 'auto' part of the name) lagged 1 period (the 'first-order' part of the name). For the rest of this course you can assume that the time series we consider are stationary (we will cover non-stationary processes later in the notes). Notice that the time series defined in the equation above has only two fundamental parameters: ϕ, called the (first-order) autoregressive coefficient, and σ^2, the variance of the 'innovation process', ε_t. All the properties of Y_t are simply functions of ϕ and σ^2, and when asked for a particular property of Y_t it should always be given as a function of ϕ and σ^2.

Problem 1

What is the unconditional mean of Y_t?

$$\begin{aligned} E\left[Y_t\right] &= E\left[\phi Y_{t-1}+\varepsilon_t\right], \text{ substituting in for } Y_t \\ &= E\left[\phi Y_{t-1}\right]+E\left[\varepsilon_t\right], \text{ property of the expectations operator} \\ &= \phi E\left[Y_{t-1}\right]+E\left[\varepsilon_t\right], \text{ since } \phi \text{ is constant} \\ &= \phi E\left[Y_{t-1}\right], \text{ because } \varepsilon_t \text{ has zero mean} \end{aligned}$$

Since Y_t is stationary, we know that $\mu = E\left[Y_t\right] = E\left[Y_{t-1}\right]$. So

$$\begin{aligned} E\left[Y_t\right] &= \phi E\left[Y_{t-1}\right] \\ \mu &= \phi\mu \\ \mu\left(1-\phi\right) &= 0 \text{ , which implies that} \\ \mu &= 0 \text{ as } \left|\phi\right| < 1 \end{aligned}$$

Problem 2

What is the unconditional variance of Y_t?

$$\begin{aligned} \gamma_0 &\equiv V\left[Y_t\right] \\ &= V\left[\phi Y_{t-1}+\varepsilon_t\right] \\ &= V\left[\phi Y_{t-1}\right]+V\left[\varepsilon_t\right]+2Cov\left[\phi Y_{t-1},\varepsilon_t\right] \\ &= \phi^2 V\left[Y_{t-1}\right]+\sigma^2+0 \end{aligned}$$

Again, since Y_t is stationary, we know that $V\left[Y_t\right] = \gamma_0 \ \forall \ t$. Then

$$\begin{aligned} \gamma_0 &= \phi^2\gamma_0+\sigma^2 \\ \gamma_0\left(1-\phi^2\right) &= \sigma^2, \text{ and so} \\ \gamma_0 &= \frac{\sigma^2}{1-\phi^2} \end{aligned}$$

Problem 3

What is the first-order autocovariance and autocorrelation of Y_t?

$$\begin{aligned}
Cov\left[Y_t, Y_{t-1}\right] &= Cov\left[\phi Y_{t-1} + \varepsilon_t, Y_{t-1}\right] \\
&= \phi Cov\left[Y_{t-1}, Y_{t-1}\right] + Cov\left[\varepsilon_t, Y_{t-1}\right] \\
&= \phi V\left[Y_{t-1}\right] + 0 \\
&= \phi\gamma_0 \\
&= \phi\frac{\sigma^2}{1-\phi^2} \\
\text{So}\quad \rho_1 &\equiv Corr\left[Y_t, Y_{t-1}\right] = \frac{Cov\left[Y_t, Y_{t-1}\right]}{V\left[Y_t\right]} \equiv \frac{\gamma_1}{\gamma_0} = \phi
\end{aligned}$$

Problem 4

What is the second-order autocovariance and autocorrelation of Y_t?

$$\begin{aligned}
\gamma_2 &\equiv Cov\left[Y_t, Y_{t-2}\right] \\
&= Cov\left[\phi Y_{t-1} + \varepsilon_t, Y_{t-2}\right] \\
&= \phi Cov\left[Y_{t-1}, Y_{t-2}\right] + Cov\left[\varepsilon_t, Y_{t-2}\right] \\
&= \phi\gamma_1, \quad \text{since } Cov\left[\varepsilon_t, Y_{t-2}\right] = 0 \\
&= \phi^2\gamma_0, \quad \text{since } \gamma_1 = \phi\gamma_0 \\
&= \phi^2\frac{\sigma^2}{1-\phi^2} \\
\text{So}\quad \rho_2 &\equiv Corr\left[Y_t, Y_{t-2}\right] = \frac{Cov\left[Y_t, Y_{t-1}\right]}{V\left[Y_t\right]} \equiv \frac{\gamma_2}{\gamma_0} = \phi^2
\end{aligned}$$

3.6 Overview of chapter

This chapter introduced some fundamental topics in time series analysis, such as autocorrelation and white noise processes. We reviewed rules for computing means, variances and covariances, and combining those rules in conjuction with implications of covariance stationarity and the law of iterated expectations, we derived some theoretical results for a first-order autoregressive (AR(1)) process.

3.7 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be:

- Describe the various forms of 'white noise' processes used in the analysis of financial data

- Use the 'law of iterated expectations' to derive unconditional means from conditional means
- Compute the mean, variance and autocovariances for an AR(1) process

3.8 Test your knowledge and understanding

1. If Y_t follows an AR(1) process:

$$\begin{aligned} Y_t &= \phi Y_{t-1} + \varepsilon_t, \quad |\phi| < 1 \\ \varepsilon_t &\sim iid\ N\left(0, \sigma^2\right) \end{aligned}$$

   then find the following quantities:

   (a) $E_t\left[Y_{t+1}\right]$

   (b) $E\left[Y_{t+1}\right]$

   (c) $E\left[\varepsilon_{t+1}\right]$

   (d) $E_t\left[E_{t+1}\left[Y_{t+2}\right]\right]$

   (e) $E_t[Y_{t+2}]$

   (f) $Cov\left[Y_t, Y_{t-j}\right]$ for any $j \geq 0$?

2. Verify explicitly that $E_t\left[Y_{t+2}\right] = E_t\left[E_{t+1}\left[Y_{t+2}\right]\right]$, as implied by the law of iterated expectations, for an AR(1) process.

3. Don't forget to check the VLE for additional practice problems for this chapter.

3.9 Solutions to activities

Activity 3.1

Part (1): If $X = -1$ with probability 1/2 and $X = +1$ with probability 1/2, and $E\left[Y|X\right] = X$, show that $E\left[Y\right] = 0$.

$$\begin{aligned} \text{If } E\left[Y|X\right] &= X, \\ \text{then } E\left[E\left[Y|X\right]\right] &= E\left[X\right], \text{ taking unconditional expectation of both sides} \\ \text{Note that } E\left[E\left[Y|X\right]\right] &= E\left[Y\right], \text{ by the LIE} \end{aligned}$$

so then we must simply solve for the unconditional mean of X

$$\begin{aligned} X &= \begin{cases} +1 & \text{with prob } 1/2 \\ -1 & \text{with prob } 1/2 \end{cases} \\ \text{so } E\left[X\right] &= 1/2\left(+1\right) + 1/2\left(-1\right) = 0 \\ \text{Thus } E\left[Y\right] &= E\left[X\right] = 0 \end{aligned}$$

Part (2): If $E[Y|X] = X^2$, and $X \sim N(0, 3^2)$, find $E[Y]$.

$$\begin{aligned} E[Y|X] &= X^2 \\ \text{so } E[E[Y|X]] &= E\left[X^2\right] \\ \text{and by the LIE we know } E[E[Y|X]] &= E[Y] \end{aligned}$$

so we just need to work out $E[X^2]$. Recall $V[X] = E[X^2] - E[X]^2$,and since $X \sim N(0, 3^2)$ we know $E[X] = 0$ and $V[X] = 9$. So $E[Y] = E[X^2] = 9$.

Activity 3.2

First we compute the elements of the mean vector:

$$\begin{aligned} E[W] &= E\left[\frac{1}{2}X + \frac{1}{2}Y\right] = \frac{1}{2}E[X] + \frac{1}{2}E[Y] = 1.5 \\ E[Z] &= E\left[\frac{3}{4}X + \frac{1}{4}Y\right] = \frac{3}{4}E[X] + \frac{1}{4}E[Y] = 1.25 \\ \text{so } E[\mathbf{U}] &= [E[W], E[Z]]' = [1.5, 1.25]' \end{aligned}$$

Next we compute the variances:

$$\begin{aligned} V[W] &= V\left[\frac{1}{2}X + \frac{1}{2}Y\right] = \frac{1}{4}V[X] + \frac{1}{4}V[Y] + \frac{2}{4}Cov[X,Y] = 1.25 \\ V[Z] &= V\left[\frac{3}{4}X + \frac{1}{4}Y\right] = \frac{9}{16}V[X] + \frac{1}{16}V[Y] + \frac{6}{16}Cov[X,Y] = 1.3125 \end{aligned}$$

and the covariances:

$$\begin{aligned} Cov[W,Z] &= Cov\left[\frac{1}{2}X + \frac{1}{2}Y, \frac{3}{4}X + \frac{1}{4}Y\right] \\ &= Cov\left[\frac{1}{2}X, \frac{3}{4}X\right] + Cov\left[\frac{1}{2}X, +\frac{1}{4}Y\right] + Cov\left[\frac{1}{2}Y, \frac{3}{4}X\right] + Cov\left[\frac{1}{2}Y, \frac{1}{4}Y\right] \\ &= \frac{3}{8}V[X] + \frac{1}{8}Cov[X,Y] + \frac{3}{8}Cov[X,Y] + \frac{1}{8}V[Y] \\ &= 1.125 \end{aligned}$$

and so

$$V[\mathbf{U}] = V\begin{bmatrix} W \\ Z \end{bmatrix} = \begin{bmatrix} V[W] & Cov[W,Z] \\ Cov[W,Z] & V[Z] \end{bmatrix} = \begin{bmatrix} 1.25 & 1.125 \\ 1.125 & 1.3125 \end{bmatrix}$$

Finally we derive the correlation:

$$\begin{aligned} Corr[W,Z] &= \frac{Cov[W,Z]}{\sqrt{V[W]V[Z]}} = \frac{1.125}{\sqrt{1.25 \times 1.3125}} = 0.878 \\ \text{so } \mathbf{R} &= \begin{bmatrix} 1 & 0.878 \\ 0.878 & 1 \end{bmatrix} \end{aligned}$$

Activity 3.3

We have been given

$$\begin{aligned} Y_{t+1} &= E_t\left[Y_{t+1}\right] + \varepsilon_{t+1} \\ \text{so } \varepsilon_{t+1} &= Y_{t+1} - E_t\left[Y_{t+1}\right] \end{aligned}$$

It is easiest if we obtain the answer to part (2) of this problem first:

$$E_t\left[\varepsilon_{t+1}\right] = E_t\left[Y_{t+1}\right] - E_t\left[Y_{t+1}\right] = 0$$

so ε_{t+1} has conditional mean zero. This implies

$$\begin{aligned} E\left[\varepsilon_{t+1}\right] &= E\left[E_t\left[\varepsilon_{t+1}\right]\right] \text{ by the LIE} \\ &= 0, \text{ since } E_t\left[\varepsilon_{t+1}\right] = 0 \end{aligned}$$

and so ε_{t+1} also has unconditional mean zero. Next show that it is serially uncorrelated:

$$\begin{aligned} Cov\left[\varepsilon_t, \varepsilon_{t-j}\right] &= E\left[\varepsilon_t \varepsilon_{t-j}\right], \text{ since } E\left[\varepsilon_{t+1}\right] = 0 \\ &= E\left[E_{t-j}\left[\varepsilon_t \varepsilon_{t-j}\right]\right] \text{ by the LIE} \\ &= E\left[E_{t-j}\left[\varepsilon_t\right] \varepsilon_{t-j}\right] \text{ since } \varepsilon_{t-j} \text{ is known at time } t-j \\ &= 0 \text{ since } E_{t-j}\left[\varepsilon_t\right] = E_{t-j}\left[E_{t-1}\left[\varepsilon_t\right]\right] = 0 \end{aligned}$$

Thus $Cov\left[\varepsilon_t, \varepsilon_{t-j}\right] = 0$ for all $j > 0$, and so ε_{t+1} is serially uncorrelated. So we have shown that ε_{t+1} is a zero-mean white noise process (part (1) of the problem) by using the answer to part (2). Now we move to part (3):

$$\begin{aligned} Cov\left[\varepsilon_{t+1}, E_t\left[Y_{t+1}\right]\right] &\equiv Cov\left[\varepsilon_{t+1}, \mu_{t+1}\right] \\ &= E\left[\varepsilon_{t+1}\mu_{t+1}\right], \text{ since } E\left[\varepsilon_{t+1}\right] = 0 \\ &= E\left[E_t\left[\varepsilon_{t+1}\mu_{t+1}\right]\right] \text{ by the LIE} \\ &= E\left[E_t\left[\varepsilon_{t+1}\right]\mu_{t+1}\right], \text{ since } \mu_{t+1} \equiv E_t\left[Y_{t+1}\right] \text{ is known at time } t \\ &= 0, \text{ since } E_t\left[\varepsilon_{t+1}\right] = 0. \end{aligned}$$

Thus ε_{t+1} is uncorrelated with the conditional mean term, $\mu_{t+1} \equiv E_t\left[Y_{t+1}\right]$.

Chapter 4
ARMA processes

4.1 Introduction

This chapter builds on the concepts and techniques introduced in the previous chatper. We will study the class of autoregressive-moving average (ARMA) processes, which are a widely-used and very useful class of models for time series data.

4.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce the most widely-used classes of time series models: autoregressive (AR), moving average (MA) and ARMA processes.

- Derive some population properties of the means, variances and autocovariances of AR, MA and ARMA processes.

4.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Compute the means and variances of some standard ARMA processes

- Derive autocorrelations for some simple ARMA processes

- Describe the various forms of 'white noise' processes used in the analysis of financial data

- Use the 'law of iterated expectations' to derive unconditional means from conditional means

4.1.3 Essential reading

- Diebold, F.X. *Elements of Forecasting.* (Thomson South-Western, Canada, 2006) fourth edition [ISBN 9780324323597], Chapters 7 and 8 (only the parts that overlap with these notes).

4.1.4 Further reading

- Stock, J.H. and M.W. Watson *Introduction to Econometrics.* (Pearson Education, Boston, 2010) third edition. [ISBN 9781408264331]. Chapter 14, Sections 1–5.
- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354], Chapter 2.

4.2 Autoregressive-moving average (ARMA) processes

4.2.1 Autoregressive (AR) processes

The previous chapter introduced the first-order autoregressive (AR(1)) process, and we now consider its more general form: a p^{th}-order AR process, or AR(p) process:

$$\begin{aligned} Y_t &\sim AR(p) \\ Y_t &= \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t,\ \varepsilon_t \sim WN(0) \end{aligned}$$

This specification posits that the dependent variable is a function of p lags (and a constant) and is useful when modelling series that have more complicated dynamics than a simple AR(1) process allows.

4.2.2 The MA(1) process

Next consider a time series process Y_t, defined as follows:

$$Y_t = \phi_0 + \varepsilon_t + \theta\varepsilon_{t-1},\ \varepsilon_t \sim WN\left(0, \sigma^2\right)$$

The above equation is another common type of time series, known as a 'moving average process of order 1', or a 'first-order moving average', or a 'MA(1)' process. By recursive substitution we can get some insight as to where this process gets its name (we drop the intercept, ϕ_0, for simplicity):

$$\begin{aligned} Y_t &= \varepsilon_t + \theta\varepsilon_{t-1} \\ &= \varepsilon_t + \theta\left(Y_{t-1} - \theta\varepsilon_{t-2}\right) = \varepsilon_t + \theta Y_{t-1} - \theta^2\varepsilon_{t-2} \\ &= \varepsilon_t + \theta Y_{t-1} - \theta^2\left(Y_{t-2} - \theta\varepsilon_{t-3}\right) = \varepsilon_t + \theta Y_{t-1} - \theta^2 Y_{t-2} + \theta^3\varepsilon_{t-3} \\ &= ... \\ &= \varepsilon_t + \sum_{i=1}^{\infty}(-1)^{i+1}\theta^i Y_{t-i} \end{aligned}$$

Thus an $MA(1)$ process can be re-written as a weighted average of all lags of Y_t plus some innovation term.

> **Activity 4.1** Let Y_t be defined as below. Find its mean, variance, first autocovariance and second autocovariance.
>
> $$Y_t = \phi_0 + \varepsilon_t + \theta\varepsilon_{t-1},\ \varepsilon_t \sim WN\left(0, \sigma^2\right)$$

4.2.3 Moving average (MA) processes

A general moving average process of order q is:

$$Y_t = \phi_0 + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ... + \theta_q\varepsilon_{t-q},\ \varepsilon_t \sim WN(0)$$

4.2.4 ARMA processes

The natural extension of both AR and MA processes is to combine them together in what is called an ARMA(p,q) process:

$$\begin{aligned} Y_t &= \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t \\ &\quad +\theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ... + \theta_q\varepsilon_{t-q},\ \varepsilon_t \sim WN(0) \end{aligned}$$

The ARMA model is the workhorse of time series forecasting. Generally p and q are set to be between 0 and 2, and so the number of parameters to be estimated is small, but the possible dynamics in the series Y_t is quite flexible.

4.3 Autocovariance functions

Recall from above that the j^{th}-order autocovariance is defined as

$$\begin{aligned} \gamma_j &= Cov[Y_t, Y_{t-j}] \\ &= E[Y_t \cdot Y_{t-j}] - \mu^2 \end{aligned}$$

If we consider γ_j as a function of the lag, j, we obtain the 'autocovariance function', or ACF, a function that is very useful for preliminary analyses of time series, prior to specifying a model.

Consider the AR(1) example we examined previously.

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

We showed that for this process the first two autocovariances are

$$\begin{aligned} \gamma_1 &= \frac{\phi\sigma^2}{1-\phi^2} = \phi\gamma_0 \\ \gamma_2 &= \frac{\phi^2\sigma^2}{1-\phi^2} = \phi^2\gamma_0 \end{aligned}$$

where $\gamma_0 = V[Y_t]$, which was denoted σ_y^2 above. It can be shown (see Activity 3.4) that for stationary AR(1) processes, the j^{th}-order autocovariance is

$$\gamma_j = \phi^j\gamma_0 \ \text{ for } j \geq 0$$

Let us examine some numerical examples. Consider the AR(1) processes:

$$\begin{aligned} Y_t &= 0.8Y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim WN(0,1) \\ Y_t &= 0.2Y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim WN(0,1) \\ Y_t &= -0.8Y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim WN(0,1) \\ Y_t &= -0.2Y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim WN(0,1) \end{aligned}$$

When looking at autocovariance plots, interpretation becomes easier if we standardise them by the variance of the process, and so transform them into autocorrelation plots.

In the top panel of Figure 4.1 we see that the processes with the smaller (in absolute value) AR coefficients have ACFs that are closer to zero, and converge to zero much quicker than those with larger AR coefficients. Note that the two AR(1) processes with a negative AR coefficient have autocorrelation coefficients that alternate sign.

The ACFs that can be generated from AR(2) processes are much more flexible than those from AR(1) processes. To see this, we first need to work out the autocovariance function for a general AR(2) process. It can be shown (but is not required for this course) that for an AR(2) the ACF is:

$$\begin{aligned}
Y_t &= \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t,\ \varepsilon_t \sim WN\left(0, \sigma^2\right) \\
\rho_1 &= \frac{\phi_1}{1-\phi_2} \\
\rho_j &= \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2} \text{ for } j \geq 2
\end{aligned}$$

We now look at a few numerical examples of AR(2) processes:

$$\begin{aligned}
Y_t &= 0.6Y_{t-1} + 0.2Y_{t-2} + \varepsilon_t \\
Y_t &= 0.1Y_{t-1} + 0.7Y_{t-2} + \varepsilon_t \\
Y_t &= 0.4Y_{t-1} - 0.4Y_{t-2} + \varepsilon_t
\end{aligned}$$

The autocorrelation functions of these processes is given in the lower panel of Figure 4.1.

Now let us look at the autocorrelations of some MA processes. We already derived the autcovariance function for an MA(1) process, if we note that $\gamma_j = 0$ for $j \geq 2$ (which is simple to show). Next consider the ACF of an MA(2) process.

> **Activity 4.2** Let Y_t be defined as below. Find its ACF.
>
> $$Y_t = \phi_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2},\ \varepsilon_t \sim WN\left(0, \sigma^2\right)$$

Let us now look at the ACFs of a few MA(1) and MA(2) processes, presented in Figure 4.2.

$$\begin{aligned}
Y_t &= \varepsilon_t + 0.8\varepsilon_{t-1},\ \varepsilon_t \sim WN(0,1) \\
Y_t &= \varepsilon_t + 0.2\varepsilon_{t-1},\ \varepsilon_t \sim WN(0,1) \\
Y_t &= \varepsilon_t - 0.5\varepsilon_{t-1},\ \varepsilon_t \sim WN(0,1) \\
Y_t &= \varepsilon_t + 0.6\varepsilon_{t-1} + 0.2\varepsilon_{t-2},\ \varepsilon_t \sim WN(0,1) \\
Y_t &= \varepsilon_t + 0.1\varepsilon_{t-1} + 0.7\varepsilon_{t-2},\ \varepsilon_t \sim WN(0,1) \\
Y_t &= \varepsilon_t + 0.4\varepsilon_{t-1} - 0.4\varepsilon_{t-2},\ \varepsilon_t \sim WN(0,1)
\end{aligned}$$

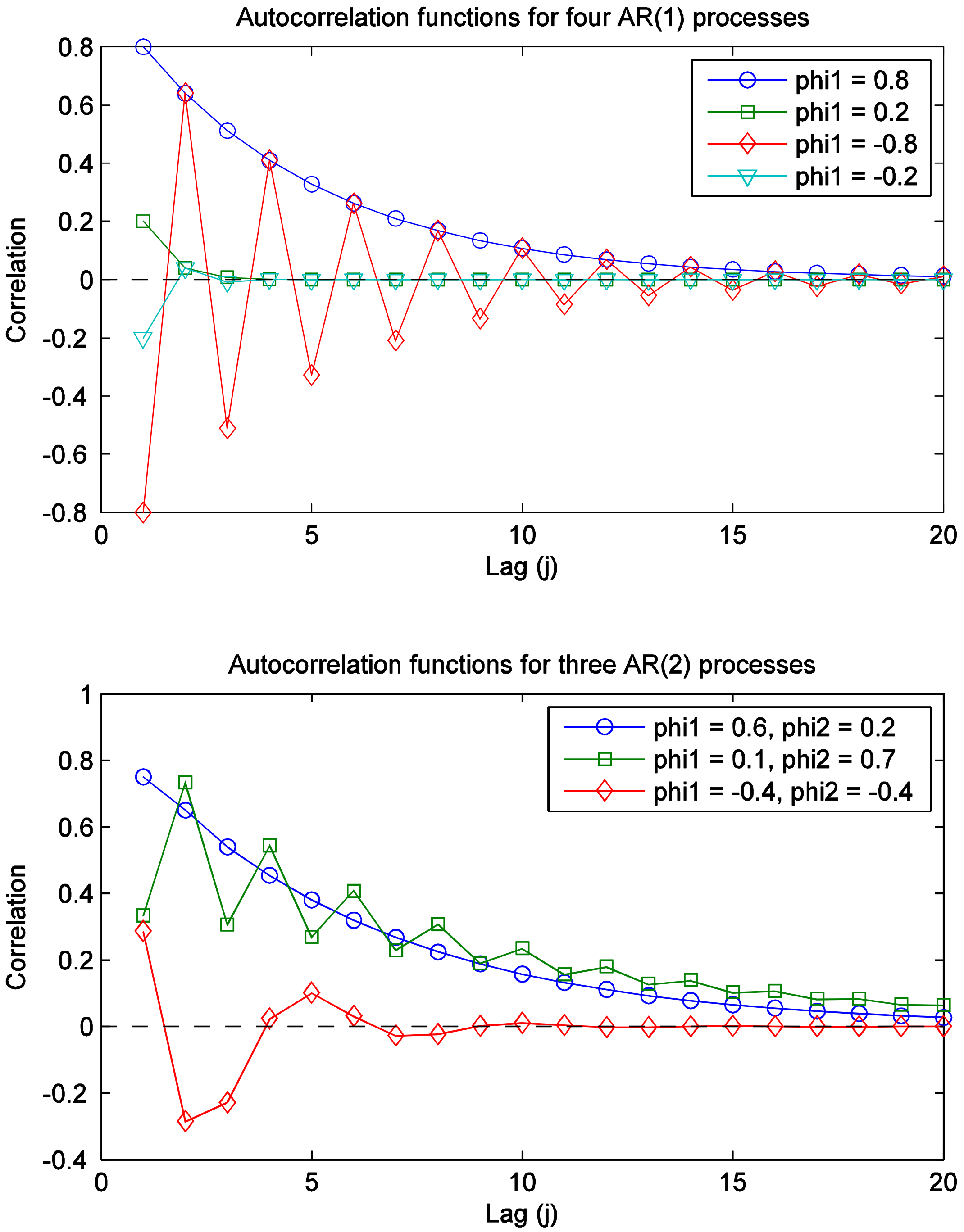


Figure 4.1: Autocorrelation functions for AR(1) and AR(2) processes.

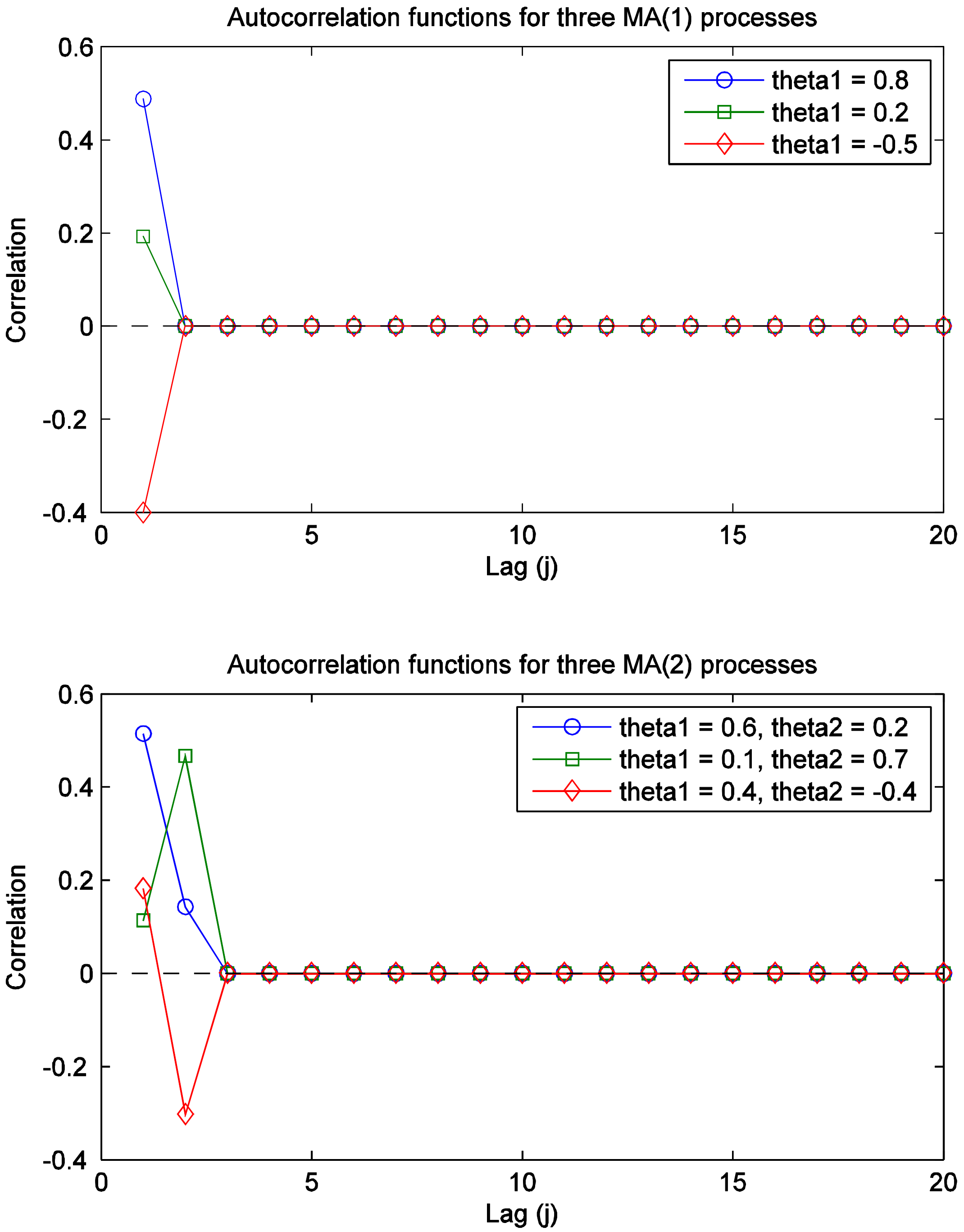


Figure 4.2: Autocorrelation functions for MA(1) and MA(2) processes.

4.4 Predictability, R^2 and ARMA processes

Recall that R^2 (R-squared) is one measure of the 'success' of a regression, measuring the proportion of variation in the dependent variable explained by the model.

$$R^2 = 1 - \frac{V[\varepsilon_{t+1}]}{V[Y_{t+1}]}$$

If the variance of the residual is exactly equal to the variance of the original variable then $R^2 = 0$, and we would conclude that the model is not very good. If the residual has very low variance, then R^2 will be close to 1, and we conclude that we have a good model. We can also use R^2 to measure the degree of (mean) predictability in a time series process.

Let's consider an AR(1) as an example:

$$Y_{t+1} = \phi_0 + \phi_1 Y_t + \varepsilon_{t+1},\ \varepsilon_{t+1} \sim WN\left(0, \sigma^2\right)$$

The variance of the residual, $V[\varepsilon_{t+1}]$, is given as σ^2. Above we determined that $V[Y_{t+1}] = \sigma^2/(1-\phi^2)$, so

$$\begin{aligned} R^2 &= 1 - \frac{V[\varepsilon_{t+1}]}{V[Y_{t+1}]} \\ &= 1 - \frac{\sigma^2}{\sigma^2/(1-\phi_1^2)} \\ &= \phi_1^2 \end{aligned}$$

So the larger the autoregressive coefficient, ϕ_1, the larger the R^2, and the greater the degree of predictability in this variable.

> **Activity 4.3** Suppose that
>
> $$\begin{aligned} Y_t &= \phi Y_{t-1} + \varepsilon_t,\ \ |\phi| < 1 \\ \varepsilon_t &\sim iid\ N(0,1) \\ \text{and }\ X_t &= \theta u_{t-1} + u_t \\ u_t &\sim iid\ N\left(0, \sigma^2\right) \end{aligned}$$

1. (a) Derive the variance of the forecast error for the optimal one-step and two-step forecasts of each of Y_t and X_t.

   (b) Find the values for θ and σ^2 that make X_t and Y_t equally predictable (according to the variance of their forecast errors) for one-step and two-step forecasts.

   (c) Given these values, which variable is easier to predict **three** steps ahead?

4.5 Choosing the best ARMA model

When fitting an ARMA(p, q) model to data, an important step is choosing the AR and MA orders (the p and q). We do not want to choose orders too low, as we might then miss some valuable information and reduce the accuracy of the forecast, but we also do not want to select orders that are too high, as the estimated parameters become less accurate the more parameters we have to estimate. Less precisely estimated parameters leads to less accurate forecasts. Thus choosing p and q involves a trade-off between *estimation error*and *goodness-of-fit.*

The most common measure of goodness-of-fit is the mean squared error, or MSE:

$$MSE = \frac{1}{T}\sum_{t=1}^{T} e_t^2$$

where e_t is the residual from the ARMA model. A lower MSE means that the errors are generally smaller, which means that the model is providing a better fit. Note that the MSE can never increase when you increase the order of an ARMA model – the worst that can happen is that the MSE stays the same. However the (potential) improvement in MSE may not come from the model being good; it may come from 'in-sample overfitting.' Over-fitting occurs when a researcher adds variables to a model that appear to be good, because they increase the R^2 or lower the MSE, but are not really useful for forecasting. Thus, the MSE goodness-of-fit measure is not useful for helping us to find a good model for forecasting, as it ignores the impact of estimation error on forecast accuracy.

Let us now consider a few alternative measures for choosing a model for forecasting. Recall that the sample variance is usually defined as the sum of squared errors divided by $(T - 1)$ to account for the fact that the sample mean is estimated. An analogue to MSE that does reflect the number of parameters estimated is s^2 :

$$\begin{aligned} s^2 &= \frac{1}{T-k}\sum_{t=1}^{n} e_t^2 \\ &= \frac{T}{T-k} \cdot MSE \end{aligned}$$

where k is the number of parameters in the regression.

It turns out that a number of other interesting goodness-of-fit measures can also be written as a function of the sample size and the number of parameters, multiplied by the MSE. The Akaike Information Criterion (AIC), Hannan-Quinn Information Criteron (HQIC) and Schwarz's Bayesian Information Criterion (known as either BIC or SIC) are:

$$\begin{aligned} AIC &= \exp\left\{\frac{2k}{T}\right\} \cdot MSE \\ HQIC &= \{\log(T)\}^{2k/T} \cdot MSE \\ BIC &= \sqrt{T}^{2k/T} \cdot MSE \end{aligned}$$

To select the best model from a given set of models, we estimate all of them and then choose the model that minimises our selection criterion: MSE, s^2, AIC, HQIC, or BIC.

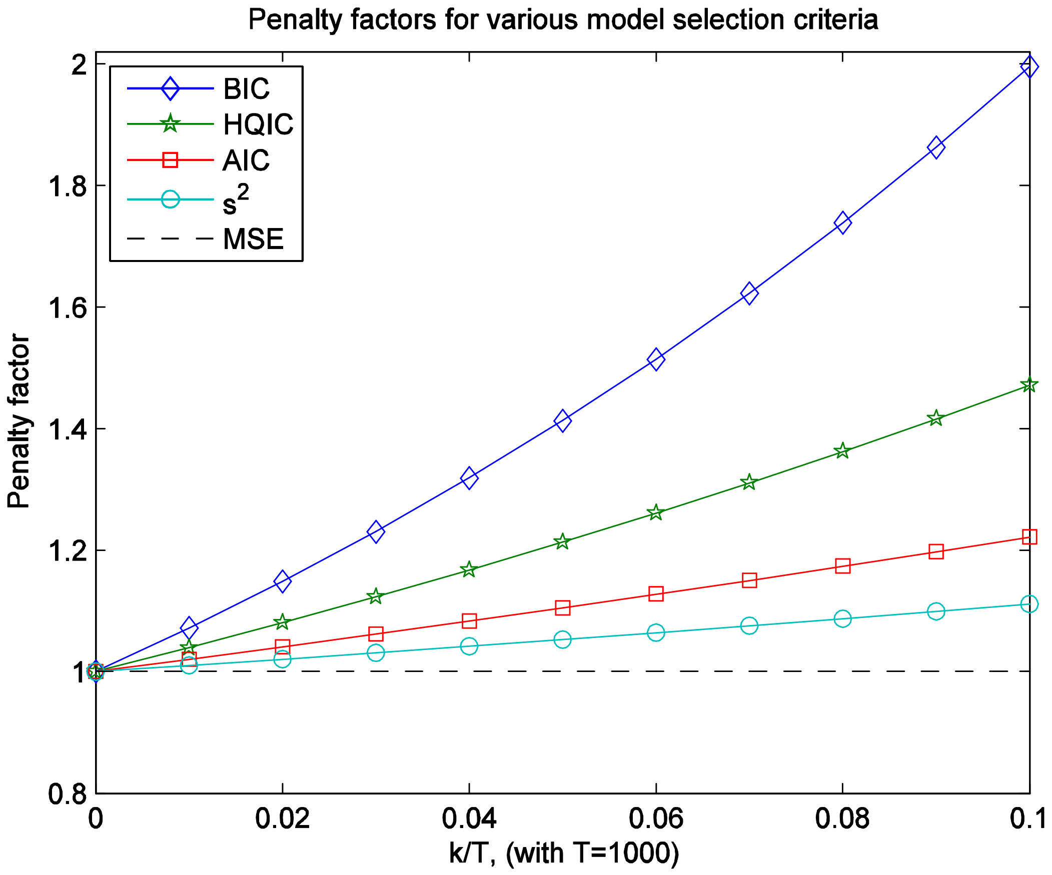


Figure 4.3: Penalty applied by various model selection criteria for the addition of an extra parameter. MSE applies no penalty, so this function is always equal to 1. Here $T = 1000$.

To see how these four measures compare, we can plot the penalty term that each of them applies for adding an extra regressor, see Figure 4.3. A penalty factor of 1 implies no penalty at all, but a penalty factor greater than 1 implies some penalty. Obviously, the MSE has a penalty factor of 1 for all k.

Figure 4.3 shows the proportion by which the MSE must decrease before the information criterion will report that the larger model is an improvement. The MSE itself simply requires that the larger model decreases the MSE by some (possibly tiny) amount. The s^2 measure requires a bit more improvement, the AIC and HQIC more still and the BIC is the strictest measure. As such, when selecting between different models, the MSE will always choose the largest, while the model that the BIC selects will generally be smaller than the model selected by the HQIC, AIC and s^2.

The AIC and BIC are the two most widely-used model selection criteria, and there are various reasons why AIC or BIC is better than the other. Most software packages usually report both, and leave it to the researcher to decide which measure to use. The BIC will pick smaller models, which is generally a good thing for forecasting, while the AIC will tend to pick larger models.

4.6 Overview of chapter

This chapter introduced some fundamental topics in time series analysis, such as autocorrelation, white noise processes and ARMA procresses. We reviewed rules for computing means, variances and covariances, and combining those rules in conjuction with implications of covariance stationarity and the law of iterated expectations, we derived some theoretical results for AR and MA processes.

4.7 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Compute the means and variances of some standard ARMA processes
- Derive autocorrelations for some simple ARMA processes
- Describe the various forms of 'white noise' processes used in the analysis of financial data
- Use the 'law of iterated expectations' to derive unconditional means from conditional means

4.8 Test your knowledge and understanding

1. Let Y_t be the following time series:

$$\begin{aligned} Y_t &= \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \\ \varepsilon_t &\sim WN\left(0, \sigma^2\right) \end{aligned}$$

   (a) Find $E\left[Y_t\right]$

   (b) Find $V\left[Y_t\right]$

   (c) Find R^2 as a function of the parameters of the process for Y_t

   (d) Find $E_t\left[Y_{t+1}\right]$

   (e) Find $E_t\left[Y_{t+3}\right]$

2. Let

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \theta \varepsilon_{t-1} + \varepsilon_t, \\ \varepsilon_t &\sim WN\left(0, \sigma^2\right) \end{aligned}$$

   (a) What type of process does Y_t follow?

   (b) Find the conditional mean of Y_t given all information available at time $t-1$.

   (c) Find the conditional mean of Y_t given all information available at time $t-2$.

   (d) Find the conditional *variance* of Y_t given all information available at time $t-1$.

3. Don't forget to check the VLE for additional practice problems for this chapter.

4.9 Solutions to activities

Activity 4.1

Let Y_t be defined as below. Find its mean, variance, first autocovariance and second autocovariance.

$$Y_t = \phi_0 + \varepsilon_t + \theta\varepsilon_{t-1},\ \varepsilon_t \sim WN\left(0, \sigma^2\right)$$

$$E\left[Y_t\right] = E\left[\phi_0 + \varepsilon_t + \theta\varepsilon_{t-1}\right] = \phi_0$$

$$\begin{aligned}
\gamma_0 &= V\left[Y_t\right] \\
&= V\left[\phi_0 + \varepsilon_t + \theta\varepsilon_{t-1}\right] \\
&= V\left[\varepsilon_t\right] + V\left[\theta\varepsilon_{t-1}\right] + 2Cov\left[\varepsilon_t, \theta\varepsilon_{t-1}\right] \\
&= \sigma^2 + \theta^2\sigma^2 + 0 \\
&= \sigma^2\left(1 + \theta^2\right)
\end{aligned}$$

$$\begin{aligned}
\gamma_1 &= Cov\left[Y_t, Y_{t-1}\right] \\
&= E\left[\left(Y_t - \phi_0\right)\left(Y_{t-1} - \phi_0\right)\right] \\
&= E\left[\left(\varepsilon_t + \theta\varepsilon_{t-1}\right)\left(\varepsilon_{t-1} + \theta\varepsilon_{t-2}\right)\right] \\
&= E\left[\varepsilon_t\varepsilon_{t-1} + \theta\varepsilon_{t-1}^2 + \theta\varepsilon_t\varepsilon_{t-2} + \theta^2\varepsilon_{t-1}\varepsilon_{t-2}\right] \\
&= E\left[\theta\varepsilon_{t-1}^2\right] \\
&= \theta\sigma^2
\end{aligned}$$

$$\begin{aligned}
\gamma_2 &= Cov\left[Y_t, Y_{t-2}\right] \\
&= E\left[\left(\varepsilon_t + \theta\varepsilon_{t-1}\right)\left(\varepsilon_{t-2} + \theta\varepsilon_{t-3}\right)\right] \\
&= E\left[\varepsilon_t\varepsilon_{t-2} + \theta\varepsilon_{t-1}\varepsilon_{t-2} + \theta\varepsilon_t\varepsilon_{t-3} + \theta^2\varepsilon_{t-1}\varepsilon_{t-3}\right] \\
&= 0
\end{aligned}$$

Activity 4.2

Let Y_t be defined as below. Find its ACF.

$$Y_t = \phi_0 + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2},\ \varepsilon_t \sim WN\left(0, \sigma^2\right)$$

$$E\left[Y_t\right] = E\left[\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}\right] = \phi_0$$

$$\begin{aligned}
\gamma_0 &= V\left[Y_t\right] \\
&= V\left[\phi_0 + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}\right] \\
&= V\left[\varepsilon_t\right] + V\left[\theta_1\varepsilon_{t-1}\right] + V\left[\theta_2\varepsilon_{t-2}\right] \\
&= \sigma^2 + \theta_1^2\sigma^2 + \theta_2^2\sigma^2 \\
&= \sigma^2\left(1 + \theta_1^2 + \theta_2^2\right)
\end{aligned}$$

$$\begin{aligned}
\gamma_1 &= Cov\left[Y_t, Y_{t-1}\right] \\
&= E\left[\left(Y_t - \phi_0\right)\left(Y_{t-1} - \phi_0\right)\right] \\
&= E\left[\left(\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}\right)\left(\varepsilon_{t-1} + \theta_1\varepsilon_{t-2} + \theta_2\varepsilon_{t-3}\right)\right] \\
&= E\left[\varepsilon_t\varepsilon_{t-1} + \theta_1\varepsilon_{t-1}^2 + \theta_2\varepsilon_{t-2}\varepsilon_{t-1}\right] + \\
&\quad E\left[\theta_1\varepsilon_t\varepsilon_{t-2} + \theta_1^2\varepsilon_{t-1}\varepsilon_{t-2} + \theta_1\theta_2\varepsilon_{t-2}^2\right] \\
&\quad + E\left[\theta_2\varepsilon_t\varepsilon_{t-3} + \theta_1\theta_2\varepsilon_{t-1}\varepsilon_{t-3} + \theta_2^2\varepsilon_{t-2}\varepsilon_{t-3}\right] \\
&= \theta_1 E\left[\varepsilon_{t-1}^2\right] + \theta_1\theta_2 E\left[\varepsilon_{t-2}^2\right] \\
&= \theta_1\sigma^2 + \theta_1\theta_2\sigma^2 \\
&= \sigma^2\theta_1\left(1 + \theta_2\right)
\end{aligned}$$

$$\begin{aligned}
\gamma_2 &= Cov\left[Y_t, Y_{t-2}\right] \\
&= E\left[\left(Y_t - \phi_0\right)\left(Y_{t-2} - \phi_0\right)\right] \\
&= E\left[\left(\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}\right)\left(\varepsilon_{t-2} + \theta_1\varepsilon_{t-3} + \theta_2\varepsilon_{t-4}\right)\right] \\
&= E\left[\varepsilon_t\varepsilon_{t-2} + \theta_1\varepsilon_{t-1}\varepsilon_{t-2} + \theta_2\varepsilon_{t-2}^2\right] + \\
&\quad E\left[\theta_1\varepsilon_t\varepsilon_{t-3} + \theta_1^2\varepsilon_{t-1}\varepsilon_{t-3} + \theta_1\theta_2\varepsilon_{t-2}\varepsilon_{t-3}\right] \\
&\quad + E\left[\theta_2\varepsilon_t\varepsilon_{t-4} + \theta_1\theta_2\varepsilon_{t-1}\varepsilon_{t-4} + \theta_2^2\varepsilon_{t-2}\varepsilon_{t-4}\right] \\
&= \theta_2 E\left[\varepsilon_{t-2}^2\right] \\
&= \theta_2\sigma^2 \\
\gamma_j &= 0 \text{ for } j \geq 3
\end{aligned}$$

Activity 4.3

(a) Derive the variance of the forecast error for the optimal one-step and two-step forecasts of each of Y_t and X_t.

⇒

$$\begin{aligned}
\hat{Y}_{t+1,t} &= E_t\left[Y_{t+1}\right] = \phi Y_t \\
\text{so } V\left[e_{t+1,t}^y\right] &= 1 \\
\hat{Y}_{t+2,t} &= E_t\left[Y_{t+2}\right] = \phi^2 Y_t \\
\text{since } Y_{t+2} &= \phi Y_{t+1} + \varepsilon_{t+2} \\
&= \phi\left(\phi Y_t + \varepsilon_{t+1}\right) + \varepsilon_{t+2} \\
&= \phi^2 Y_t + \phi\varepsilon_{t+1} + \varepsilon_{t+2} \\
\text{so } V\left[e_{t+2,t}^y\right] &= 1 + \phi^2 \\
\hat{X}_{t+1,t} &= E_t\left[X_{t+1}\right] = \theta u_t \\
\text{so } V\left[e_{t+1,t}^x\right] &= \sigma^2 \\
\hat{X}_{t+2,t} &= E_t\left[X_{t+2}\right] = 0 \\
\text{so } V\left[e_{t+2,t}^x\right] &= \sigma^2\left(1 + \theta^2\right)
\end{aligned}$$

(b) Find the values for θ and σ^2 that make X_t and Y_t equally predictable (according to the variance of their forecast errors) for one-step and two-step forecasts.

⇒

$$
\begin{aligned}
V\left[e^y_{t+1,t}\right] &= V\left[e^x_{t+1,t}\right] \Rightarrow \sigma^2 = 1 \\
V\left[e^y_{t+2,t}\right] &= V\left[e^x_{t+2,t}\right] \Rightarrow 1+\phi^2 = 1+\theta^2 \Rightarrow \theta = \pm\phi
\end{aligned}
$$

(c) Given these values, which variable is easier to predict **three** steps ahead?

⇒

$$
\begin{aligned}
\hat{Y}_{t+3,t} &= E_t\left[Y_{t+3}\right] = \phi^3 Y_t \\
\text{so } V\left[e^y_{t+3,t}\right] &= 1+\phi^2+\phi^4 \\
\hat{X}_{t+3,t} &= E_t\left[X_{t+3}\right] = 0 \\
\text{so } V\left[e^x_{t+3,t}\right] &= \sigma^2\left(1+\theta^2\right) \\
&= 1+\phi^2, \text{ if } \sigma^2 = 1 \text{ and } \theta = \pm\phi \\
&\leq 1+\phi^2+\phi^4
\end{aligned}
$$

so if $\phi \neq 0$, then X_t is easier to predict three steps ahead than Y_t, in terms of error variance. In fact, notice that X_t is in fact **not predictable at all** three steps ahead (the forecast is always equal to zero, its unconditional mean), while Y_t is at least partially predictable three steps ahead. Thus if we measured predictability by percentage of variation explained by the forecast, (similar to R^2) rather than by error variance, Y_t would be easier to predict three steps ahead than X_t.

Chapter 5
Empirical features of financial asset returns

5.1 Introduction

This chapter will start to build your familiarity with working with real financial data. We will discuss summary statistics that are routinely presented in empirical work prior to more detailed or complicated analyses, and we will discuss the Jarque-Bera test for normality, which is based on these summary statistics.

5.1.1 Aims of the chapter

The aims of this chapter are to:

- Present common summary statistics for financial asset returns
- Introduce the Jarque-Bera test for normality
- Increase familiarity with financial data by illustrating all of the main ideas with three financial time series.

5.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Compute the sample moments and sample quantiles that make up a table of summary statistics
- Compute the Jarque-Bera test statistic and determine whether the null hypothesis of normality is rejected or not

5.1.3 Essential reading

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487], Chapter 3 Sections 3–4.

5.1.4 Further reading

- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354], Chapter 1.

- Taylor, Stephen J. *Asset Price Dynamics, Volatility and Prediction.* (Princeton University Press, Oxford, 2005) [ISBN 9780691134796], Chapter 4.

5.1.5 References cited

- Jarque, C. M. and A. K. Bera, 'A test for normality of observations and regression residuals,' *International Statistical Review*, 1987, 55(2), pp.163–172.

5.2 Summary statistics

Before undertaking any detailed analysis, it is good empirical practice to gather 'summary statistics' for the data to be studied. We will review a few standard summary statistics here. Recall from Chapter 2 that we almost always study *returns* not prices, and so before doing anything further, we transform our price series into a return series, and we scale it up by 100 so that the returns are in percentages.

$$R_t = 100 \times (\log P_t - \log P_{t-1})$$

5.2.1 Sample moments

It is common to report the sample mean, standard deviation, skewness and kurtosis of the returns to be studied. These four quantities are usually referred to as the first four sample moments, even though strictly only the mean is a moment; the others are transformations of moments.

$$\begin{aligned}
\bar{R} &= \frac{1}{T}\sum_{t=1}^{T} R_t, \\
\hat{\sigma} &= \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(R_t - \bar{R}\right)^2} \\
\hat{S} &= \frac{1}{\hat{\sigma}^3}\frac{1}{T}\sum_{t=1}^{T}\left(R_t - \bar{R}\right)^3 \\
\hat{K} &= \frac{1}{\hat{\sigma}^4}\frac{1}{T}\sum_{t=1}^{T}\left(R_t - \bar{R}\right)^4
\end{aligned}$$

When the sampling frequency (how often a price is observed) is less than annual, it is common to 'annualise' the mean and the standard deviation. This is done so that returns computed over different frequencies (e.g., daily and monthly) can be more easily compared. The idea behind 'annualisation' is to infer the mean and standard deviation that would be obtained for annual sampling, and the most common way to annualise is to do so assuming that returns are *serially uncorrelated*. This is not true (as we will see below) but this assumption provides a quick approximation and is widely-used anyway. Under this assumption, average returns are scaled by the number of sampling periods within a year (it is generally assumed that there are 252 trade days per year, and 22 trade days per month). The standard deviation is scaled by the *square-root* of the number of sampling periods.

$$\begin{aligned} \bar{R}_{ann} &= h \times \bar{R} \\ \hat{\sigma}_{ann} &= \sqrt{h} \times \hat{\sigma} \end{aligned}$$

Let's now see where these scaling factors come from. Do to so, we will assume that returns satisfy

$$Cov\left[R_t, R_{t-j}\right] = 0 \ \forall \ j \neq 0$$

Then define *aggregated returns* (eg, annual returns) as:

$$X_t = \sum_{j=0}^{h-1} R_{t-j}$$

Let's now derive the mean of X_t:

$$E\left[X_t\right] = E\left[\sum_{j=0}^{h-1} R_{t-j}\right] = \sum_{j=0}^{h-1} E\left[R_{t-j}\right] = \sum_{j=0}^{h-1} \mu = h\mu$$

Thus the mean of X_t is just h times the mean of R_t. This motivates the scaling factor for the mean. Next we derive the variance of X_t:

$$\begin{aligned} V\left[X_t\right] &= V\left[\sum_{j=0}^{h-1} R_{t-j}\right] \\ &= \sum_{j=0}^{h-1} V\left[R_{t-j}\right] + 2\sum_{j=0}^{h-2}\sum_{k=j+1}^{h-1} Cov\left[R_{t-j}, R_{t-k}\right] \\ &= \sum_{j=0}^{h-1} V\left[R_{t-j}\right], \ \text{since } Cov\left[R_t, R_{t-j}\right] = 0 \ \forall \ j \neq 0 \\ &= \sum_{j=0}^{h-1} \sigma^2 = h\sigma^2 \\ \text{so } \sqrt{V\left[X_t\right]} &= \sqrt{h\sigma^2} = \sqrt{h}\sigma \end{aligned}$$

which motivates the scaling factor for the standard deviation.

> **Activity 5.1** Using the mean and standard deviation scaling formulas above to fill in the table below. (Assume that there are 22 trading days in a month, and 252 in a year.)

Return summary statistics

	Daily	*Weekly*	*Monthly*	*Annual*
Mean	0.0311			
Std dev	1.1483			

Activity 5.2 The following table contains the level of the FTSE 100 index on the last trading day of each year from 2001 to 2013. Convert these prices to continuously compounded percentage returns and compute the first four sample moments.

End of year	P_t	R_t
2001	5242.4	–
2002	3900.6	
2003	4470.4	
2004	4820.1	
2005	5638.3	
2006	6241.0	
2007	6476.9	
2008	4392.7	
2009	5397.9	
2010	5971.0	
2011	5566.8	
2012	5925.4	
2013	6731.3	

Mean	
Std dev	
Skewness	
Kurtosis	

5.2.2 Sample quantiles

A different way to summarise the sample distribution of a series of returns is through 'quantiles' (also known as 'percentiles'). An example of a quantile is the median (which is the 0.5 quantile), which is the number such that 50% of the observations lie below it. More generally a population quantile is defined implicitly as

$$\Pr\left[R \leq Q_\alpha\right] = \alpha, \text{ for some } \alpha \in (0,1).$$

Common choices of α are 0.05 or 0.1, 0.25 (the first 'quartile'), 0.5 (the median), 0.75 (the third quartile) and 0.9 or 0.95. (We will see in Chapter 10 that quantiles are related to a measure of risk known as 'Value-at-Risk.')

If the *cdf* is continuous, then we can use the fact that $\Pr\left[R \leq Q_\alpha\right] \equiv F\left(Q_\alpha\right)$ to express the population quantile as the *inverse cdf:*

$$Q_\alpha = F^{-1}\left(\alpha\right).$$

We obtain *sample* quantiles by ordering our sample of data from smallest to largest:

$$R_{(1)} \leq R_{(2)} \leq \cdots \leq R_{(T-1)} \leq R_{(T)}$$

where $R_{(1)}$ is the smallest return in our sample, $R_{(2)}$ is the second-smallest, and $R_{(T)}$ is the largest return. Then the α-quantile is equal to

$$\hat{Q}_\alpha = R_{(\lfloor \alpha T \rfloor)}$$

where $\lfloor \alpha T \rfloor$ rounds αT to the nearest integer. For example, if $T = 90$ and $\alpha = 0.1$, then $\alpha T = 9$, and we would report the 9^{th} smallest observation as our sample 0.10 quantile. If $\alpha = 0.25$ then $\alpha T = 22.5$, and so we round and report the 23^{rd} smallest observation as our sample 0.25 quantile. A convention is that if $\alpha < 1/T$ then $\hat{Q}_\alpha$ is just the smallest observed return (i.e., the minimum) and similarly if $\alpha > (T - 1/2)/T$ then $\hat{Q}_\alpha$ is the largest observed return.

> **Activity 5.3** Using the returns you computed in Activity 4.3, compute the sample quantiles for the table below
>
> | α | $\hat{Q}_\alpha$ |
> |---|---|
> | 0.1 | |
> | 0.33 | |
> | 0.5 | |
> | 1 | |

5.3 Jarque-Bera test for normality

The Jarque-Bera (JB) test examines whether a given sample of data is normally distributed. Recall that all normally distributed random variables, regardless of their mean and variance, have skewness equal to zero and kurtosis of three. The JB normality test is based on how far the sample skewness and kurtosis are from zero and three respectively. The JB test statistic is simple to compute:

$$JB = \frac{T}{6}\left(\hat{S}^2 + \frac{1}{4}\left(\hat{K} - 3\right)^2\right).$$

Under the null hypothesis that the data come from a Normal distribution the JB test statistic has the χ^2_2 distribution, and so the 95% critical value for the JB test statistic is 5.99. If the JB test statistic is greater than 5.99 then we reject the null and conclude that the data do not come from a normal distribution.

> **Activity 5.4** Compute the JB test statistic for the sample of returns you computed in Activity 4.3, and determine whether this suggests that we should reject normality or not.

5.4 A detailed look at some financial data

Now we will apply the tools we have learned to a few daily financial time series. The series we will consider are:

Name	**Sample period**	T
Euro/USD exchange rate	Jan 4, 1999 - Dec 31, 2009	2767
S&P 500 index	Jan 3, 1980 - Dec 31, 2009	7570
US 3-month T-bill rate	Jan 3, 1989 - Dec 31, 2009	5982

We will examine the continuously compounded returns of these variables. For the exchange rate and the stock index this is simply:

$$R_{t+1} = 100 \times (\log P_{t+1} - \log P_t) \equiv 100 \times \Delta \log P_{t+1}$$

The continuously compounded return on an asset is sometimes called the 'log-difference' of the price series. For stock returns the natural variable to look at is indeed just the log-difference, as this represents the continuously compounded return on the asset over the period. For exchange rates the log-difference represents the return one would receive if one had bought the foreign currency yesterday, stuck it under the bed, and then exchanged it back to the domestic currency today. In reality, of course, people don't put money under the bed, rather they would invest it in the foreign overnight risk-free asset, so an economically more interesting variable might be the sum of the log-difference in the exchange rate and the foreign overnight risk-free rate. Failing to include the foreign overnight risk-free rate may change the results in certain circumstances, but it is not uncommon to ignore (as we will) the foreign overnight risk-free rate.

For the T-bill we first convert the rate into a price, using the formula:

$$P_t = \frac{100}{(1 + R_t)^{0.25}}$$

and then study the log-difference of the price series, which represents the continuously-compounded return from holding a 3-month T-bill for one day. Plots of these three returns series are given in Figure 5.1.

Table 5.1 presents summary statistics for the three return series introduced above. The first four sample moments reveal wide differences in the volatility of each series, ranging from 18% annualised for the S&P 500 index returns to just 0.2% annualised for the 3-month T-bill returns. We also see that the stock index returns have negative skewness, while the other two series have positive skewness. All three series have excess kurtosis (they are 'leptokurtotic').

The second panel of Table 5.1 below reports quantiles of the sample distribution. Specifically, I report the minimum value, the 5% quantile, the median (the 50% quantile), the 95% quantile and the maximum. These statistics provide an alternative view of the distribution. By reporting the minimum and maximum values we are also able to see whether there are data errors or outliers that need to be examined more closely.

In the bottom panel I report the Jarque-Bera test statistic and associated p-value for the test of Normality. On daily asset returns we almost invariably see strong rejections of Normality, and this is confirmed for the three series considered here.

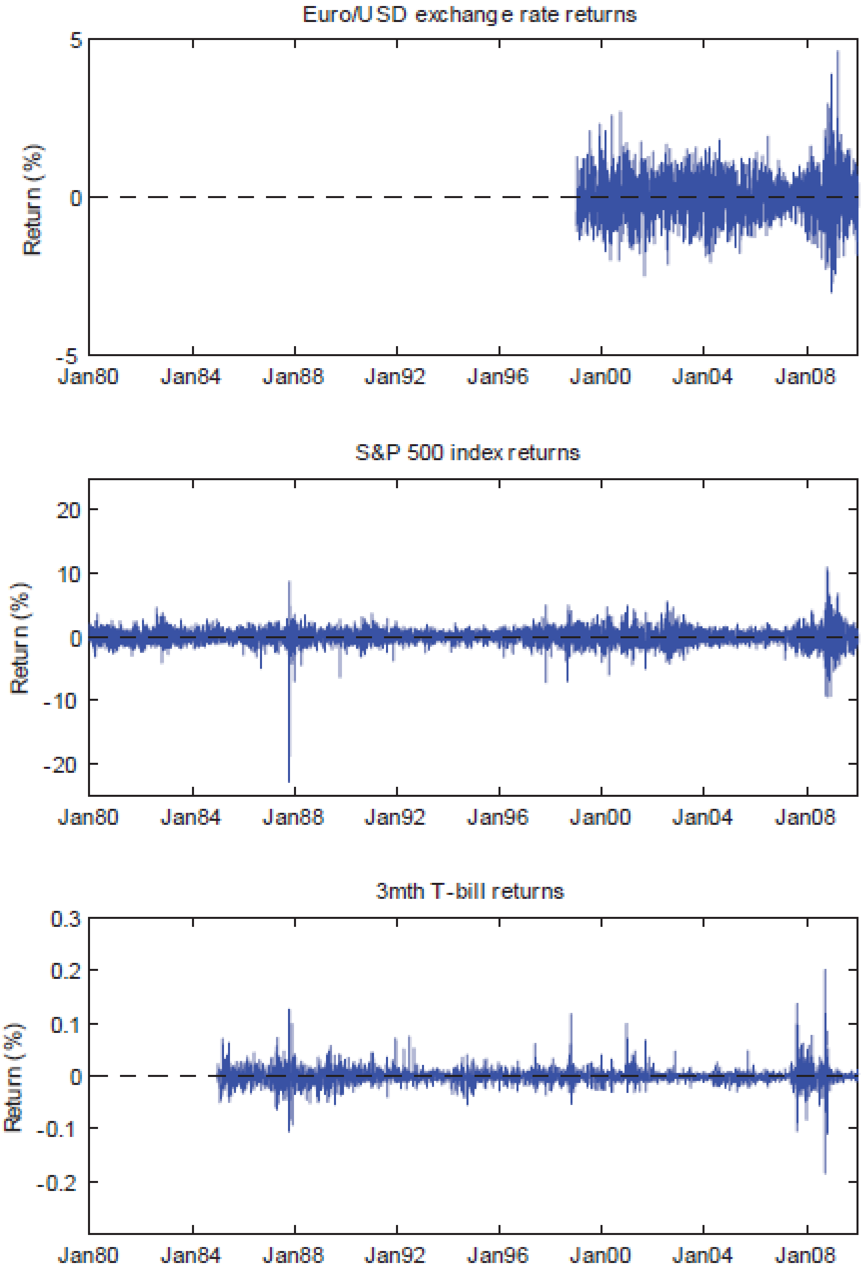


Figure 5.1: Time series of daily returns (in per cent) for the euro/US dollar exchange rate (Jan 99-Dec 2009), S&P 500 index (Jan 1980 - Dec 2009) and the 3-month T-bill (Jan 1985 - Dec 2009).

Table 5.1: Summary statistics

	Ex rate	*Stock index*	*T-bill*
Mean	0.0071	0.0311	0.0005
Std dev	0.6503	1.1483	0.0140
Annualised mean	1.8012	7.8414	0.1159
Annualised std dev	10.3223	18.2295	0.2216
Skewness	0.1714	-1.2471	0.8178
Kurtosis	5.5155	31.8615	28.8995
Minimum	-3.0031	-22.8997	-0.1839
5% quantile	-1.0501	-1.6843	-0.0188
Median	0.0000	0.0533	0.0000
95% quantile	1.0631	1.6744	0.0192
Maximum	4.6208	10.9572	0.2016
JB statistic	740.8	264546.2	167735.1
JB p-value	0.0000	0.0000	0.0000

5.5 Overview of chapter

This chapter introduced some common summary statistics used when working with financial asset returns, and presented the Jarque-Bera test for normality. We computed these statistics for some samples of time series of financial asset returns.

5.6 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Compute the sample moments and sample quantiles that make up a table of summary statistics
- Compute the Jarque-Bera test statistic and determine whether the null hypothesis of normality is rejected or not

5.7 Test your knowledge and understanding

1. Show that the sample skewness and sample kurtosis ($\hat{S}$ and $\hat{K}$) of a return series are not affected when the returns are scaled by some positive constant.

2. The table below presents some summary statistics and Jarque-Bera tests for

returns on the FTSE 100 index across four different sampling frequencies.

	Daily	*Weekly*	*Monthly*	*Annually*
Mean	0.01	0.05	0.16	2.08
Std dev	1.24	2.55	4.70	17.59
Skewness	-0.13	-1.36	-1.90	-1.35
Kurtosis	10.17	17.08	9.70	3.48
JB test stat	6724.67	5364.23	351.06	3.76

(a) Recalling that the 95% critical value for the χ_2^2 distribution is 5.99, interpret these test results.

(b) Comment on how accurate the scaling formula $\mu_h = h \times \mu$ is for the sample means at different frequencies.

(c) Comment on how accurate the scaling formula $\sigma_h = \sqrt{h} \times \sigma$ is for the sample standard deviations at different frequencies.

3. Don't forget to check the VLE for additional practice problems for this chapter.

5.8 Solutions to activities

Activity 5.1

Return summary statistics

	Daily	*Weekly*	*Monthly*	*Annually*
Mean	0.0311	0.1555	0.6842	7.7750
Std dev	1.1483	2.5677	5.3860	18.1562

Activity 5.2

End of year	P$_t$	R_t
2001	5242.4	–
2002	3900.6	-29.6
2003	4470.4	13.6
2004	4820.1	7.5
2005	5638.3	15.7
2006	6241.0	10.2
2007	6476.9	3.7
2008	4392.7	-38.8
2009	5397.9	20.6
2010	5971.0	10.1
2011	5566.8	-7.0
2012	5925.4	6.2
2013	6731.3	12.8

Mean	2.08
Std dev	17.59
Skewness	-1.35
Kurtosis	3.48

Activity 5.3

α	$\hat{Q}_\alpha$
0.1	-38.8
0.33	3.7
0.5	7.5
1	20.6

Activity 5.4

From Activity 4.2 the sample skewness is -1.3 and the sample kurtosis is 3.5. Thus the JB test statistic is

$$\begin{aligned} JB &= \frac{T}{6}\left(\hat{S}^2 + \frac{1}{4}\left(\hat{K}-3\right)^2\right) \\ &= \frac{12}{6}\left((-1.3)^2 + \frac{1}{4}(3.5-3)^2\right) \\ &= 3.76 \end{aligned}$$

The 95% critical value for the JB test comes from the χ^2_2 distribution and is 5.99. The test statistic is *less than* the critical value, and so we *fail to reject* the null of normality. Thus we conclude that the annual returns on the FTSE 100 index over the last 12 years are not significantly non-normal. (This is somewhat common for annual returns; for monthly and daily returns we usually find strong evidence against normality.)

Chapter 6
Testing for predictability in financial time series

6.1 Introduction

This chapter will continue to build your familiarity with working with real financial data. We will consider sample autocorrelations and hypothesis tests based on these to test for the presence of predictability in a financial asset return series.

6.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce sample autocorrelations
- Discuss testing for predictability based on single or multiple autocorrelations
- Increase familiarity with financial data by illustrating all of the main ideas with three financial time series.

6.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Compute sample autocorrelations
- Interpret plots of sample autocorrelations and associated confidence intervals, and interpret outputs of tests for autocorrelations

6.1.3 Essential reading

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487], Chapter 3 Sections 3–4.

6.1.4 Further reading

- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354], Chapter 1.
- Taylor, Stephen J. *Asset Price Dynamics, Volatility and Prediction.* (Princeton University Press, Oxford, 2005) [ISBN 9780691134796]. Chapter 4.

6.1.5 References cited

- Bartlett, M., 'On the theoretical specification of sampling properties of autocorrelated time series,' *Journal of the Royal Statistical Society B*, 1946, 8, pp.27–41.
- White, H., 'A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,' *Economerica*, 1980, 48(4), pp.817–838.
- Newey, W. K., and K. D. West, 'A Simple, Positive Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,' *Econometrica*, 1987, 55(3), pp.703–708.
- Ljung, G. M. and G. E. P. Box, 'On a measure of lack of fit in time-series models,' *Biometrika*, 1978, 65, pp.297–303.

6.2 Sample autocorrelations

The simplest and most widely-cited measure of time series predictability is autocorrelation, which reveals whether a time series is (linearly) predictable using its own lags. When we estimate the autocorrelation function of a time series given a data set, we would like to know which of the lags are significant and which are not, i.e., we would like to conduct hypothesis tests on the estimated autocorrelation coefficients.

Recall that the autocorrelation coefficient is the ratio of the autocovariance coefficient to the variance:

$$\rho_j = \frac{\gamma_j}{\gamma_0} = \frac{Cov\left[R_t, R_{t-j}\right]}{V\left[R_t\right]}$$

Denote the sample mean as $\bar{R} = \frac{1}{T}\sum_{t=1}^{T} R_t$, then the standard estimate of the autocorrelation coefficient for a given data set is:

$$\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0} = \frac{\frac{1}{T-j}\sum_{t=j+1}^{T}\left(R_t - \bar{R}\right)\left(R_{t-j} - \bar{R}\right)}{\frac{1}{T}\sum_{t=1}^{T}\left(R_t - \bar{R}\right)^2}$$

Note that when computing the j^{th}-order autocorrelation coefficient we must drop the first j observations from the summation in the numerator. One implication of this is that we can only compute autocorrelations up to order $(T-1)$, however, it is not a good idea to try to compute autocorrelations of very high order relative to T, because such estimates are not very precise. (The $(T-1)^{th}$-order autocorrelation is an estimate based on only 1 observation, clearly not a very precise estimate.) Some authors suggest $T^{1/3}$ as the largest number of autocorrelations to estimate.

> **Activity 6.1** Show that the population *and* sample autocorrelations are unaffected by re-scaling the data (eg, working with returns in decimals or in percentages).

> **Activity 6.2** Assume that for a given time series, we find $\hat{\rho}_1 = -0.4$. Using your results on the theoretical autocorrelations of an AR(1) and an MA(1) find the method-of-moments estimator for the parameters of those models. (That is, find the ϕ and θ for the AR(1) and MA(1), separately, that would generate $\rho = -0.4$.)

6.3 Tests on individual autocorrelation coefficients

When the data are *iid* and Normally distributed it can be shown that the asymptotic distribution of the autocorrelation estimate is:

$$\hat{\rho}_j \sim N\left(0, \frac{1}{T}\right)$$

And so the standard error on $\hat{\rho}_j$ is $1/\sqrt{T}$. This is known as 'Bartlett's' standard error for the sample autocorrelation coefficient. This distribution is useful when we want to test the following null hypothesis:

$$\begin{aligned} H_0 &: \rho_j = 0 \\ \text{vs. } H_a &: \rho_j \neq 0 \end{aligned}$$

Recall that a 95% confidence interval is a pair of numbers, $[L, U]$, such that the probability under the null hypothesis that the interval includes the true value is 0.95. Using the asymptotic distribution we can obtain a 95% confidence interval for the sample autocorrelation coefficient:

$$95\%\ CI: \left[\hat{\rho}_j - \frac{1.96}{\sqrt{T}}, \hat{\rho}_j + \frac{1.96}{\sqrt{T}}\right]$$

If this interval does *not* include zero, then we conclude that the j^{th}-order autocorrelation coefficient is significantly different from zero at the 5% level. When looking at many autocorrelations, it is convenient to re-centre these intervals on zero, and if these re-centred intervals do not include $\hat{\rho}_j$ then we similarly conclude that the j^{th}-order autocorrelation coefficient is significantly different from zero at the 5% level.

Equivalently, we could construct a t-statistic for $\hat{\rho}_j$:

$$tstat = \frac{\hat{\rho}_j}{\sqrt{V[\hat{\rho}_j]}} = \frac{\hat{\rho}_j}{1/\sqrt{T}} = \sqrt{T}\hat{\rho}_j$$

If $\hat{\rho}_j \sim N(0, 1/T)$ under the null hypothesis that $\rho_j = 0$, then $tstat = \sqrt{T}\hat{\rho}_j \sim N(0, 1)$. We conduct a t-test by comparing the t-statistic with the 95% critical values of the $N(0, 1)$ distribution, which are ±1.96. If $|tstat| > 1.96$ we conclude that we have evidence against the null hypothesis that $\rho_j = 0$ and we say that the autocorrelation coefficient is 'significantly different from zero' (or just 'significant').

> **Activity 6.3** Using the returns data you computed in Activity 5.2, compute the first three sample autocorrelations and the Bartlett t-statistics on these. Which, if any, of these autocorrelations are significant?

In Figure 6.1 I have plotted the sample autocorrelation functions (SACF) for two data sets. The first exhibits only mild serial dependence, while the second exhibits strong serial dependence. The bars are the sample autocorrelation coefficients, while the stars represent the 95% confidence intervals for the *individual* autocorrelation coefficients. The first plot exhibits only very weak evidence of serial correlation (only one significant

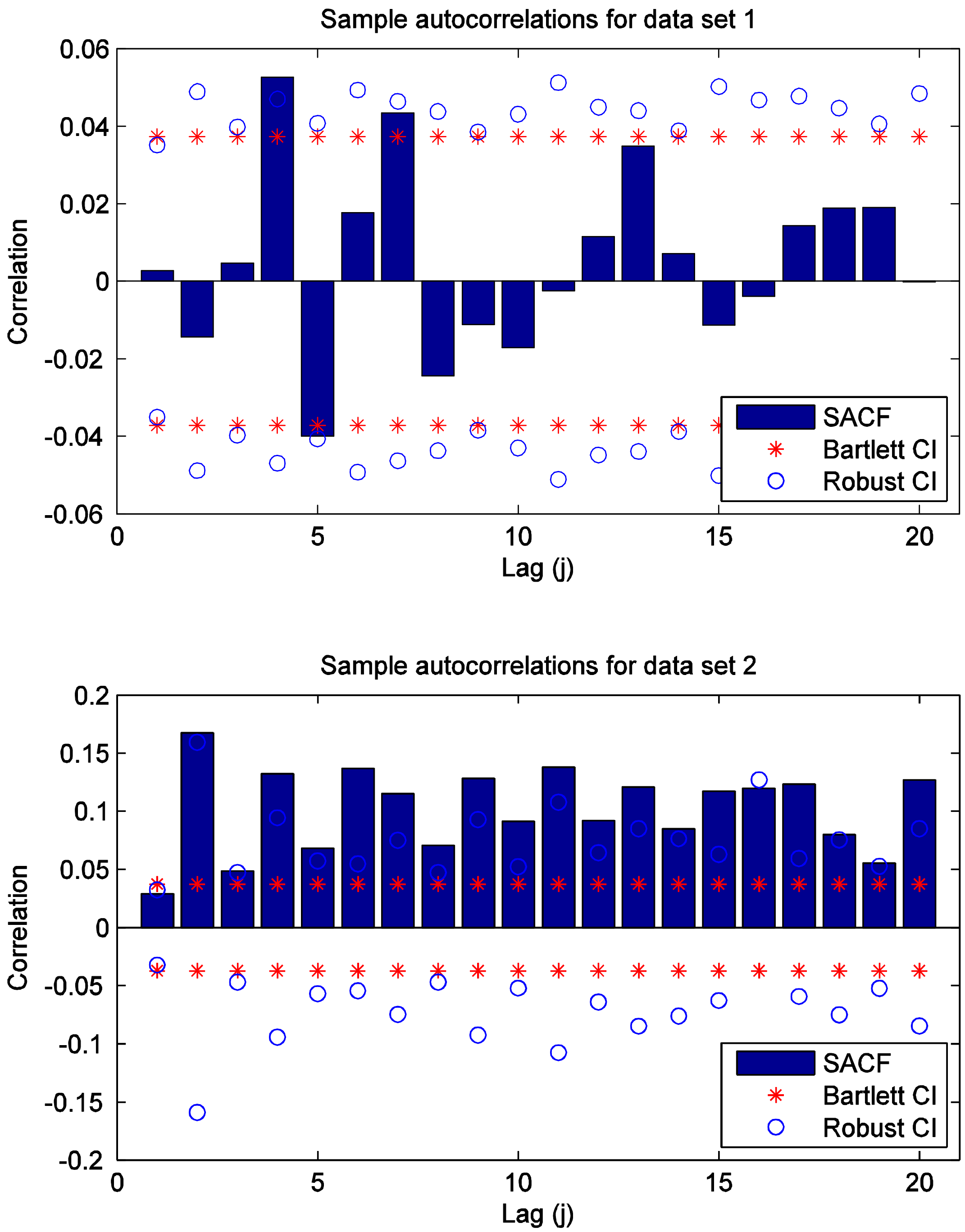


Figure 6.1: Sample autocorrelation functions for two data sets. The bars represent the estimated autocorrelation coefficients, the asterisks represent the 95% Bartlett confidence intervals, and the circles represent White's robust 95% confidence interval.

lag out of 20) whereas the second plot suggests that the series exhibits substantial serial correlation: all but 2 of the 20 lags are individually significant.

When the data under analysis exhibit heteroskedasticity (non-constant variance) then the above test is not appropriate, because Bartlett's standard errors only apply if the data are *iid*. Many asset returns exhibit conditional heteroskedasticity and so it is preferable to use a 'robust' test for autocorrelation, using the White (1980) or Newey-West (1987) estimators of standard errors. We can do this by running a regression:

$$Y_t = \beta_0 + \beta_j Y_{t-j} + e_t$$

and testing

$$\begin{aligned} H_0 &: \beta_j = 0 \\ \text{vs. } H_a &: \beta_j \neq 0 \end{aligned}$$

which is achieved via a simple t-test. Most econometric software packages offer 'robust' standard errors as an option. If robust standard errors are used to compute the t-statistic then the test is appropriate for heteroskedastic data. Note that the confidence intervals for each autocorrelation need not be the same, unlike Bartlett confidence intervals.

From Figure 6.1 we see that for these two variables the robust standard errors are not too different for the first data set (upper panel), with the robust confidence intervals being generally just slightly wider than the Bartlett confidence intervals. In the lower panel we see that the robust confidence intervals are quite a bit wider than the Bartlett confidence intervals, suggesting that this data set is further from the *iid* assumption required for the simpler confidence intervals.

6.4 Joint tests on many autocorrelations

A 5% level hypothesis test has, by construction, a 5% chance of falsely rejecting the null hypothesis, so when the null hypothesis is true there is a 1 in 20 chance that a 5% level test will suggest that the null hypothesis is false. Why should we care about this here? If we look at an SACF out to 20 lags, and observe that only the 6^{th} lag is significant according to the confidence intervals for *individual* autocorrelation coefficients, what should we conclude? Is lag 6 significant, or is it the one 'false rejection' that we would expect when conducting 20 individual tests? (This question relates to the subject of data mining, which we will discuss in the next chapter.)

One way of overcoming this concern is to conduct a *joint* test that *all* autocorrelation coefficients up to lag L are zero. One such test is based on the Ljung-Box Q-statistic. This is a widely-used test, but, like Bartlett's standard errors, is based on the assumption that the data are *iid* Normal, and thus again is not applicable to heteroskedastic data. This test statistic is for the following null and alternative hypotheses:

$$\begin{aligned} H_0 &: \rho_1 = \rho_2 = ... = \rho_L = 0 \\ \text{vs. } H_a &: \rho_j \neq 0 \text{ for some } j = 1, 2, ..., , L \end{aligned}$$

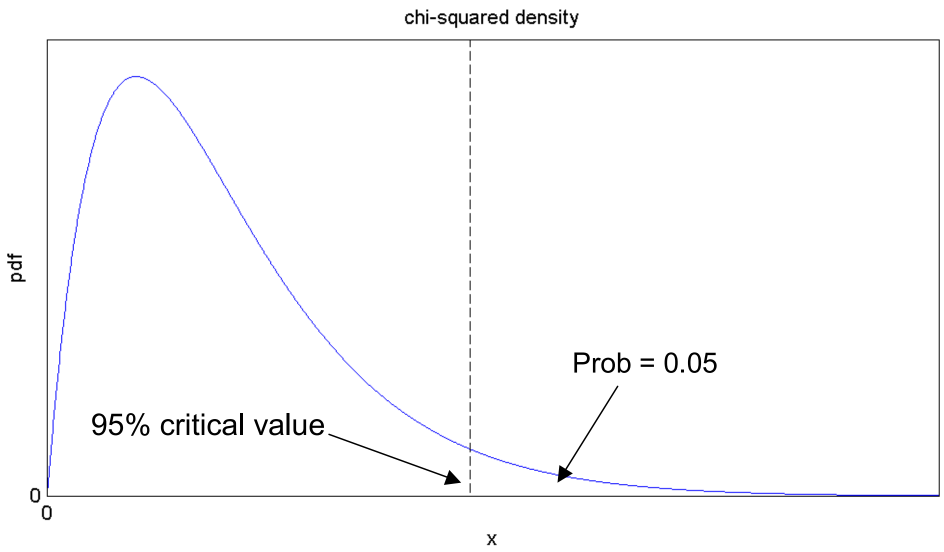


Figure 6.2: Chi-squared density, with 95% critical value.

The Ljung-Box Q-statistic, denoted $Q_{LB}(L)$, is:

$$Q_{LB}(L) = T(T+2)\sum_{j=1}^{L}\left(\frac{1}{T-j}\right)\hat{\rho}_j^2$$

The $Q_{LB}(L)$ statistic is simply a weighted sum of the squared autocorrelation coefficients, with j ranging from 1 to L. Under the null hypothesis, the $Q_{LB}(L)$ statistic is distributed as χ_L^2, a chi-squared random variable with L degrees of freedom. An example of a χ^2 density is given in Figure 6.2. With a chi-squared test we reject the null hypothesis if the test statistic (in our case, $Q_{LB}(L)$) is larger than the 95% critical value of a χ_L^2 random variable. The critical values for some values of L are given below:

95% Critical Values for the χ_L^2 distribution

L	**Critical value**
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
8	15.51
10	18.31
12	21.03
13	22.36
20	31.41

Note that the LB statistic requires the researcher to choose L, the number of autocorrelations we want to examine. This choice is usually somewhat arbitrary. On daily data I would recommend using 10 lags (possibly 20), on weekly data I would recommend 4 or 8 lags, on monthly data I would recommend 6 or 12 lags.

The asymptotic distribution of the Ljung-Box test statistic is based on the same *iid* assumption as the Bartlett standard error for a single sample autocorrelation. A better, robust, approach to test for serial correlation up to a given lag L is via a regression-based test: Estimate the following regression

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_L Y_{t-L} + e_t$$

and obtain robust standard errors for the parameter estimates. Then test the joint hypothesis:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = ... = \beta_L = 0 \\ \text{vs. } H_a &: \beta_j \neq 0 \text{ for some } j = 1, 2, ..., L \end{aligned}$$

via a χ^2-test. The null hypothesis imposes L restrictions on the parameters, and so the appropriate critical value will be that of a χ^2 variable with L degrees of freedom, as it is for the Ljung-Box test statistic.

The results of the Ljung-Box test and the robust test for autocorrelation for various choices of L are given below.

Ljung-Box tests for serial correlation						
L	5		10		20	
Crit val	*11.07*		*18.31*		*31.41*	
Test stat	*LB*	*Robust*	*LB*	*Robust*	*LB*	*Robust*
Data set 1	12.76*	9.02	21.86*	15.72	28.02	20.32
Data set 2	148.71*	17.79*	320.51*	44.74*	647.54*	72.51*

The table reveals that the conclusions for data set 1 depend heavily on whether the LB test or a robust test is used: in the former case significant autocorrelation is found when $L = 5$ or $L = 10$ (though not for $L = 20$), while in the latter case it is not found for any choice of L. For data set 2 the conclusions are the same whether the LB test or a robust test is used, and across all three choices of L. Thus we would conclude that there is significant evidence of autocorrelation in the second data set, but no significant evidence for the first data set.

6.5 Testing for predictability in our financial data

The sample autocorrelations of the three time series discussed above are presented in Figure 6.3.

Tests for autocorrelation are presented in the table below. Using the simple Ljung-Box test we would conclude that there is significant autocorrelation (up to lags 5 or 10) for all three series. However using a robust test for autocorrelation we only find evidence of

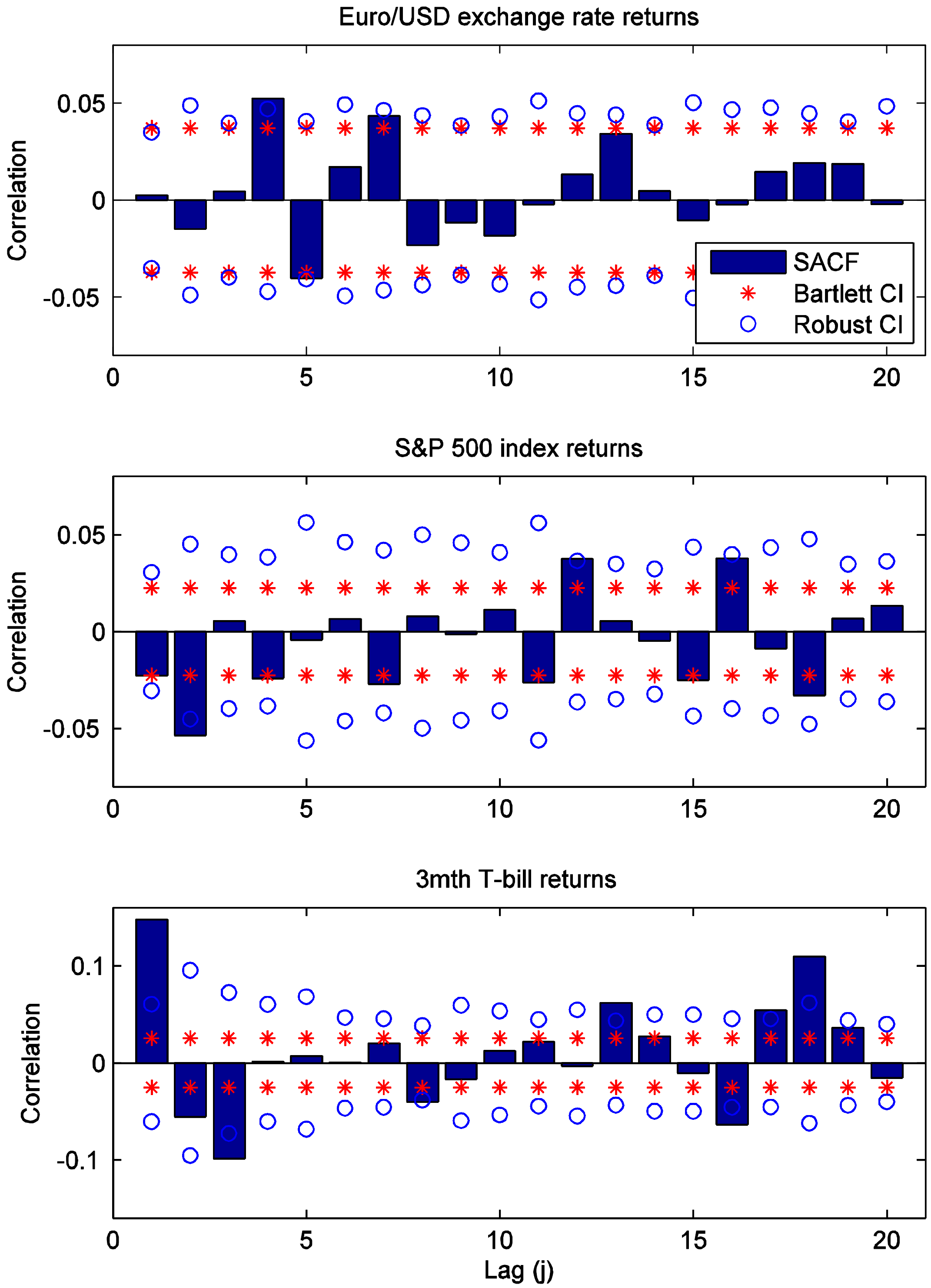


Figure 6.3: Sample autocorrelation functions of the daily returns on the exchange rate, stock index and T-bill, with Bartlett and robust confidence interval bounds.

significant autocorrelation for the T-bill rate. This implies that the exchange rate and the stock index returns are apparently not predictable using historical data, whereas the T-bill return is at least partially predictable.

Ljung-Box tests for serial correlation						
L	5		10		20	
Crit val	*11.07*		*18.31*		*31.41*	
Test stat	*LB*	*Robust*	*LB*	*Robust*	*LB*	*Robust*
FX	12.76*	9.02	21.86*	15.72	28.02	20.32
Stock	30.19*	5.80	37.39*	8.08	79.27*	21.50
T-bill	207.42*	32.68*	222.23*	38.31*	376.69*	87.69*

6.6 Overview of chapter

This chapter introduced some common summary statistics used when working with financial asset returns, and presented the Jarque-Bera test for normality. We also examined sample autocorrelations and hypothesis tests on these used to look for predictability of asset returns.

6.7 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Compute sample autocorrelations
- Interpret plots of sample autocorrelations and associated confidence intervals
- Interpret tests for predictability based on sample autocorrelations

6.8 Test your knowledge and understanding

1. Figure 6.4 presents the sample autocorrelation function for daily returns on the FTSE 100 index over the period January 1, 2002 to December 31, 2013.
   (a) Interpret this figure.
   (b) Describe how to test *jointly* for autocorrelation in squared residuals up to lag L using a regression-based approach.
   (c) The table below presents the results from such a test for three choices of L. Interpret these results.

L	5	10	20
Crit val	*11.07*	*18.31*	*31.41*
Test statistic	19.32	26.47	36.21

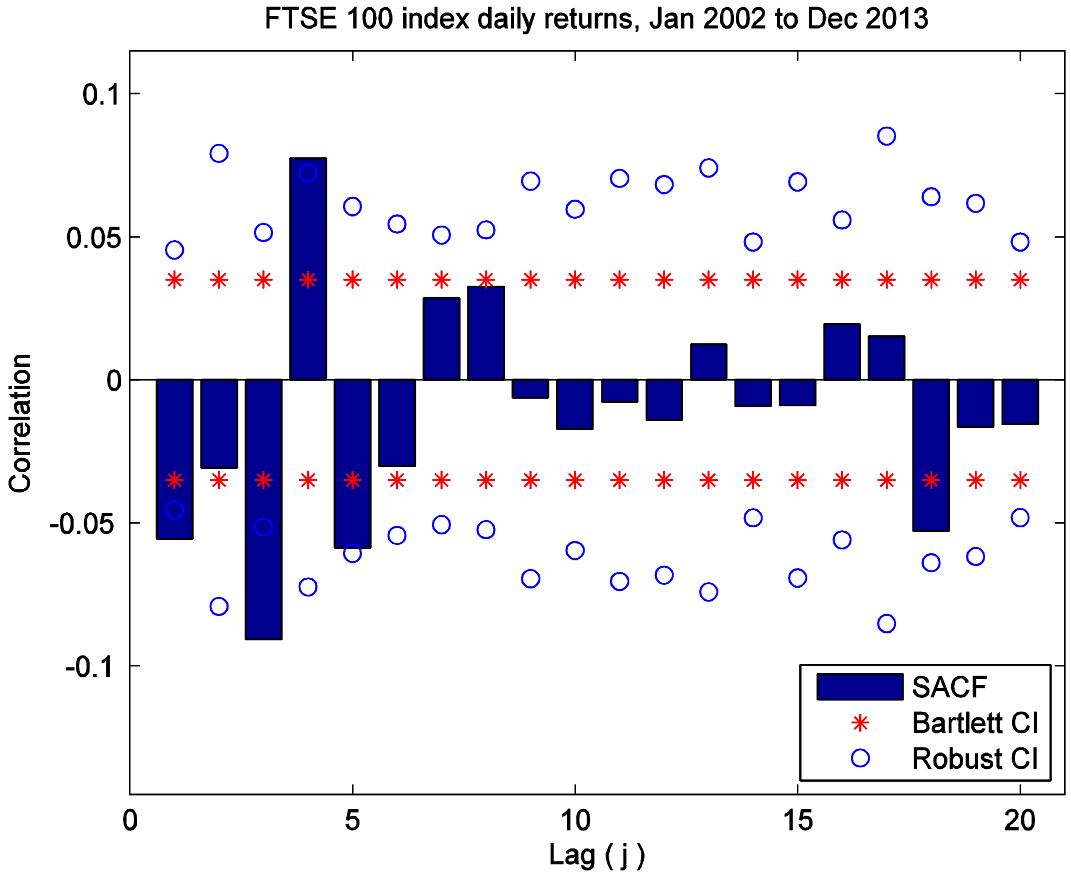


Figure 6.4: Sample autocorrelations for daily returns on the FTSE 100 index.

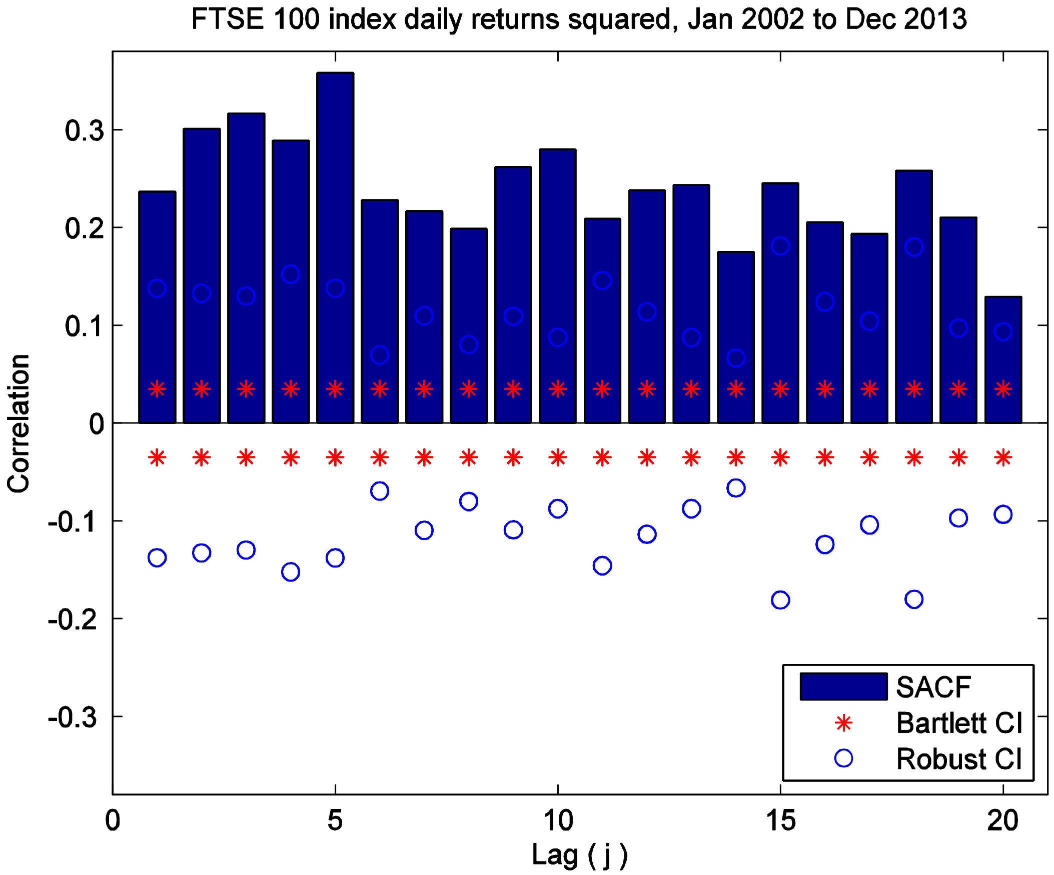


Figure 6.5: Sample autocorrelations for squared daily returns on the FTSE 100 index.

2. Figure 6.5 presents the sample autocorrelation function for *squared* daily returns on the FTSE 100 index over the period January 1, 2002 to December 31, 2013.

   (a) Interpret this figure.

   (b) Describe how to test *jointly* for autocorrelation in squared residuals up to lag L using a regression-based approach.

   (c) The table below presents the results from such a test for three choices of L. Interpret these results.

L	5	10	20
Crit val	*11.07*	*18.31*	*31.41*
Test statistic	57.89	80.63	121.27

3. Don't forget to check the VLE for additional practice problems for this chapter.

6.9 Solutions to activities

Activity 6.1

Population autocorrelations:

$$\rho_j = \frac{\gamma_j}{\gamma_0} = \frac{Cov\left[R_t, R_{t-j}\right]}{V\left[R_t\right]}$$

Now consider scaling returns by some constant $a \neq 0$:

$$\begin{aligned}
\rho_j\left(a\right) &\equiv \frac{Cov\left[aR_t, aR_{t-j}\right]}{V\left[aR_t\right]} \\
&= \frac{a^2 Cov\left[R_t, R_{t-j}\right]}{a^2 V\left[R_t\right]}, \text{ by properties of covariance and variance} \\
&= \frac{Cov\left[R_t, R_{t-j}\right]}{V\left[R_t\right]} = \rho_j
\end{aligned}$$

Now consider sample autocorrelations:

$$\begin{aligned}
\hat{\rho}_j &= \frac{\hat{\gamma}_j}{\hat{\gamma}_0} = \frac{\frac{1}{T-j}\sum_{t=j+1}^{T}\left(R_t - \bar{R}\right)\left(R_{t-j} - \bar{R}\right)}{\frac{1}{T}\sum_{t=1}^{T}\left(R_t - \bar{R}\right)^2} \\
\text{where } \bar{R} &\equiv \frac{1}{T}\sum_{t=1}^{T} R_t
\end{aligned}$$

Then we find

$$\overline{aR} = \frac{1}{T}\sum_{t=1}^{T} aR_t = a\left(\frac{1}{T}\sum_{t=1}^{T} R_t\right) = a\bar{R}$$

and

$$\begin{aligned}
\hat{\rho}_j\left(a\right) &= \frac{\frac{1}{T-j}\sum_{t=j+1}^{T}\left(aR_t - \overline{aR}\right)\left(aR_{t-j} - \overline{aR}\right)}{\frac{1}{T}\sum_{t=1}^{T}\left(aR_t - \overline{aR}\right)^2} \\
&= \frac{\frac{1}{T-j}\sum_{t=j+1}^{T}\left(a\left(R_t - \bar{R}\right)\right)\left(a\left(R_{t-j} - \bar{R}\right)\right)}{\frac{1}{T}\sum_{t=1}^{T}\left(a\left(R_t - \bar{R}\right)\right)^2}, \text{ since } \overline{aR} = a\bar{R} \\
&= \frac{a^2\frac{1}{T-j}\sum_{t=j+1}^{T}\left(\left(R_t - \bar{R}\right)\right)\left(R_{t-j} - \bar{R}\right)}{a^2\frac{1}{T}\sum_{t=1}^{T}\left(R_t - \bar{R}\right)^2} \\
&= \hat{\rho}_j
\end{aligned}$$

Let $X_t = aR_t$, where a is some positive constant. First we need to derive the mean and standard deviation of X_t :

$$
\begin{aligned}
\bar{X} &= \frac{1}{T}\sum_{t=1}^{T} X_t = \frac{a}{T}\sum_{t=1}^{T} R_t = a\bar{R} \\
\hat{\sigma}_x &= \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(X_t - \bar{X}\right)^2} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(aR_t - a\bar{R}\right)^2} \\
&= \sqrt{\frac{a^2}{T}\sum_{t=1}^{T}\left(R_t - \bar{R}\right)^2} = a\hat{\sigma}
\end{aligned}
$$

Then we use these to compute the skewness and kurtosis:

$$
\begin{aligned}
\hat{S}_x &= \frac{1}{\hat{\sigma}_x^3}\frac{1}{T}\sum_{t=1}^{T}\left(X_t - \bar{X}\right)^3 = \frac{1}{a^3\hat{\sigma}^3}\frac{1}{T}\sum_{t=1}^{T} a^3\left(R_t - \bar{R}\right)^3 = \frac{1}{\hat{\sigma}^3}\frac{1}{T}\sum_{t=1}^{T}\left(R_t - \bar{R}\right)^3 \equiv \hat{S} \\
\hat{K} &= \frac{1}{\hat{\sigma}_x^4}\frac{1}{T}\sum_{t=1}^{T}\left(X_t - \bar{X}\right)^4 = \frac{1}{a^4\hat{\sigma}^4}\frac{1}{T}\sum_{t=1}^{T} a^4\left(R_t - \bar{R}\right)^4 = \frac{1}{\hat{\sigma}^4}\frac{1}{T}\sum_{t=1}^{T}\left(R_t - \bar{R}\right)^4 \equiv \hat{K}
\end{aligned}
$$

Thus while the mean and standard deviations change, the skewness and kurtosis does not.

Activity 6.2

Recall that for an AR(1)

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN\left(0, \sigma^2\right)$$

we have

$$\rho_1 = \phi_1.$$

Thus the method-of-moments estimator of ϕ_1 is simple

$$\hat{\phi}_1 = \hat{\rho}_1 = -0.4.$$

Note that given just the sample first-order autocorrelation we cannot get an estimate of the intercept, ϕ_0. (For that we would need some additional information, such as the sample mean of Y_t.)

For an MA(1)

$$Y_t = \mu + \theta\varepsilon_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN\left(0, \sigma^2\right)$$

we have

$$\rho_1 = \frac{\theta}{1+\theta^2}$$

rearranging and solving this equation for θ yields:

$$
\begin{aligned}
\rho_1\theta^2 - \theta + \rho &= 0 \\
\theta &= \frac{1 \pm \sqrt{1-4\rho^2}}{2\rho}
\end{aligned}
$$

so we have *two* possible method-of-moments estimators of θ :

$$\hat{\theta} = \frac{1 \pm \sqrt{1 - 4 \times 0.4^2}}{2 \times -0.4} = -2 \text{ or } -0.5.$$

We usually take the value of θ that is less than 1 in absolute value (this is known as the 'invertible' solution) and so we would set $\hat{\theta} = -0.5$. Note that we again do not have enough information to identify the intercept, μ.

Activity 6.3

Lag (j)	$\hat{\rho}_j$	*t-stat*$_j$
1	-0.25	-0.86
2	-0.26	-0.91
3	-0.07	-0.23

Thus all three of these autocorrelations are negative, with the second lag being (just slightly) the strongest. Negative correlations indicate that good years (above average returns) tend to be followed by bad years (below-average returns), and vice versa. However all t-statistics are less than 1.96 in absolute value, and so we fail to reject each null that the given autocorrelation is equal to zero.

Chapter 7
The efficient markets hypothesis and market predictability

7.1 Introduction

In this chapter we relate the concept of efficient markets, defined in various ways, to the evidence of predictability of financial variables. Much of modern quantitative finance relates to methods and models for predicting aspects of asset returns, and yet the classical theory of efficient markets may appear to suggest that asset returns should be completely unpredictable. This chapter will reconcile the empirical evidence for asset return predictability with the concept of an efficient market.

7.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce the 'efficient markets hypothesis' (EMH) and discuss it with respect to forecasting in financial markets
- Discuss the various forms of the EMH, and refinements of the EMH that have been proposed
- Introduce the idea of 'data snooping' and how this relates to the EMH

7.1.2 Learning outcomes

By the end of this chapter, and having completed the activities, you should be able to:

- Discuss the differences between weak-form, semi strong-form and strong-form efficiency of markets.
- Discuss some recent refinements of the concept of market efficiency, with reference to the growing set of forecasting models and 'ephemeral predictability.'
- Discuss how 'data snooping' may explain some apparent evidence against market efficiency.

7.1.3 Essential reading

These notes serve as the essential reading for this topic.

7.1.4 Further reading

More advanced treatments of the efficient markets hypothesis are provided in:

- Campbell, John Y., Andrew W. Lo and A. Craig Mackinlay *The Econometrics of Financial Markets.* (Princeton University Press, Princeton, New Jersey, 1997) [ISBN 0691043019] Chapter 2 Section 1, and Chapter 1 Section 5.

- Granger, Clive W. J. and Allan Timmerman, 'Efficient Market Hypothesis and Forecasting', *International Journal of Forecasting*, 20(1) 2004, pp. 15-27.

For an introductory treatment of the efficient markets hypothesis, you may consult an introductory investments text book, as most of those include a chapter on the efficient markets hypothesis. For example, see:

- Bodie, Z., A. Kane and A.J. Marcus *Investments.* (McGraw-Hill, U.S.A., 2013) ninth edition [ISBN 0073530700] Chapter 11.

- Elton, E.J., M.J. Gruber, S.J. Brown and W.N. Goetzmann *Modern Portfolio Theory and Investment Analysis.* (John Wiley & Sons, New York, 2009) eighth edition [ISBN 0470388323] Chapter 17.

- The SEC and CFTC joint report on the 2010 'flash crash' is available at `http://goo.gl/0etv`, [accessed 20 October 2014].

- Further information on the 2013 Nobel prize in economics is available at `http://goo.gl/XRuxjA`, and more technical background on the work of the recipients of the prize is available at `http://goo.gl/BNZ28Z`, [both accessed 20 October 2014].

Research articles on data snooping and statistical methods to control for it are:

- For the econometric theory:
  - White, Halbert. 'A Reality Check for Data Snooping,' *Econometrica*, 68, 2000, pp. 1097-1126.
- For applications:
  - Sullivan, Ryan, Allan Timmermann and Halbert White 'Data-Snooping, Technical Trading Rules and the Bootstrap,' *Journal of Finance*, 54, 1999, pp. 1647-1692.
  - Sullivan, Ryan, Allan Timmermann and Halbert White 'Dangers of Data Mining: The Case of Calendar Effects in Stock Returns,' *Journal of Econometrics*, 2001, pp.249-286.

7.1.5 References cited

- Black, F. 'Noise,' *Journal of Finance*, 1986, 41, pp.529 – 543.
- Jensen, M., 'Some anomalous evidence regarding market efficiency,' *Journal of Financial Economics*, 1978, 6, pp.95 – 101.
- Malkiel, B., 'Efficient Market Hypothesis,' in Newman, P., M. Milgate and J. Eatwell (eds), *New Palgrave Dictionary of Money and Finance*, 1992, Macmillan, London.
- Roberts, H. 'Statistical versus clinical prediction of the stock market,' Unpublished manuscript, Center for Research in Security Prices, University of Chicago, 1967.
- Timmermann A. and C. W. J. Granger, 'Efficient markt hypothesis and forecasting,' *International Journal of Forecasting*, 2004, 20, 15–27.

7.2 The efficient markets hypothesis

Broadly stated, the efficient markets hypothesis (EMH) postulates that market prices of financial assets fully reflect all available information about the value of the asset. In an efficient market, then, one might expect that it is impossible to predict the value of a financial asset, as any information that might be used to predict the value of the asset will already be incorporated into the current price. This hypothesis is around a half-century old, and since its inception it has been a topic of debate, with researchers conducting various studies and finding, or not finding, apparent violations of EMH.

The EMH has serious policy implications as well: if markets truly are efficient, then the best way to determine the value of an asset is to simply observe its market price; there is no need for external or independent studies of the value of the asset, or for regulation of the price of the asset. This interpretation has been called into question many times in the past decade: the dot.com bubble saw firms valued at billions of dollars crashing into worthlessness with little apparent change in their structure or business model; the recent financial crisis witnessed complicated derivative securities trading at what appear (*ex post*) to be grossly over-valued prices.

This chapter takes the EMH, or variations of it, as a benchmark for thinking about prices of financial assets. Below we link ideas from the EMH with topics from the forecasting literature, seeking to reconcile the EMH with evidence of predictability in financial markets.

7.3 Standard definitions of market efficiency

The literature on market efficiency started in the 1960s, and in that decade and the next, several papers proposed formal definitions of 'market efficiency.' We will first consider a simple definition of market efficiency, and then later in this chapter we will consider more recent refinements of the EMH.

7.3.1 Roberts, Jensen and Malkiel definitions

Definition 7.1 (Market Efficiency, Jensen 1978) A market is efficient with respect to information set Ω_t if it is impossible to make economic profits by trading on the basis of information set Ω_t.

This was one of the first definitions of market efficiency. This definition was refined by Malkiel (1992), who stated that a capital market is efficient if it fully and correctly reflects all relevant information in determining securities prices. Further, he noted that if the market is efficient with respect to some information set, Ω_t, then securities prices would be unaffected by revealing that information to all participants.

Both of the above definitions emphasise that there are **three** elements of importance in defining market efficiency:

1. **The information set:** what information are we considering? (This will be discussed further below.)
2. **The ability to exploit the information in a trading strategy:** can we use the information to form a successful trading strategy/investment rule?
3. **The performance measure for the trading strategy is *economic profits*:** that is, risk-adjusted (because investors are risk-averse) and net of transaction costs.

Roberts (1967) defined three types of market efficiency, depending on the 'size' of Ω_t:

- **Weak-form efficiency:** The information set contains only historical values of the asset price, dividends (and possibly volume) up until time t.
- **Semi Strong-form efficiency:** The information set contains all *publicly available* information as at time t.
- **Strong-form efficiency:** The information set contains all information known to any market participant (i.e., all *public and private* information) up until time t.

Publicly available information at time t includes:

1. Historical values of the price of the asset under analysis
2. Historical values of any other financial asset, such as
   (a) market indices like the FTSE100 and S&P500,
   (b) interest rates,
   (c) option prices,
   (d) exchange rates, etc.
3. Historical values of other variables that might influence stock prices, such as
   (a) unemployment rates,
   (b) GDP figures,

   (c) indices of consumer confidence

   (d) numbers of new housing starts

4. Absolutely any other piece of information that you could obtain if you looked for it in a publicly available source

Privately available information includes things like:

5. Planned closures of plants or factories that have not yet been reported publicly

6. Knowledge that the CEO is leaving, not yet reported publicly

7. Any other piece of information that may affect the future stock price that has not been made public. (That is, all the types of information that it is usually illegal to trade on before it is made public.)

Weak-form efficiency is based on the information set in point (1) above. Semi strong-form efficiency is based on points (1) to (4), and strong-form efficiency is based on points (1) to (7). Financial market regulators usually have laws against trading on the basis of private information, such as those in points (5) to (7). Since we can never really hope to know all privately available information, strong-form efficiency is not widely studied. The weak-form and semi strong-form efficiency of markets have been widely studied.

> **Activity 7.1** Suppose we are at time t, and we are interested in the market for a given stock. Let Ω_t^W be the weak-form efficient markets information set at time t, Ω_t^{SS} be the semi strong-form efficient markets information set at time t, and Ω_t^S be the strong-form efficient markets information set at time t. To which information set, if any, do the following variables belong?
> a. The stock price today.
> b. The risk-free interest rate today.
> c. The unemployment rate last year.
> d. Next year’s production figures just approved by the company’s board of directors.
> e. The value today of an option on the stock, which expires in three months’ time.
> f. The value of the stock at time $t + 1$.
> g. The number of shares Warren Buffett purchased today of the stock.

7.4 The EMH and ‘bubbles’

A reasonable summary of the EMH is that it implies the absence of arbitrage opportunities, given some information set. Black (1986) proposed that, in addition to the no-arbitrage condition, a market efficiency definition should include some reference to possible deviations of the market price from the fundamental value. He proposed that the price should be within a factor of two (an admittedly arbitrary figure) of the intrinsic value of the assets. This relates to asset price ‘bubbles’, where the price of an asset far exceed the fundamental value of the asset.

According to Black’s definition a market may be efficient if asset price bubbles do not grow ‘too big’ or last ‘too long’ before they are corrected. The problem with empirically testing this definition is that the fundamental value of most assets is hard to measure and so any test would be a *joint test* of Black's market efficiency definition and the model used to determine the intrinsic value of the assets.

An absence of arbitrage opportunities does not rule out *all* forms of predictability in asset returns. Let us now consider some situations where some form of the EMH may hold, and predictability may exist.

7.5 Transaction costs and other market frictions

The presence of market frictions, such as transactions costs (stamp duty, costs of information gathering and processing, etc.), or trading constraints (like short selling constraints) imply that even if the current market price is slightly different from what it 'should' be (on the basis of forecasted future returns and dividends), an arbitrage opportunity may not exist.

1. **Transaction costs:** If the stock price is £0.10 below what you think it should be, but it's going to cost you £0.20 to buy the stock in transaction fees, then you won't bother. Thus small deviations from efficient markets are possible when one considers the presence of transactions costs. Predictable patterns in asset returns only violate the EMH if they are large enough to cover transaction costs. It is important to note here that transaction costs have changed through time (as technology has made it cheaper to trade many assets) and they vary across investors: small investors (like university professors) usually face higher transaction costs than large investors (like mutual funds or hedge funds), for example.

2. **Information processing costs:** If it costs £40,000 per year to hire a graduate to analyse the data and determine that there is an arbitrage opportunity, then many small opportunities will not be picked up, a point made by Grossman and Stiglitz (1980). This relates to one of the explanations for the presence of market analysts: if the benefits of hiring an analyst to find and exploit market inefficiencies exactly matches the salary he or she charges, then analysts may still be employed in a fully efficient market. More recently, the growth of 'algorithmic trading' (trading in stocks with only limited human intervention) has highlighted a different source of information processing costs, namely the cost of high speed computers and fast connections to the exchange.

3. **Market impact:** In practice it may be difficult or impossible to take a £0.10 deviation of the price of a stock from its 'true' value and buy 1 million stocks so as to make £100,000: if an investor really tried to do this, the bid-ask spread would widen and the market supply of the asset would shrink. The 'price impact' of such a strategy would gradually get larger, to the point where the transaction cost would out-weigh the profit.

4. **Trading restrictions:** If short-selling constraints, a common type of trading restriction, are present then certain types of predictability may exist and no arbitrage opportunities be present. For example, if you predicted that a particular stock was going to fall in price over the coming month you would want to short sell

> it (i.e., borrow the stock from a broker today, sell it on the market today, then re-purchase the stock in one month's time and give it back to the broker). However, if you were barred from short selling (e.g., you work at a mutual fund) then you would not be able to follow this trading strategy.

In practice, many market participants are indeed barred from short selling (e.g., certain mutual funds) or face high fees to short sell (e.g., small investors). However, some participants are able to short sell (e.g., hedge funds) and so while not everyone can follow this type of trading strategy, some investors *can* and so we would not expect the arbitrage opportunity to exist for long. Indeed, the ability of hedge funds to take short positions, when many other market participants cannot, is one of the possible economic benefits of the presence of hedge funds in the market. The recent astronomical growth of algorithmic trading is thought to have made small deviations of asset prices from their 'fundamental' values even shorter lived.

7.6 New forecasting models and model selection

Many researchers, when proposing new statistical models for forecasting asset returns, apply the model to historical data to see how it works. Is it reasonable to conclude that finding that a new model is able to forecast historical asset returns constitutes evidence against the EMH? For example, many flexible and complicated models are now available (we will look at some of these later in this course), some of which are computationally intensive to estimate. If we go back and look at data from the 1920s and find that an investor using a such models could have made economic profits, should we conclude that the market was inefficient?

Most people think the answer is 'no'. Timmermann and Granger (2004) propose refining the definition of market efficiency to capture this idea. They do so by considering a set of forecasting models, $\mathcal{M}_t$, that can grow over time as researchers develop new methods and as computing power allows for increasingly complicated models.

When we expand the definition of market efficiency to consider an evolving set of forecasting models, $\mathcal{M}_t$, we are also led to think about how we *choose* one of these models to produce a forecast; the 'search technology.' Do we pick the one that performed best on average? The one that performed best over the most recent n months? What metric do we use to determine 'best'? (We will look at a variety of methods for choosing forecasting models in this course, including information criteria, out-of-sample measures, and statistical tests, some of which are computationally intensive and would not have been feasible 50 years ago.) Let $\mathcal{S}_t$ denote the 'search technology' that is available at time t for selecting a forecasting model from some set of models.

Definition 7.2 (Market Efficiency, Timmermann and Granger, 2004) A market is efficient with respect to the information set Ω_t, search technologies $\mathcal{S}_t$, and forecasting models $\mathcal{M}_t$ if it is impossible to make economic profits by trading on the basis of a forecasting model in $\mathcal{M}_t$, selected with a search technology in $\mathcal{S}_t$, based on predictor variables in Ω_t.

This definition takes into account the fact that certain forecasting models were not always available. The computationally intensive models of today were not available and would have been impossible to work with given the computing technology of 50 or 100 years ago. Thus efficiency should be defined with reference to an information set, a set of forecast models, and a method for choosing a forecasting model.

7.7 Ephemeral forecastability

Once a particular predictive relation has become public knowledge we would expect it to disappear. For example, suppose someone discovered that small firms pay higher returns in recessions. The discoverer of this fact would then buy small firms' stock during recessions, driving up their stock prices, and thus driving down their returns until the point where the returns on small firms is 'right'. So even if the deviation from market efficiency (small firms paying too high a return) was true in some sample period, it may not be true in a subsequent sample period. (This appears to be what happened after Banz (1981) published his paper on the 'size' effect.)

This idea may lead to a further refinement of the definition of market efficiency to include some reference to the holding period, or horizon. This definition below allows a market to be defined 'efficient', and still exhibit short periods of deviations from efficiency.

Definition 7.3 (Market Efficiency for long horizons) A market is efficient with respect to the information set Ω_t, search technologies $\mathcal{S}_t$, and forecasting models $\mathcal{M}_t$ for horizons greater than τ if it is impossible to make economic profits by trading on the basis of a forecasting model in $\mathcal{M}_t$, selected with a search technology in $\mathcal{S}_t$, based on predictor variables in Ω_t over periods of duration greater than τ.

The concept of ephemeral forecastability has three applications. The first relates to the incorporation of information into market prices. Some research in market microstructure (the analysis of asset returns within a single trade day) suggests that there are arbitrage opportunities at very short horizons (say, 5 or 10 minutes) but they disappear at the 1 hour horizon. Recent developments in algorithmic trading may be shrinking these periods of ephemeral forecastability – 'algos' now trade at the millisecond frequency. Thus the existence of short periods of predictability in this case relates more to information, and links with the ideas of Grossman and Stiglitz (1980) on the profits of gathering price-relevant information.

The second application relates to short-lived periods of predictability that are *not* information based. A prominent example of this is the 'flash crash' of 2010. On May 6 of that year, many U.S. equity indices (such as the S&P 500 and the Dow Jones Industrial Average) lost around 4% in value over the course of morning trading, and suddenly fell a further 5 to 6% at around 2:30pm. Within 20 minutes, however, both indices recovered almost all of the sudden loss. Many individual U.S. equities followed a similar trajectory, falling between 5 and 15% in a brief period around 2:30pm, and recovering very quickly afterwards. (See the SEC and CFTC joint report on the 'Market Events of May 6, 2010' for a more detailed description.) The generally accepted explanation for these events is that a large sell order placed by an algorithmic trader triggered a flurry of other sell orders and led to a loss of liquidity in many markets. This meant that the price was able to move (fall) very quickly. While a 5-10% drop is not overly large, the fact that it occurred in such a short period of time, and without any apparent reason, was a cause of much concern. This relates to Black's definition of market efficiency, that prices should not deviate too far, or for too long, from the intrinsic value of the asset. I suspect that Black would rule that the 'flash crash' was indeed a period of market inefficiency.

The thid application of this definition of market efficiency relates to the incorporation of statistical relationships, via models, into market prices, related to the example of small firms given at the start of this section. If a new model is able to capture an economically profitable relationship between financial assets, but this relationship is only profitable for a finite period, then it also is an example of ephemeral forecastability. Presumably the period of time that a new forecasting model is profitable, if it is ever profitable, is on the order of months or years, not minutes or hours. This idea helps explain the economic incentive to develop new forecast models (and there must be an incentive - there are many people out there working on new forecast models or new trading strategies).

7.8 Data snooping

Many papers have claimed to have found a relationship between future returns and some observable variable. Some examples:

- Returns on the stock market are lower in October
- Returns on the stock market are lower on Mondays
- Returns on the stock market are higher in January
- Temperature in Papua New Guinea for forecasting the Dow Jones Industrial Average

One interpretation of the above results is that these relationships represent evidence against the EMH. An alternative explanation is that the findings are the outcomes of individual or collective data snooping. The problem of data snooping (otherwise known as 'data mining') comes from the fact that most of the data we have in economics is observational not experimental, thus everyone is using and re-using the same data sets. If an individual searches hard enough for a pattern/correlation in a given data set, he/she can usually *seem to* find something.

7.8.1 Refresher: Type I and Type II errors

Recall that in a hypothesis test that we postulate a *null* (H_0) and an *alternative* (H_a or H_1) hypothesis. The following matrix shows the four possible combinations of true hypothesis and decision made on the basis of a statistical test.

		Decision	
		Fail to reject H_0	Reject H_0
Truth	H_0 is true	Correct decision	Type I error
	H_1 is true	Type II error	Correct decision

With a statistical test we know we will sometimes make a mistake, and prior to running the test we make a choice on how often we are prepared to falsely reject the null hypothesis: we choose the level (α) of the test. Setting $\alpha = 0.05$, for example, means we are prepared to make a Type I error 5% of the time. The 'power' of a test is the proportion of times we reject the null hypothesis when the null is false. It is equal to one minus the Type II error rate, and a good test has high power.

> **Activity 7.2** Consider a test that has a Type I error rate of 5%, and power of 50%. Suppose, before running the test, that the researcher thinks that both the null and the alternative are equally likely.
>
> 1. If the test indicates a rejection of the null hypothesis, what is the probability that the null is false?
>
> 2. If the test indicates a failure to reject the null hypothesis, what is the probability that the null is true?

7.8.2 Individual vs. collective data snooping

Collective data snooping (also known as the 'file drawer problem') comes from the fact that (usually) only statistically significant relationships are given attention and/or published. Consider the case that 100 researchers are each thinking of one particular variable as being useful for forecasting the stock market. Two of these researchers find a significant result, and the other 98 do not. The two significant results will attract attention, while the remaining 98 generally will not. (The remaining 98 researchers put their non-significant results in their file drawers and forget about them.) If all of the 100 researchers conducted 5% level tests, then even if *none* of the 100 variables were truly useful for forecasting we would expect around 5 researchers to find a significant relationship just by chance. The finance literature may be scattered with examples of apparently significant predictive relationships that are due just to Type I errors.

Individual data snooping occurs when the *same* researcher considers all 100 (or 1000, or 1 million, etc.) possible predictor variables himself/herself, and then only reports the significant results. Statistical methods to adjust the critical values when testing a large number of possible predictor variables have been proposed to account for this problem. (See the references given in the 'Further Reading' section above.) When one of these new methods was applied to testing whether the 'calendar effects' (significantly better or worse returns on certain days of the week, months of the year, etc.) that had been reported in the literature were truly significant *none* were confirmed; all of them were dismissed as being the result of data snooping (though not everyone is convinced).

The difference between individual data snooping and collective data snooping is subtle. In terms of the probability of seeing a result that is significant but not truly important there is no difference. The main difference is that it is possible to control for individual data snooping, by conducting joint tests or other tests that take into account the number of models that were tried, whereas there is no real way of controlling for collective data snooping. One possible way to minimise the impact of collective data snooping is to maintain a healthy scepticism of results that are statistically significant but based on weak or strange economic arguments.

> **Activity 7.3** The *Journal of Finance* publishes approximately 15 articles per issue, and 6 issues per year. Approximately 80% of the articles in the *Journal of Finance* are empirical (as opposed to theoretical articles which use no data). Assume that one-half of the empirical articles are published because they report some new, statistically significant (at the 5% level), relationship between financial variables. If the tests reported in these articles are independent of each other, how many articles per year are published simply because of a Type I error?

7.9 Predictability of other properties of asset returns

Up to this point we have mostly focussed on the forecasting of asset prices (or equivalently, returns). But there are other properties of asset returns that may be of interest, for example: the risk of an asset return. As we have all seen, financial markets tend to go through periods of great turbulence and of relative tranquility. Thus the risk (or volatility, sometimes measured as variance) of asset returns is time-varying, and might possibly be forecasted. (Indeed, we will focus on forecasting risk later in this guide.) Forecasting the risk of an asset is useful for portfolio decisions, risk management, option pricing, amongst other things.

There are many other properties of the distribution of asset returns that may be of interest in addition to the return itself: risk (as we just mentioned), crash probability, positive/negative return probability, skewness, or even the entire distribution itself. Forecasting objects other than returns themselves (such as the characteristics just mentioned) is a relatively new and very active area of research in financial econometrics. These other properties of returns often relate to risk, and are of much interest in risk management. Risk management is an important function within most organisations, both financial and non-financial, and the rewards to good risk management are thought to be large.

If there are no trading strategies that exist to make economic profits from an accurate risk forecast (e.g., financial assets to exploit this type of information may simply not exist) then highly forecastable asset return risk will not violate the EMH.

A recent development related to this point is the growth in the markets for ‘variance swaps’ and ‘volatility swaps’. These are derivative securities written on the sample variance (or standard deviation) of some underlying asset over a fixed interval of time. The availability of these securities makes it possible to directly implement trading strategies based on risk forecasts, rather than price forecasts. These securities, however, are currently only available on a limited set of underlying assets, although the set of underlying assets may grow in the future.

7.10 Random walks versus EMH

Some early work on efficient markets equated the 'random walk' model for asset prices with the EMH. The random walk model for stock prices (or the log of stock prices) is:

$$\begin{aligned} P_{t+1} &= P_t + \varepsilon_{t+1} \\ \text{where} \quad \varepsilon_{t+1} &\sim iid\, WN(0) \\ \text{or} \quad \varepsilon_{t+1} &\sim WN(0) \end{aligned}$$

This model suggests that tomorrow's price is equal to today's price plus some innovation term, ε_{t+1}. There are two versions of the random walk model: the first assumes that the innovation term is *iid* with zero mean; the second assumes only that the innovation term is a mean-zero white noise process. The idea behind the statement that in an efficient market prices should follow a random walk is that at each point in time the price of an asset should reflect all available information, so $E[P_{t+1}|P_t, P_{t-1}, ...] = P_t$, and thus it should not be possible to profit by trading on the basis of old information. More recent work does *not* accept that the EMH implies that prices follow a random walk. For example:

1. Under the random walk model we have: $E_t[P_{t+1}] = P_t$ and so $E_t[R_{t+1}] = 0$ which violates basic finance theory that investors require a reward (in terms of expected return) for holding a risky asset. At the very least the model should be extended to $P_{t+1} = \mu + P_t + \varepsilon_{t+1}$ to allow a reward for holding a risky asset.

2. There is substantial evidence of predictability in volatility, i.e. the variance of the innovation term ε_{t+1}, which would reject the assumption that ε_{t+1} is *iid* while it would not reject the EMH if there exists no asset which can turn this predictability into economic profits. This evidence is not sufficient to reject the second type of random walk models above.

7.11 Conclusion

What do theories of financial market efficiency have to tell us about whether we should bother learning forecasting techniques? Campbell, *et al.* (1997) concluded 'Recent econometric advances and empirical evidence seem to suggest that financial asset returns are predictable to some degree. Thirty years ago this would have been tantamount to an outright rejection of market efficiency. However, modern financial economics teaches us that other, perfectly rational, factors may account for such predictability.' Thus studying forecasting techniques may still have some use in an efficient market.

The EMH tells us only a little about forecasting other properties of asset returns, such as their risk, and it turns out that forecasting risk is much easier than forecasting returns (in terms of the success of the models, not necessarily in terms of the complexity of the models). Further, we have seen that forecasting historical prices will probably be easier than forecasting future prices (but it is not so impressive to tell people you would have made money using your forecasts if only you could go back in time).

Finally, the field of 'behavioural finance' is an active one, and many apparent deviations from fully rational investor behavior have been documented. These deviations include such things as under- or over-reaction to earnings announcements, an apparent lack of international diversification in portfolios, or apparent IPO mis-pricing. (These phenomena are empirically well-documented, but there exist both behavioral and 'rational' explanations for them.) The presence of predictability in financial markets is no longer surprising to believers in efficient markets, but was never surprising in the first instance to behavioral economists.

The 2013 Nobel prize in economics was awarded to three professors for their work on understanding asset prices. One of these, Eugene Fama, is one of the 'founding fathers' of efficient markets theory, and his work in the 1960s showed that it is very difficult to predict stock returns over short horizons, providing evidence consistent with the weak-form efficient markets hypothesis. A second recipient, Robert Shiller, was awarded the prize for his research on longer-run predictability of asset markets, and for rational and behavioral explanations for this predictability. The third recipient, Lars Peter Hansen, received the prize for developing econometric tools (most famously, the 'generalised method of moments') for estimating and testing (rational) models of asset prices. The awarding of the Nobel prize jointly to these three researchers, with their somewhat conflicting views of asset market behaviour, serves to illustrate the many ways that may simultaneously exist to view asset market behaviour.

7.12 Overview of chapter

This chapter introduced the 'efficient markets hypothesis' (EMH) and discussed its relationship with evidence of predictability of financial variables. We discussed different 'forms' of the EMH (weak form, semi-strong form, and strong form) as well as various extensions and refinements that have been proposed in the literature. We also discussed how the EMH relates to the idea of 'bubbles' in financial markets, and to the problem of data snooping.

7.13 A reminder of your learning outcomes

By the end of this chapter you should be able to:

- Discuss the differences between weak-form, semi strong-form and strong-form efficiency of markets.
- Discuss some recent refinements of the concept of market efficiency, with reference to the growing set of forecasting models and 'ephemeral predictability'.
- Discuss how 'data snooping' may explain some apparent evidence against market efficiency.

7.14 Test your knowledge and understanding

1. Why is testing Black's definition of market efficiency difficult to test in practice?

2. What is Granger and Timmermann's definition of market efficiency? What was their motivation for proposing their refinement of the standard definition?

3. What is the difference between 'collective' and 'individual' data snooping?

4. When would being able to perfectly forecast the *variance* of tomorrow's return on Company A violate the EMH?

5. What is the 'efficient markets hypothesis', and what are the three key elements of the original definition?

6. Discuss *one* of the modifications/extensions/refinements of the original definition of the efficient markets hypothesis.

7. Consider a test that has a Type I error rate of 5%, and power of 33%. Suppose, before running the test, that the researcher thinks that both the null and the alternative are equally likely.

   (a) If the test indicates a rejection of the null hypothesis, what is the probability that the null is false?

   (b) If the test indicates a failure to reject the null hypothesis, what is the probability that the null is true?

8. Don't forget to check the VLE for additional practice problems for this chapter.

7.15 Solutions to activities

Activity 7.1

1. (a) The stock price this today $\Rightarrow \Omega_t^W$

   (b) The risk-free interest rate today $\Rightarrow \Omega_t^{SS}$

   (c) The unemployment rate last year $\Rightarrow \Omega_t^{SS}$ (and also Ω_{t-1}^{SS}, Ω_{t-2}^{SS}, .. all the way back to Ω_{t-365}^{SS} when the rate was announced)

   (d) Next year's production figures just approved by the company's board of directors $\Rightarrow \Omega_t^S$

   (e) The value today of an option on the stock, which expires in 3 months' time $\Rightarrow \Omega_t^{SS}$

   (f) The value of the stock at time $t+1 \Rightarrow$ none, it belongs to Ω_{t+1}^W, the weak-form information set for *tomorrow.*

   (g) The number of shares Warren Buffett purchased today of the stock $\Rightarrow \Omega_t^S$, assuming his trades are not public knowledge

Activity 7.2

Let $R = \{\text{reject null}\}$, $A = \{\text{fail to reject null}\}$, $T = \{\text{null is true}\}$, $F = \{\text{null is false}\}$. Then we are given:

$$\begin{aligned}
\Pr[F] &= \Pr[T] = 0.5 \\
\Pr[R|T] &= 0.05, \text{ so } \Pr[A|T] = 1 - 0.05 = 0.95 \\
\Pr[R|F] &= 0.5, \text{ so } \Pr[A|F] = 1 - 0.5 = 0.5
\end{aligned}$$

Now we want to find $\Pr[F|R]$. First we need to find $\Pr[R]$:

$$\begin{aligned}
\Pr[R] &= \Pr[R|T]\Pr[T] + \Pr[R|F]\Pr[F], \text{ by the `law of total probability'} \\
&= 0.05 \times 0.5 + 0.5 \times 0.5 \\
&= 0.275
\end{aligned}$$

Then:

$$\begin{aligned}
\Pr[F|R] &= \frac{\Pr[R|F]\Pr[F]}{\Pr[R]}, \text{ by Bayes' Rule} \\
&= \frac{0.5 \times 0.5}{0.275} \\
&= 0.91
\end{aligned}$$

Thus from an initial view that $\Pr[F] = \Pr[T] = 0.5$, a test rejection leads the researcher to update these probabilities to $\Pr[F|R] = 0.91$ and $\Pr[T|R] = 0.09$.

Now we compute $\Pr[T|A]$, and we use the fact that $\Pr[A] = 1 - \Pr[R] = 0.725$

$$\begin{aligned}
\Pr[T|A] &= \frac{\Pr[A|T]\Pr[T]}{\Pr[A]} \\
&= \frac{0.95 \times 0.5}{0.725} \\
&= 0.66
\end{aligned}$$

Activity 7.3

$$\begin{aligned}
\text{Number of articles per year} &= 15 \times 6 = 90 \\
\text{Number of empirical articles per year} &= 90 \times 0.8 = 72 \\
\text{Number of articles per year with new significant relationship} &= 72 \times \frac{1}{2} = 36 \\
\text{Number of articles with a Type I error} &= 36 \times 0.05 = 1.8
\end{aligned}$$

So approximately 2 articles per year will be published on the basis of an significant result attributable to a Type I error. (Unfortunately we can not tell, without more data, which 2 articles they are!)

Chapter 8
Modelling asset return volatility: Introduction

8.1 Introduction

Risk plays a central role in financial decision making, and it is thus no surprise that a great deal of effort has been devoted to the study of the volatility of asset returns. This effort has paid large dividends: volatility modelling and forecasting methods have been shown to be very useful in many economic applications. In this chapter we will cover some of the most widely-used models for modelling volatility, discuss the estimation of these models, and methods of testing for volatility predictability.

8.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce some widely-used models for volatility
- Discuss the estimation of volatility models
- Discuss methods for testing for volatility predictability

8.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Describe some of the most popular models for asset return volatility.
- Explain how to test for the presence of 'volatility clustering' in asset returns.
- Describe how to estimate a volatility model using maximum likelihood.

8.1.3 Essential reading

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487]. Chapter 4.

8.1.4 Further reading

- Diebold, F.X. *Elements of Forecasting.* (Thomson South-Western, Canada, 2006) fourth edition [ISBN 9780324323597]. Chapter 14.
- Taylor, Stephen J. *Asset Price Dynamics, Volatility and Prediction.* (Princeton University Press, Oxford, 2005) [ISBN 9780691134796]. Chapters 8 and 9.
- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354]. Chapter 3.

8.1.5 References cited

- Bollerslev, T. 'Generalized Autoregressive Conditional Heteroskedasticity,' *Journal of Econometrics*, 1986, 31(3), pp.307–327.
- Engle, R. F., 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation,' *Econometrica*, 1982, 50(4), pp.987–1008
- McLeod, A. I. and W. K. Li, 'Diagnostic checking of ARMA time series models using squared residual autocorrelations,' *Journal of Time Series Analysis*, 1983, 4, pp.269–273.

8.2 Implications of homoskedasticity

Consider a standard ARMA(1,1) model for an asset return:

$$\begin{aligned} Y_{t+1} &= \phi_0 + \phi_1 Y_t + \varepsilon_{t+1} + \theta \varepsilon_t \\ \varepsilon_{t+1} &\sim WN\left(0, \sigma^2\right) \end{aligned}$$

which implies that

$$\begin{aligned} V_t\left[Y_{t+1}\right] &= V_t\left[\phi_0 + \phi_1 Y_t + \varepsilon_{t+1} + \theta \varepsilon_t\right] \\ &= V_t\left[\varepsilon_{t+1}\right], \text{ since the other terms are known at time } t \\ &= \sigma^2 \text{ by the assumption } \varepsilon_{t+1} \sim WN\left(0, \sigma^2\right) \end{aligned}$$

So in standard models $V_t\left[Y_{t+1}\right] = \sigma^2$, a constant, which means that Y_{t+1} is (conditionally) *homoskedastic.*

The squared residual can reveal information about volatility, and just as we considered the time series Y_t we can consider the series of squared residuals, ε_t^2, as a time series. So let

$$\begin{aligned} \varepsilon_{t+1}^2 &= E_t\left[\varepsilon_{t+1}^2\right] + \eta_{t+1}, \eta_{t+1} \sim WN\left(0\right) \\ &= V_t\left[\varepsilon_{t+1}\right] + E_t\left[\varepsilon_{t+1}\right]^2 + \eta_{t+1}, \text{ by the definition of variance} \\ &= V_t\left[\varepsilon_{t+1}\right] + \eta_{t+1}, \text{ since } E_t\left[\varepsilon_{t+1}\right] = 0 \\ &= \sigma^2 + \eta_{t+1}, \text{ since } \varepsilon_{t+1} \sim WN\left(0, \sigma^2\right) \end{aligned}$$

That is, the time series ε_{t+1}^2 can be decomposed into two parts: the conditional mean and a mean-zero white noise innovation series, denoted here η_{t+1}. If the conditional variance truly was constant, what would the ACF of ε_{t+1}^2 look like?

$$\begin{aligned}
\gamma_j &= Cov\left[\varepsilon_{t+1}^2, \varepsilon_{t+1-j}^2\right] \\
&= Cov\left[\sigma^2 + \eta_{t+1}, \sigma^2 + \eta_{t+1-j}\right] \\
&= Cov\left[\eta_{t+1}, \eta_{t+1-j}\right] \\
&= 0 \ \ \forall\ j \neq 0, \ \text{ since } \eta_{t+1} \text{ is white noise}
\end{aligned}$$

Thus if Y_t has constant conditional variance, the ACF of ε_{t+1}^2 would be zero for all lags. Let us now check whether this is true empirically.

8.3 Predictability of asset return volatilities

For illustration purposes we will consider the continuously compounded returns on a few example financial time series. We will use the same series as in the previous chapter:

Name	**Sample period**	T
Euro/USD exchange rate	Jan 4, 1999 - Dec 31, 2009	2767
S&P 500 index	Jan 3, 1980 - Dec 31, 2009	7570
US 3-month T-bill rate	Jan 3, 1989 - Dec 31, 2009	5982

Before examining the conditional variance of these returns we must first capture any dynamics in the conditional mean. From last chapter the optimal models ARMA(p, q) models, for p and q between 0 and 5, when using Schwarz's Bayesian information criterion (BIC) were found to be $(0, 0)$, $(0, 0)$ and $(0, 3)$ for these three series respectively. We will use those models for the conditional mean.

In Figure 8.1 I present the sample autocorrelation functions of the squared residuals from the optimal ARMA(p,q) model for each of these three series according to the BIC. The SACFs of the squared returns on the three assets clearly indicate significant serial correlation. This evidence is particularly strong for the stock index, and weakest for the interest rate (though still significant). This feature of financial time series gained attention with Nobel laureate Robert Engle's 1982 article, and is now one of the generally accepted stylised facts about asset returns: there is a substantial amount of predictability in return *volatility*. (This is known as (the presence of) conditional *heteroskedasticity*.) Studies of conditional volatility have formed a large part of the financial econometrics literature.

If we can somehow capture the predictability in volatility, we may be able to improve our portfolio decisions, risk management decisions, option pricing, amongst other things. We will now turn to a very popular and successful model for conditional variance: the ARCH model.

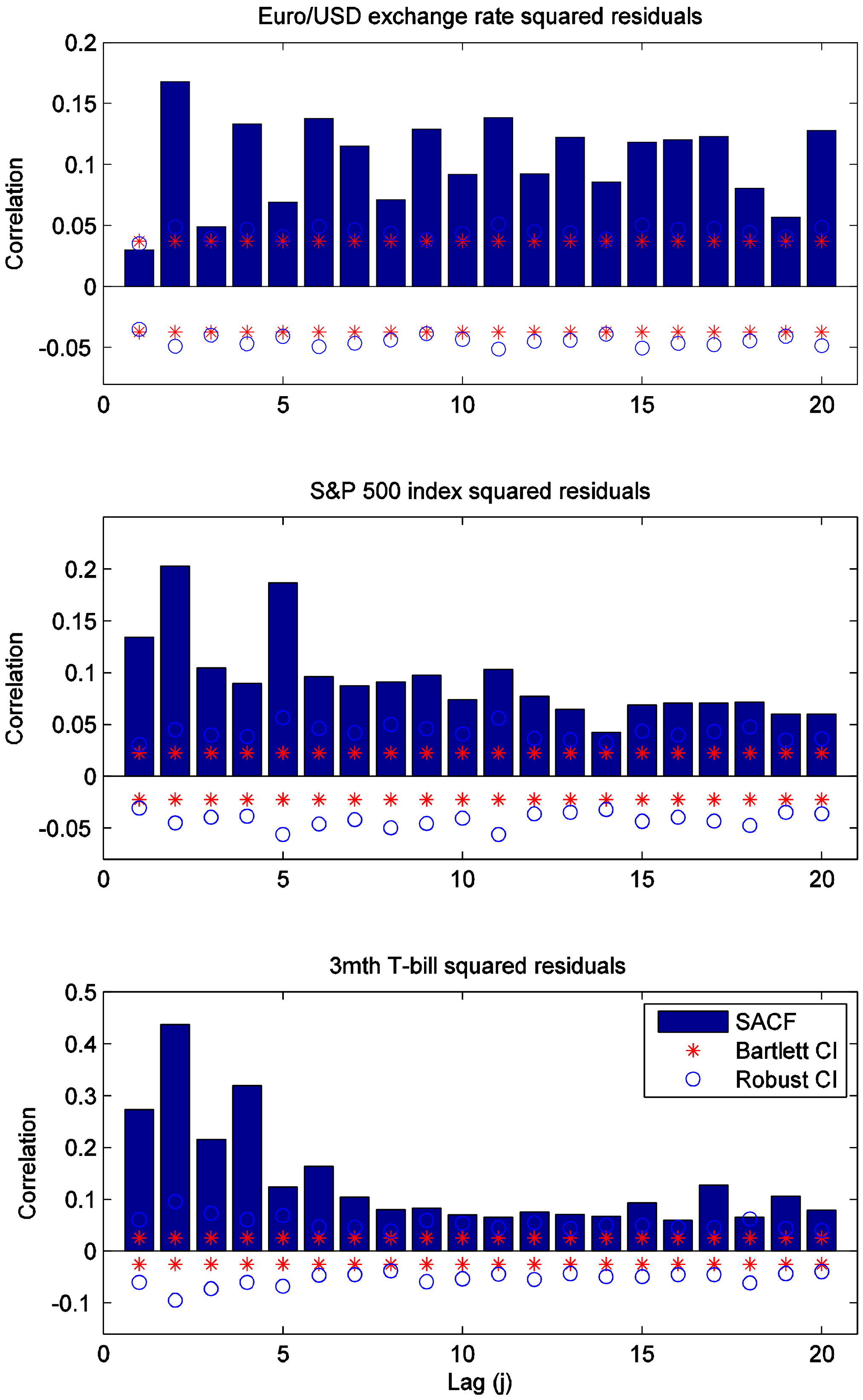


Figure 8.1: Sample autocorrelation functions of the squared residuals for the exchange rate, stock index and T-bill.

8.4 Autoregressive conditional heteroscedasticity (ARCH) processes

From our SACFs for squared returns we saw strong evidence of serial dependence. This suggests that the assumption of constant conditional variance of Y_{t+1}, or equivalently, the assumption of a constant conditional mean for ε_{t+1}^2, is false. If the conditional variance is not constant, how might we model it? A place to start might be an AR model for ε_{t+1}^2:

$$\varepsilon_{t+1}^2 = \omega + \alpha\varepsilon_t^2 + \eta_{t+1},\ \ \eta_{t+1} \sim WN\left(0\right)$$

What would this specification imply for the conditional variance function?

$$\begin{aligned}
\text{Let }\ \sigma_{t+1}^2 &\equiv V_t\left[Y_{t+1}\right] = E_t\left[\varepsilon_{t+1}^2\right] \\
&= E_t\left[\omega + \alpha\varepsilon_t^2 + \eta_{t+1}\right] \\
&= \omega + \alpha\varepsilon_t^2 + 0, \text{ because } \eta_{t+1} \text{ is white noise} \\
\text{so }\ \sigma_{t+1}^2 &= \omega + \alpha\varepsilon_t^2
\end{aligned}$$

The equation above is the famous ARCH(1) model of Engle (1982). It states that the conditional variance of tomorrow's return is equal to a constant, plus some fraction of today's residual squared. This a simple and powerful model for capturing the predictability in volatility.

If an AR(1) model for ε_{t+1}^2 leads to an ARCH(1) model for the conditional variance, what would a more flexible ARMA(1,1) model for ε_{t+1}^2 lead to?

$$\varepsilon_{t+1}^2 = \omega + \gamma\varepsilon_t^2 + \lambda\eta_t + \eta_{t+1}, \eta_{t+1} \sim WN\left(0\right)$$

What would this specification imply for the conditional variance function?

$$\begin{aligned}
\text{Let }\ \sigma_{t+1}^2 &\equiv V_t\left[Y_{t+1}\right] = E_t\left[\varepsilon_{t+1}^2\right] \\
&= E_t\left[\omega + \gamma\varepsilon_t^2 + \eta_{t+1} + \lambda\eta_t\right] \\
&= \omega + \gamma\varepsilon_t^2 + \lambda\eta_t + 0, \text{ since } E_t\left[\eta_{t+1}\right] = 0 \\
&= \omega + \gamma\varepsilon_t^2 + \lambda\left(\varepsilon_t^2 - E_{t-1}\left[\varepsilon_t^2\right]\right), \text{ substituting in for } \eta_t \\
&= \omega + \gamma\varepsilon_t^2 + \lambda\left(\varepsilon_t^2 - \sigma_t^2\right) \\
&= \omega + \left(\gamma + \lambda\right)\varepsilon_t^2 - \lambda\sigma_t^2 \\
\text{so }\ \sigma_{t+1}^2 &= \omega + \alpha\varepsilon_t^2 + \beta\sigma_t^2
\end{aligned}$$

where we redefine the coefficients as $\alpha = (\gamma + \lambda)$ and $\beta = -\lambda$.

This is the famous GARCH(1,1) model due to Bollerslev (1986). The ARMA(1,1)-GARCH(1,1) model is a work-horse in financial time series analysis, which we can now write as:

$$\begin{aligned}
Y_{t+1} &= \mu_{t+1} + \varepsilon_{t+1},\ \ \varepsilon_{t+1} \sim WN\left(0, \sigma_{t+1}^2\right) \\
\mu_{t+1} &= E_t\left[Y_{t+1}\right] = \phi_0 + \phi_1 Y_t + \lambda\varepsilon_t \\
\sigma_{t+1}^2 &= V_t\left[Y_{t+1}\right] = \omega + \alpha\varepsilon_t^2 + \beta\sigma_t^2
\end{aligned}$$

We will cover the estimation of the parameters of this model in Section 8.6 below.

> **Activity 8.1** Show that an AR(p) model for ε_{t+1}^2 leads to an ARCH(p) model for the conditional variance.

8.5 Stationarity, moments, and restrictions on parameters

In this course we will generally focus on 'covariance stationary' time series, which are those where:

$$\begin{aligned} E\left[Y_t\right] &= \mu \; \forall \; t \\ V\left[Y_t\right] &= \sigma_y^2 \; \forall \; t \\ Cov\left[Y_t, Y_{t+j}\right] &= \gamma_j \; \forall \; t, \; j \end{aligned}$$

and so unconditional second moments are assumed to be constant through time. GARCH models are used for the conditional variance, but without further restrictions they can lead to a violation of covariance stationarity. The requirements for a GARCH(1,1) process to be covariance stationary are given below. Let

$$\begin{aligned} Y_{t+1} &= \mu_{t+1} + \varepsilon_{t+1} \\ \varepsilon_{t+1} &= \sigma_{t+1}\nu_{t+1} \\ \nu_{t+1}|\mathcal{F}_t &\sim F\left(0,1\right) \\ \sigma_{t+1}^2 &= \omega + \beta\sigma_t^2 + \alpha\varepsilon_t^2 \end{aligned}$$

then we need:

$$\begin{aligned} \text{Condition 1} &: \omega > 0, \;\; \alpha, \beta \geq 0, \text{ for positive variance} \\ \text{Condition 2} &: \beta = 0 \text{ if } \alpha = 0, \text{ for identification} \\ \text{Condition 3} &: \alpha + \beta < 1, \text{ for covariance stationarity} \end{aligned}$$

If $\alpha + \beta < 1$ then

$$\begin{aligned} E\left[\sigma_{t+1}^2\right] &\equiv \bar{\sigma}^2 = \omega + \beta E\left[\sigma_t^2\right] + \alpha E\left[\varepsilon_t^2\right] \\ &= \omega + \beta E\left[\sigma_t^2\right] + \alpha E\left[E_{t-1}\left[\varepsilon_t^2\right]\right] \\ &= \omega + \beta E\left[\sigma_t^2\right] + \alpha E\left[\sigma_t^2\right] \\ \text{so } \; E\left[\sigma_{t+1}^2\right] &= E\left[\sigma_t^2\right] = \frac{\omega}{1-\alpha-\beta} \end{aligned}$$

This quantity is interesting because:

$$E\left[\sigma_t^2\right] = E\left[E_{t-1}\left[\varepsilon_t^2\right]\right] = E\left[\varepsilon_t^2\right] = V\left[\varepsilon_t\right]$$

the unconditional variance of the residuals, ε_t. Note that this is only one part of the unconditional variance of Y_t :

$$\begin{aligned} \mu &= E\left[Y_t\right] = E\left[E_{t-1}\left[Y_t\right]\right] = E\left[\mu_t\right] \\ \sigma_y^2 &= V\left[Y_t\right] \\ &= V\left[\mu_t + \varepsilon_t\right] \\ &= V\left[\mu_t\right] + V\left[\varepsilon_t\right] + 2Cov\left[\mu_t, \varepsilon_t\right] \\ &= V\left[\mu_t\right] + V\left[\varepsilon_t\right] \\ &= V\left[\mu_t\right] + E\left[\sigma_t^2\right] \end{aligned}$$

and so the unconditional variance of the returns is equal to the sum of the unconditional variance of the conditional mean term and the unconditional variance of the innovation term (which was computed above for the GARCH(1,1) case). If the conditional mean is constant, then the first term is zero and the unconditional variance of returns is equal to the unconditional variance of the residuals. In practice, the conditional mean of asset returns varies much less (at high frequencies, at least) than the average variance of the residual, and so the total variance is close to the variance of the residual. (The ratio of the second term to the first term depends on the variables included in the model for the conditional mean, but has been estimated at something between 100 and 700 for daily stock returns, around 250 for monthly stock returns, around 14 for annual stock returns and around 2 for 4-year returns. Note that this pattern is *not* consistent with a simple ARMA-type model for daily stock returns.)

> **Activity 8.2** Derive an expression for the unconditional variance of a GARCH(2,2) process, assuming that the process is covariance stationary.

> **Activity 8.3** a. Show that a stationary GARCH(1,1) model can be re-written as a function of the unconditional variance, $\sigma_y^2 = E\left[\varepsilon_t^2\right]$, and the deviations of the lagged conditional variance and lagged squared residual from the unconditional variance.
> b. Using the alternative expression for a GARCH(1,1) from part (a), derive the two-step ahead predicted variance for a GARCH(1,1) as a function of the parameters of the model and the one-step forecast. That is, let $\sigma_{t+1}^2 \equiv \sigma_{t+1,t}^2 = E_t\left[\varepsilon_{t+1}^2\right]$, and derive $\sigma_{t+2,t}^2 = E_t\left[\varepsilon_{t+2}^2\right]$ as a function of (ω, α, β) and $\sigma_{t+1,t}^2$.
> c. Derive the two-step ahead predicted variance for a GARCH(1,1), $\sigma_{t+3,t}^2 = E_t\left[\varepsilon_{t+3}^2\right]$, and infer the general expression for a h-step ahead forecast, $\sigma_{t+h,t}^2$. In what financial application might we be interested in a h-step ahead forecast?

8.6 Maximum likelihood estimation of GARCH models

GARCH models are relatively easily estimated if we are willing to make an assumption about the distribution of ε_{t+1}, which allows us to employ maximum likelihood. The most common distributional assumption is that of normality:

$$\varepsilon_{t+1}|\mathcal{F}_t \sim N\left(0, \sigma_{t+1}^2\right)$$

Combined with some model for the conditional mean, say an ARMA(1,1), this implies that the time series is conditionally normally distributed:

$$\begin{aligned}
Y_{t+1} &= \mu_{t+1} + \varepsilon_{t+1} \\
\mu_{t+1} &= E_t\left[Y_{t+1}\right] = \phi_0 + \phi_1 Y_t + \lambda\varepsilon_t \\
\sigma_{t+1}^2 &= V_t\left[Y_{t+1}\right] = \omega + \alpha\varepsilon_t^2 + \beta\sigma_t^2 \\
Y_{t+1}|\mathcal{F}_t &\sim N\left(\phi_0 + \phi_1 Y_t + \theta\varepsilon_t, \omega + \alpha\varepsilon_t^2 + \beta\sigma_t^2\right)
\end{aligned}$$

We can estimate the unknown parameters $\theta \equiv [\phi_0, \phi_1, \lambda, \omega, \alpha, \beta]'$ by maximum likelihood. To do this, we first obtain the joint density of a sequence of observations $(y_1, ..., y_T)$. The approach for doing so differs from the *iid* example in the previous section as these observations are *serially dependent* (indeed, this serial dependence is what the ARMA and GARCH models are designed to capture). With serially dependent data it is useful to break the joint density into the product of conditional densities (recall that $f_{xy}(x,y) = f_{x|y}(x|y)\, f_y(y)$ and that $f_{xyz}(x,y,z) = f_{x|yz}(x|y,z)\, f_{yz}(y,z) = f_{x|yz}(x|y,z)\, f_{y|z}(y|z)\, f_z(z)$):

$$\begin{aligned}
\mathcal{L}(\theta|y_1,y_2,...,y_t) &= f(y_1,...,y_T) \\
&= f_{y_1}(y_1) \times f_{y_2|y_1}(y_2|y_1) \times ... \times f_{y_T|y_{T-1},...,y_1}(y_T|y_{T-1},...,y_1) \\
&= f_{y_1}(y_1) \times \prod_{t=2}^{T} f_{y_t|y_{t-1},...,y_1}(y_t|y_{t-1},...,y_1)
\end{aligned}$$

We decompose the joint density into the product of conditional densities because the structure of the model tells us that the conditional densities are all Normal. Thus $f_{y_2|y_1},...,f_{y_T|y_{T-1},...,y_1}$ are all just Normal densities with different means and variances. We do not know the unconditional distribution of y_1 (it is *not* Normal). What is commonly done is to maximise the *conditional* likelihood, conditioning on the first observation. The conditional likelihood is:

$$\begin{aligned}
f(y_2,...,y_T|y_1;\theta) &= \prod_{t=2}^{T} f_{y_t|y_{t-1},...,y_1}(y_t|y_{t-1},...,y_1;\theta) \\
&= \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \cdot \exp\left\{-\frac{\varepsilon_t^2}{2\sigma_t^2}\right\} \\
\text{where } \varepsilon_t &= Y_t - \phi_0 - \phi_1 Y_{t-1} - \lambda\varepsilon_{t-1} \\
\sigma_t^2 &= \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \\
\text{and we assume } \varepsilon_1 &= 0 \\
\sigma_1^2 &= \frac{\omega}{1-\alpha-\beta}
\end{aligned}$$

So the conditional log-likelihood is:

$$\begin{aligned}
\frac{1}{T-1}\log f(y_2,...,y_T|y_1;\theta) &= \frac{1}{T-1}\log\mathcal{L}(\theta|y_1,y_2,...,y_T) \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2(T-1)}\sum_{t=2}^{T}\log\sigma_t^2 \\
&\quad -\frac{1}{2(T-1)}\sum_{t=2}^{T}\frac{\varepsilon_t^2}{\sigma_t^2}
\end{aligned}$$

and the MLE is

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta}\ \log\mathcal{L}(\theta|y_1,y_2,...,y_T) \\
&= \arg\max_{\theta}\ \left\{-\frac{1}{2}\log(2\pi) - \frac{1}{2(T-1)}\sum_{t=2}^{T}\log\sigma_t^2 - \frac{1}{2(T-1)}\sum_{t=2}^{T}\frac{\varepsilon_t^2}{\sigma_t^2}\right\}
\end{aligned}$$

No analytical solution for the maximum likelihood estimates is available and so we instead use numerical methods to maximise the likelihood. This can be done using Matlab, or one of many other standard econometric software packages such as EViews.

It should be pointed out that there are other ways we could deal with ε_1 and σ_1^2. Above we simply set them equal to their unconditional means. As the sample size grows $(T \to \infty)$ the impact of the treatment of the first observation becomes vanishingly small, and in fact the asymptotic distribution of the conditional MLE is identical to that of the exact MLE.

8.7 Testing for volatility clustering

Before going to the trouble of specifying and estimating a volatility model it is a good idea to test for the presence of volatility clustering the data. Two simple tests are available. McLeod and Li (1983) suggest using the Ljung-Box test on the squared residuals (or squared returns, if they have conditional mean zero) to test jointly for evidence of serial correlation. This test can be implemented using the description of the Ljung-Box test given in the previous chapter.

An alternative, similar, test is the ARCH test of Engle (1982), which involves running the regression:

$$e_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + ... + \alpha_L e_{t-L}^2 + u_t$$

and then performing a χ_L^2 test of the hypothesis[1]:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_L = 0$$

Note that robust standard errors should be used in this regression, making this a robust test for volatility clustering.

If we apply the ARCH test to the three data series considered above, using the ARMA(p,q) models suggested by the BIC model selection criterion for the conditional mean, we obtain the following test statistics:

ARCH tests for volatility clustering

	Lag length, L		
Series	5	10	20
95% Critical value	*11.07*	*18.31*	*31.41*
Euro/USD exchange rate	26.59*	62.76*	163.12*
S&P 500 index	205.39*	568.71*	1040.20*
US 3-month T-bill rate	186.26*	208.33*	430.61*

Thus we have significant statistical evidence of volatility clustering in all three time series. Having found significant evidence of volatility clustering, we now estimate a GARCH(1,1) model on each of these series. The results are presented below. (I have scaled the intercept, ω, by 10^3 as the estimates are all zero to three decimal places.) Notice that the estimated α and β for the T-bill residuals sum to 1, which violates the

[1] Engle (1982) presented this test as a Lagrange multiplier test and it is sometimes known as the ARCH LM test. The way we have discussed it here involves a Wald test rather than a LM test, since we estimate the model under the alternative. This implementation is the simplest and most common, and we still call this test Engle's ARCH test.

condition for stationarity. Some computer software (eg, Matlab) may still return standard errors for estimated parameters, and these are reported below, but they are not reliable in this case. Further, we cannot compute $\bar{\sigma}^2$ when $\alpha + \beta = 1$. We will see below whether a different volatility model for these series works better.

GARCH(1,1) parameter estimates

	$\omega \times 10^3$	α	β	$\alpha + \beta$	$\sqrt{252\bar{\sigma}^2}$
Euro/USD exchange rate	1.173	0.029	0.969	0.998	11.294
	(0.723)	(0.004)	(0.005)		
S&P 500 index	11.560	0.072	0.920	0.992	18.854
	(4.540)	(0.020)	(0.021)		
US 3-month T-bill rate	0.002	0.165	0.835	1.000	–
	(0.001)	(0.023)	(0.021)		

In Figure 8.2 we present the annualised conditional standard deviations produced by the estimated models. We see that these three series have quite different average levels of volatility, with the exchange rate volatility fluctuating around 10% (annualised), stock volatility around 20%, and T-bill volatility around 0.2%. For all three series we observe that the 2007-08 financial crisis led to very high volatility, particularly when compared with earlier that decade when volatility was low by historical standards. For the S&P 500 index we see that the heightened volatility during the financial crisis was longer-lived, but not as large, as the volatility around the October 19, 1987 stock market crash ('Black Monday'). T-bill returns experienced both higher volatility and longer-lived high volatility in the recent financial crisis compared with the stock market crash of 1987.

8.8 Overview of chapter

This chapter presented some of the most widely-used models for modelling volatility. We discussed the use of maximum likelihood estimation of GARCH models, and we covered tests for volatility predictability.

8.9 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

1. Describe some of the most popular models for asset return volatility
2. Explain how to test for the presence of 'volatility clustering' in asset returns
3. Describe how to estimate a volatility model using maximum likelihood.

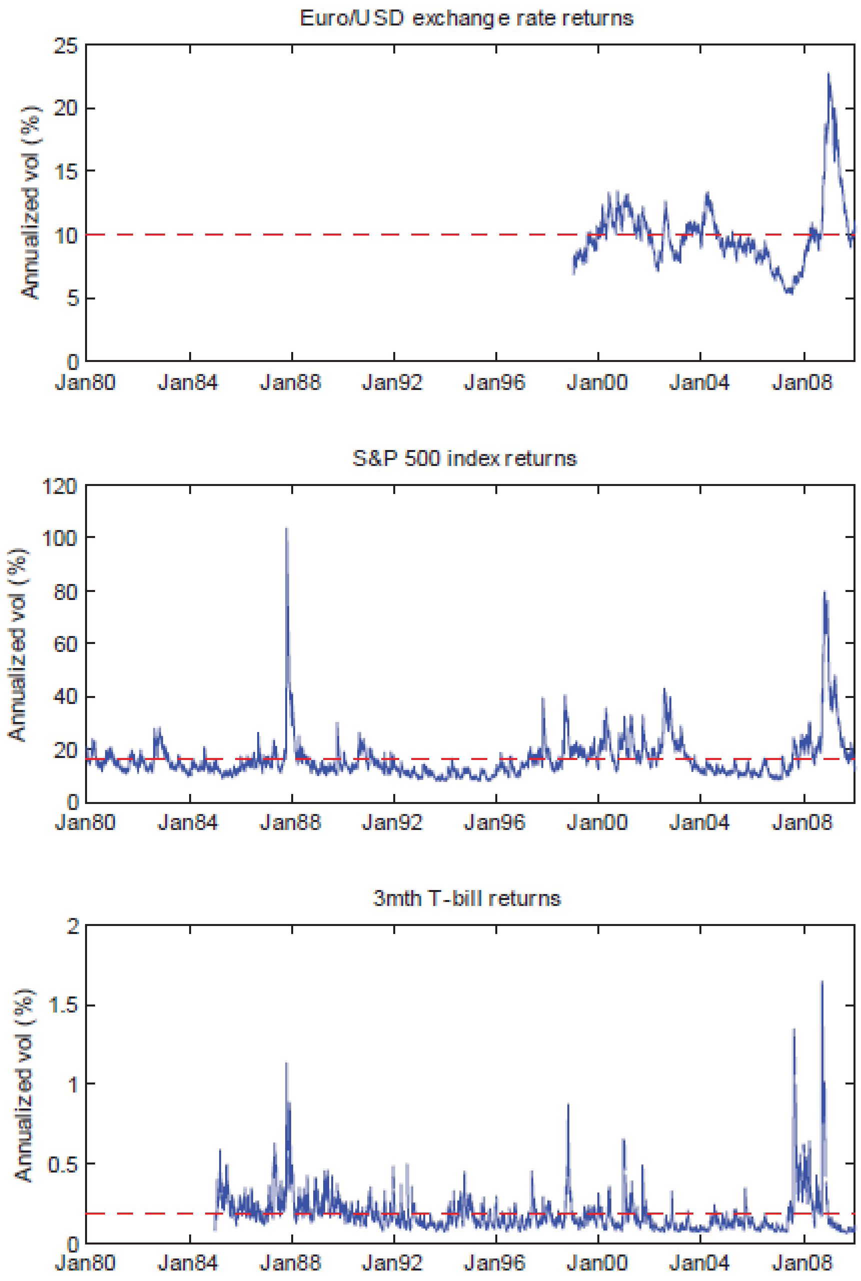


Figure 8.2: Time series of daily conditional volatility (in annualised percent) for the euro/US dollar exchange rate (Jan 99-Dec 2009), S&P 500 index (Jan 1980 - Dec 2009) and the 3-month T-bill (Jan 1985 - Dec 2009). Average annualized volatility is plotted as a dashed line.

8.10 Test your knowledge and understanding

1. a. Assume that

$$\begin{aligned} Y_{t+1} &= \phi_0 + \phi_1 Y_t + \varepsilon_{t+1} \\ \varepsilon_{t+1} &\sim iid\ N\left(0, \sigma^2\right) \end{aligned}$$

Find the first-order autocorrelation of ε_{t+1}^2, and interpret.

b. If we instead assume an AR(1) for ε_{t+1}^2, what would we obtain for the conditional variance of Y_{t+1} from part (a)? Why would we employ such a model?

2. Describe one graphical method and one formal test for detecting volatility clustering in a time series of asset returns.

3. Don't forget to check the VLE for additional practice problems for this chapter.

8.11 Solutions to activities

Activity 8.1

Show that an AR(p) model for ε_{t+1}^2 leads to an ARCH(p) model for the conditional variance.

$$\begin{aligned} Y_{t+1} &= \phi_0 + \phi_1 Y_t + \varepsilon_{t+1} + \theta \varepsilon_t \\ \varepsilon_{t+1}^2 &= \omega + \alpha_1 \varepsilon_t^2 + \alpha_2 \varepsilon_{t-1}^2 + ... + \alpha_p \varepsilon_{t+1-p}^2 + \eta_{t+1} \\ \eta_{t+1} &\sim WN(0) \end{aligned}$$

Then, following the notes:

$$\begin{aligned} V_t\left[Y_{t+1}\right] &= E_t\left[\varepsilon_{t+1}^2\right] \equiv \sigma_{t+1}^2, \text{ so} \\ \sigma_{t+1}^2 &= E_t\left[\omega + \alpha_1 \varepsilon_t^2 + \alpha_2 \varepsilon_{t-1}^2 + ... + \alpha_p \varepsilon_{t+1-p}^2 + \eta_{t+1}\right] \\ &= \omega + \alpha_1 \varepsilon_t^2 + \alpha_2 \varepsilon_{t-1}^2 + ... + \alpha_p \varepsilon_{t+1-p}^2 + 0, \text{ because } \eta_{t+1} \text{ is white noise} \\ \sigma_{t+1}^2 &= \omega + \alpha_1 \varepsilon_t^2 + \alpha_2 \varepsilon_{t-1}^2 + ... + \alpha_p \varepsilon_{t+1-p}^2 \end{aligned}$$

which is the ARCH(p) model of Engle (1982).

Activity 8.2

Derive an expression for the unconditional variance of a GARCH(2,2) process, assuming that the process is covariance stationary.

$$\begin{aligned}
\sigma^2_{t+1} &= \omega + \beta_1\sigma^2_t + \beta_2\sigma^2_{t-1} + \alpha_1\varepsilon^2_t + \alpha_2\varepsilon^2_{t-1} \\
E\left[\sigma^2_{t+1}\right] &= \omega + \beta_1 E\left[\sigma^2_t\right] + \beta_2 E\left[\sigma^2_{t-1}\right] + \alpha_1 E\left[\varepsilon^2_t\right] + \alpha_2 E\left[\varepsilon^2_{t-1}\right] \\
&= \omega + \beta E\left[\sigma^2_t\right] + \beta_2 E\left[\sigma^2_{t-1}\right] \\
&\quad +\alpha_2 E\left[E_{t-1}\left[\varepsilon^2_t\right]\right] + \alpha_2 E\left[E_{t-2}\left[\varepsilon^2_{t-1}\right]\right] \\
&= \omega + \beta_1 E\left[\sigma^2_t\right] + \beta_2 E\left[\sigma^2_t\right] + \alpha_1 E\left[\sigma^2_t\right] + \alpha_2 E\left[\sigma^2_t\right] \\
\text{so } E\left[\sigma^2_t\right] &= \frac{\omega}{1-\beta_1-\beta_2-\alpha_1-\alpha_2}
\end{aligned}$$

Actvity 8.3

a.

$$\begin{aligned}
\sigma^2_{t+1} &= \omega + \beta\sigma^2_{t,t-1} + \alpha\varepsilon^2_t \\
&= \omega + (\alpha+\beta)\sigma^2_y + \beta\left(\sigma^2_{t,t-1} - \sigma^2_y\right) + \alpha\left(\varepsilon^2_t - \sigma^2_y\right) \\
&= \sigma^2_y + \beta\left(\sigma^2_{t,t-1} - \sigma^2_y\right) + \alpha\left(\varepsilon^2_t - \sigma^2_y\right) \\
\text{Recalling that } \sigma^2_y &= \frac{\omega}{1-\alpha-\beta}
\end{aligned}$$

This shows that the GARCH(1,1) forecast can be thought of as a weighted average of the unconditional variance, the deviation of last period's forecast from the unconditional variance, and the deviation of last period's squared residual from the unconditional variance.

b. Next we work out the two-step ahead forecast:

$$\begin{aligned}
\sigma^2_{t+2,t} &\equiv E_t\left[\varepsilon^2_{t+2}\right] \\
&= E_t\left[E_{t+1}\left[\varepsilon^2_{t+2}\right]\right] \text{ by the LIE} \\
&= E_t\left[\sigma^2_y + \beta\left(\sigma^2_{t+1,t} - \sigma^2_y\right) + \alpha\left(\varepsilon^2_{t+1} - \sigma^2_y\right)\right] \\
&= \sigma^2_y + \beta\left(\sigma^2_{t+1,t} - \sigma^2_y\right) + \alpha\left(E_t\left[\varepsilon^2_{t+1}\right] - \sigma^2_y\right) \\
&= \sigma^2_y + (\alpha+\beta)\left(\sigma^2_{t+1,t} - \sigma^2_y\right)
\end{aligned}$$

c. Similar to part (b), we can derive that

$$\begin{aligned}
\sigma^2_{t+3,t} &\equiv E_t\left[\varepsilon^2_{t+2}\right] \\
&= E_t\left[E_{t+1}\left[\varepsilon^2_{t+3}\right]\right] \text{ by the LIE} \\
&= E_t\left[\sigma^2_y + (\alpha+\beta)\left(\sigma^2_{t+2,t+1} - \sigma^2_y\right)\right] \\
&= \sigma^2_y + (\alpha+\beta)\left(E_t\left[\sigma^2_{t+2,t+1}\right] - \sigma^2_y\right) \\
&= \sigma^2_y + (\alpha+\beta)\left(E_t\left[\sigma^2_y + \beta\left(\sigma^2_{t+1,t} - \sigma^2_y\right) + \alpha\left(\varepsilon^2_{t+1} - \sigma^2_y\right)\right] - \sigma^2_y\right) \\
&= \sigma^2_y + (\alpha+\beta)\left(\sigma^2_y + \beta\left(\sigma^2_{t+1,t} - \sigma^2_y\right) + \alpha\left(E_t\left[\varepsilon^2_{t+1}\right] - \sigma^2_y\right) - \sigma^2_y\right) \\
&= \sigma^2_y + (\alpha+\beta)^2\left(\sigma^2_{t+1,t} - \sigma^2_y\right)
\end{aligned}$$

Following similar arguments we can then show the general formula:

$$\sigma^2_{t+h,t} = \sigma^2_y + (\alpha+\beta)^{h-1}\left(\sigma^2_{t+1,t} - \sigma^2_y\right), h \geq 1$$

Multi-step volatility forecasts ($h > 1$) are useful when the decision horizon is longer than the measurement horizon. For example, when we measure returns daily (close-to-close returns, for instance), but we want to re-balance our portfolio only once per week. In that case, we would be interested in the variance of our portfolio over the coming five days:

$$\begin{aligned}
V_t\left[\sum_{j=1}^{5} r_{t+j}\right] &= \sum_{j=1}^{5} V_t\left[r_{t+j}\right], \text{ if returns are serially uncorrelated} \\
&\equiv \sum_{j=1}^{5} \sigma^2_{t+j,t} \\
&= \sum_{j=1}^{5} \sigma^2_y + (\alpha+\beta)^{j-1}\left(\sigma^2_{t+1,t} - \sigma^2_y\right), \text{ using the formula above} \\
&= 5\sigma^2_y + \left(\sigma^2_{t+1,t} - \sigma^2_y\right)\sum_{j=1}^{5}(\alpha+\beta)^{j-1} \\
&= 5\sigma^2_y + \left(\sigma^2_{t+1,t} - \sigma^2_y\right)\cdot\frac{1-(\alpha+\beta)^5}{1-(\alpha+\beta)}
\end{aligned}$$

Chapter 9
Modelling asset return volatility: Extensions

9.1 Introduction

In this chapter we discuss extensions of the basic ARCH/GARCH class of models. Univariate extensions have been proposed to capture more detailed features of asset return volatility, such as the so-called 'leverage effect'. We then discuss ways of choosing the 'best' volatility model.

9.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce some of the most useful extensions of the baseline models for asset return volatility
- Show how the squared residual may be used as a 'proxy' for volatility
- Discuss methods for choosing a volatility model in practice

9.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Describe two univariate extensions of the basic GARCH model
- Discuss using of squared residuals as a 'volatility proxy'
- Compare and contrast various methods for choosing one volatility model over another

9.1.3 Essential reading

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487]. Chapter 4.

9.1.4 Further reading

- Taylor, Stephen J. *Asset Price Dynamics, Volatility and Prediction.* (Princeton University Press, Oxford, 2005) [ISBN 9780691134796]. Chapters 9–10. (Harder)
- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354]. Chapter 3.

9.1.5 References cited

- Andersen, T. G., T. Bollerslev, F. X. Diebold and H. Ebens, ‘The Distribution of Realized Stock Return Volatility,’ *Journal of Financial Economics*, 2001, 61, pp.43–76.
- Andersen, T. G., T. Bollerslev, F. X. Diebold and P. Labys, ‘Modeling and Forecasting Realized Volatility,’ *Econometrica*, 2003, 71, pp.579–625.
- Barndorff-Nielsen, O. E. and N. Shephard ‘Econometric analysis of realised covariation: high frequency based covariance, regression and correlation in financial economics,’ *Econometrica*, 2004, 72, pp.885–925.
- Bollerslev, T. ‘Generalized Autoregressive Conditional Heteroskedasticity,’*Journal of Econometrics*, 1986, 31(3), pp.307–327.
- Bollerslev, T. ‘Glossary to ARCH (GARCH),’ in *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle* (eds. T. Bollerslev, J. R. Russell and M. W. Watson), Chapter 8, pp.137-163. Oxford, U.K.: Oxford University Press, 2010.
- Engle, R. F., ‘Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation,’ *Econometrica*, 1982, 50(4), pp.987–1008
- Engle, R. F., and V. K. Ng, ‘Measuring and Testing the Impact of News on Volatility,’ *Journal of Finance*, 1993, 48(5), 1749–1778.
- Engle, R. F., D. M. Lilien and R. P. Robbins, ‘Estimating Time Varying Risk Premia in the Term Structure: The Arch-M Model,’ *Econometrica*, 1987, 55(2), pp.391–407.
- Fleming, J., C. Kirby and B. Ostdiek, ‘The Economic Value of Volatility Timing,’ *Journal of Finance*, 2001, 56(1), pp.329–352.
- Glosten, L. R., R. Jagannathan and D. E. Runkle, ‘On the relation between the expected value and the volatility of the nominal excess return on stocks,’ *Journal of Finance*, 1993, 48(5), pp.1779–1801.
- Hansen, P. R. and A. Lunde, ‘A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?,’ *Journal of Applied Econometrics*, 2005, 20(7), pp.873–889.
- Nelson, D. B., ‘Conditional Heteroskedasticity in Asset Returns: A New Approach,’ *Econometrica*, 1991, 59(2), 347–370.

- West, K. D., H. J. Edison and D. Cho, 1993, 'A Utility Based Comparison of Some Models of Exchange Rate Volatility,' *Journal of International Economics*, 1993, 35, pp.23–46.

9.2 Extensions of the univariate ARCH model

We have already considered the most widely used extension of the ARCH model, namely the Generalized ARCH, or GARCH model of Bollerslev (1986). In this chapter we will consider two important directions for improving the ARCH/GARCH model for asset returns, and in the next chapter we will move on to considering *multivariate* models for volatility.

Recall the notation from the previous chapter: returns are denoted Y_t, the conditional mean of returns given time $t-1$ information is denoted μ_t, the residual is denoted ε_t and the conditional variance of Y_t is denoted σ_t^2:

$$\begin{aligned} Y_t &= \mu_t + \varepsilon_t \\ \mu_t &= E_{t-1}\left[Y_{t+1}\right] \\ \sigma_t^2 &= V_{t-1}\left[Y_t\right] \end{aligned}$$

This chapter explores additional models for σ_t^2.

9.2.1 Models with a 'leverage effect'

Black (1976) was perhaps the first to observe that stock returns are negatively correlated with changes in volatility: that is, volatility tends to rise (or rise more) following bad news (a negative return) and fall (or rise less) following good news (a positive return). This is called the 'leverage effect,' as firms' use of leverage can provide an explanation for this correlation: if a firm uses both debt and equity then as the stock price of the firm falls its debt-to-equity ratio rises. This will raise equity return volatility if the firm's cashflows are constant. Thus negative returns should lead to higher future volatility, and positive returns the opposite. The leverage effect has since been shown to provide only a partial explanation to observed correlation, but the name persists, and 'asymmetric volatility' models are sometimes referred to as models with a 'leverage effect.'

Given the above relationship, we might want to allow negative returns to have a different impact on volatility tomorrow than positive returns. This behaviour cannot be captured by a standard GARCH model:

$$\sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \alpha\varepsilon_t^2$$

which assumes that tomorrow's volatility is quadratic in today's residual, so the sign of today's residual does not matter. The simplest extension to accommodate this relation is the model of Glosten, Jagannathan and Runkle (1993) (so-called GJR-GARCH, sometimes known as Threshold-GARCH):

$$GJR\text{-}GARCH : \sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \alpha\varepsilon_t^2 + \delta\varepsilon_t^2\mathbf{1}\left\{\varepsilon_t < 0\right\}$$

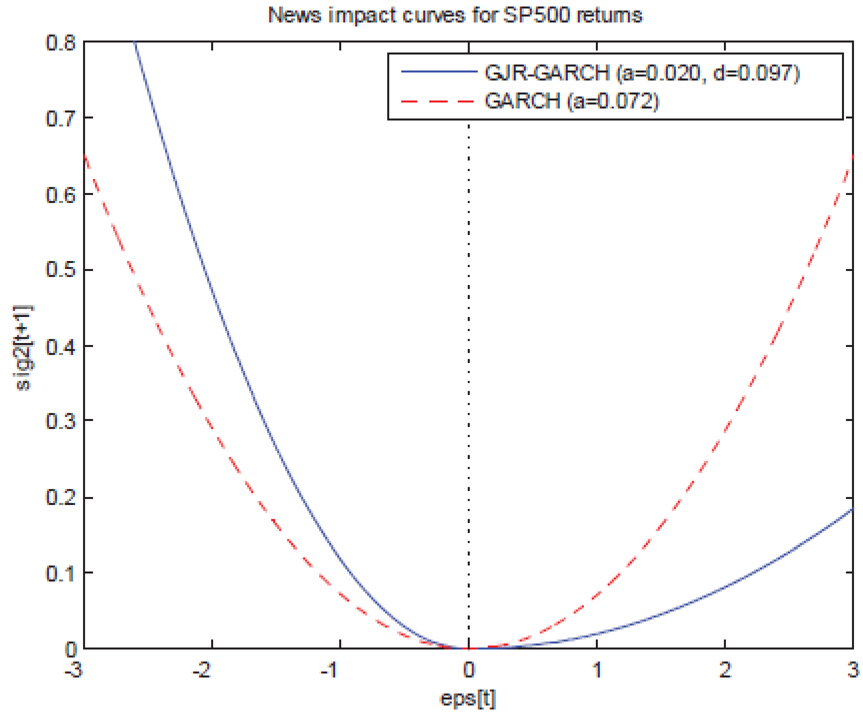


Figure 9.1: The news impact curves of a GJR-GARCH and a standard GARCH model.

If $\delta > 0$ then the impact on tomorrow's volatility of today's residual is greater if today's residual is negative.

One way of illustrating the difference between a volatility models is via their '*news impact curves*', see Engle and Ng (1993). This curve plots σ_{t+1}^2 as the value of ε_t varies, leaving everything else in the model fixed, and normalising the function to equal zero when $\varepsilon_t = 0$. This is illustrated for the S&P 500 index returns in Figure 9.1. The news impact curve for a standard GARCH model is simply the function $\alpha\varepsilon_t^2$, while for the GJR-GARCH model it is $\alpha\varepsilon_t^2 + \delta\varepsilon_t^2\mathbf{1}\left\{\varepsilon_t < 0\right\}$.

The results from estimating a GJR-GARCH(1,1) on the three series analysed in the previous chapter are presented below. From this table we see that the asymmetry parameter, δ, is positive and significant for the stock return, negative and significant for the interest rate, and not significant for the exchange rate. A positive and significant δ indicates that negative shocks lead to higher future volatility than do positive shocks of the same magnitude. This is the pattern observed for the stock index returns, but an opposite relationship holds for the T-bill.

GJR-GARCH(1,1) parameter estimates

	$\omega \times 10^3$	α	β	δ	δ tstat
Euro/USD exchange rate	1.266 (0.739)	0.023 (0.006)	0.969 (0.005)	0.011 (0.008)	1.375
S&P 500 index	15.963 (5.253)	0.020 (0.006)	0.917 (0.017)	0.097 (0.027)	3.593
US 3-month T-bill rate	0.002 (0.001)	0.120 (0.016)	0.839 (0.023)	−0.083 (0.036)	−2.306

The other widely-used GARCH model that allows for leverage effects is the 'exponential GARCH', or 'EGARCH' model of Nelson (1991). Nelson proposed this model to remedy two shortcomings of the standard GARCH model. The first is its inability to capture the leverage effect, and the second is the conditions have to be imposed on the parameters of the GARCH model to ensure a positive volatility estimate. The EGARCH model deals with both:

$$\log \sigma_{t+1}^2 = \omega + \beta \log \sigma_t^2 + \alpha \left| \frac{\varepsilon_t}{\sigma_t} \right| + \gamma \frac{\varepsilon_t}{\sigma_t}$$

By modelling $\log \sigma_{t+1}^2$ rather than σ_{t+1}^2 we are ensured a positive estimate of σ_{t+1}^2. Further by allowing γ to differ from zero the leverage effect can be captured. The parameter estimates for EGARCH(1,1) models estimated on our three data sets are presented below.

EGARCH(1,1) parameter estimates

	ω	α	β	γ	γ tstat
Euro/USD exchange rate	−0.056 (0.008)	0.068 (0.010)	0.996 (0.002)	−0.010 (0.007)	1.429
S&P 500 index	−0.099 (0.018)	0.128 (0.024)	0.984 (0.004)	−0.080 (0.017)	−4.706
US 3-month T-bill rate	−0.481 (0.073)	0.295 (0.003)	0.970 (0.007)	0.061 (0.016)	3.813

The EGARCH model results indicate that the asymmetry parameter, γ in this case, is significant and negative for the stock return, significant and positive for the interest rate, but not significant for the exchange rate. This is similar to the results we obtained using the GJR-GARCH model, when we note from the EGARCH specification that a negative γ implies higher future volatility following a negative shock (which would lead to a positive δ in the GJR-GARCH model).

> **Activity 9.1** Suppose you had been working with a trader in the market for a particular stock, and you observed that participants in this market regarded a return of 0.01 as 'no news', but any return above or below this was 'news' and seemed to lead to higher volatility. Write down an extension of the basic GARCH model that would capture this.

9.2.2 ARCH-in-mean model

A central idea in finance is that the return on a risky security should be positively related to its risk. This led Engle, Lilien and Robins (1987) to develop the 'ARCH in mean' (or 'ARCH-M') model which posits that the conditional mean of a return is dependent on some function of its conditional variance or conditional standard deviation. The particular function of conditional variance that enters the conditional mean (i.e., level of variance, standard deviation, log-variance, etc.) is left to the researcher. Using the standard deviation has the benefit that it is in the same units as returns, and so parameters are unaffected by scale (such as multiplying returns by 100).

If we think that the mean has an AR(1) term, for example, then we might use one of the following models:

$$\begin{aligned} r_{t+1} &= \phi_0 + \phi_1 r_t + \gamma\sigma_{t+1}^2 + \varepsilon_{t+1}, \text{ or} \\ r_{t+1} &= \phi_0 + \phi_1 r_t + \gamma\sigma_{t+1} + \varepsilon_{t+1}, \text{ or} \\ r_{t+1} &= \phi_0 + \phi_1 r_t + \gamma \log\sigma_{t+1} + \varepsilon_{t+1}, \text{ with} \\ \sigma_{t+1}^2 &= \omega + \beta\sigma_t^2 + \alpha\varepsilon_t^2 \end{aligned}$$

Estimating the ARCH-M models above (with the conditional standard deviation added to the mean equation) for the three series studied so far yielded the following parameter estimates. From this table we see that the 'ARCH-in-mean' term is borderline significant (*t-stat* of -1.73) for the exchange rate, and positive and significant for the stock index return and the T-bill return. This is a relatively positive result for this model, as in many applications of the ARCH-M model researchers have found that the volatility term in the mean equation is not significant, or not of the expected sign.

AR(1)-GARCH-M(1,1) parameter estimates

	ϕ_0	ϕ_1	γ	$\omega \times 10^3$	α	β	γ tstat
Euro/USD ex rate	0.084 (0.043)	−0.003 (0.018)	−0.130 (0.075)	1.151 (0.730)	0.030 (0.004)	0.968 (0.005)	−1.733
S&P 500 index	−0.043 (0.029)	0.016 (0.009)	0.080 (0.035)	11.869 (4.619)	0.073 (0.020)	0.919 (0.021)	2.286
US 3-month T-bill rate	−0.001 (0.000)	−0.063 (0.016)	0.020 (0.003)	0.002 (0.001)	0.159 (0.023)	0.841 (0.022)	6.667

9.2.3 NARCH, PARCH, QARCH, STARCH...

The line of research that started with a simple ARCH model and the plain vanilla GARCH(1,1), has since been extended in numerous directions. Some of the many flavours of GARCH models are given below, see Bollerslev (2009) for a review of many more extensions. The notation we use is:

$$\begin{aligned} Y_{t+1} &= \mu_{t+1} + \varepsilon_{t+1} \\ \varepsilon_{t+1} &= \sigma_{t+1}\nu_{t+1} \\ \nu_{t+1}|\mathcal{F}_t &\sim iid\ F(0,1) \end{aligned}$$

IGARCH (Engle and Bollerslev, 1986)

$$\sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + (1-\beta)\,\varepsilon_t^2$$

PARCH (Ding, Granger and Engle, 1993)

$$\sigma_{t+1}^\gamma = \omega + \beta\sigma_t^\gamma + \alpha\varepsilon_t^{2\gamma}$$

A**PARCH** (Ding, Granger and Engle, 1993)

$$\sigma_{t+1}^\gamma = \omega + \beta\sigma_t^\gamma + \alpha\gamma\left(|\varepsilon_t| - \delta\varepsilon_t\right)^\gamma$$

PARCH (Engle and Bollerslev, 1986)

$$\sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \alpha\varepsilon_t^\delta$$

SQR-GARCH (Taylor, 1986 and Schwert, 1989)

$$\sigma_{t+1} = \omega + \beta\sigma_t + \alpha\left|\varepsilon_t\right|$$

QGARCH (Sentana, 1991)

$$\sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \alpha\varepsilon_t^2 + \delta\varepsilon_t$$

NARCH (Higgins and Bera, 1992)

$$\sigma_{t+1}^2 = \left(\phi_0\omega^\delta + \phi_1\varepsilon_t^{2\delta} + \phi_2\varepsilon_{t-1}^{2\delta} + ... + \phi_p\varepsilon_{t-p+1}^{2\delta}\right)^{1/\delta}$$

All-in-the-family GARCH (Hentschel, 1995)

$$\frac{\sigma_{t+1}^\gamma - 1}{\gamma} = \omega + \alpha\sigma_t^\gamma\left[\left|\nu_t - \delta\right| - \lambda\left(\nu_t - \delta\right)\right] + \beta\frac{\sigma_t^\gamma - 1}{\gamma}$$

SQARCH (Ishida and Engle, 2001)

$$\sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \alpha\sigma_t\left(\nu_t^2 - 1\right)$$

9.2.4 Does anything beat a GARCH(1,1)?

In the two decades or so since the first ARCH model was proposed over a dozen extensions have been published. With increased computing power has come increasingly complex volatility models. A very reasonable question to ask is "does anything beat the benchmark GARCH(1,1) volatility model?" A recent paper by Hansen and Lunde (2005) addresses this question. They consider a total of 330 different ARCH-type models for the Deutsche mark-U.S. dollar exchange rate and for IBM equity returns. For the exchange rate they find no evidence against the simple GARCH(1,1).

For the equity return they find that the 'Asymmetric Power GARCH(2,2)', or APARCH(2,2), model performed best. This model is:

$$\sigma_{t+1}^\delta = \omega + \sum_{i=1}^{2}\alpha_i\left(\left|\varepsilon_t\right| - \gamma_i\varepsilon_t\right)^\delta + \sum_{j=1}^{2}\beta_i\sigma_t^\delta$$

The APARCH model is one of the most complicated in use. It allows for a leverage effect (when $\gamma \neq 0$). Allowing δ to differ from 2 enables the model to use the fact that serial correlation in the absolute value of returns to the power δ (with $\delta < 2$) tends to be stronger than that in squared returns. We will discuss the ways Hansen and Lunde (2005) measured the performance of these volatility models below.

9.3 Choosing a volatility model

Above we reviewed some of the numerous volatility models available to a financial forecaster, raising the question of how we choose the ‘best’ model. The choice of volatility model should depend on what we intend to do with it. For example, if we intend to use it for out-of-sample forecasting of volatility, which is perhaps the most common use for a volatility model, then the correct way to choose a volatility model should be according to some measure of its out-of-sample forecast performance. If we instead want to determine whether there is statistical evidence of a leverage effect, then we should pick the model that gives the best in-sample fit and thus (hopefully) the most precise parameter estimates. Below we we will discuss the selection of models based on out-of-sample forecast performance, but first we will focus on choosing a model using in-sample information.

9.3.1 Comparing nested models

If the two models being compared are ‘nested’, in the sense that for certain choices of parameters the models are identical, then we can conduct statistical tests to see if the models are significantly different. The simplest example of this is comparing a GJR-GARCH model with a GARCH model:

$$\begin{aligned} \textit{GJR-GARCH} &: \sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \alpha\varepsilon_t^2 + \delta\varepsilon_t^2\mathbf{1}\left\{\varepsilon_t < 0\right\} \\ \textit{GARCH} &: \sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \alpha\varepsilon_t^2 \end{aligned}$$

When $\delta = 0$ the GJR-GARCH is simply the GARCH model, and so the GJR-GARCH nests the GARCH model. Using the formulas in Section 10.4.2 of Taylor (2005), we can compute the standard errors for the GJR-GARCH parameters and we can test:

$$\begin{aligned} H_0 &: \delta = 0 \\ \text{vs. } H_a &: \delta \neq 0 \end{aligned}$$

If we reject the null hypothesis then we conclude that the GJR-GARCH is significantly better than the GARCH model, at least in-sample. In out-of-sample forecast comparisons it is often the case that more parsimonious models perform best, even if a more flexible model is significantly better in-sample. If the more flexible model is not significantly better in-sample (e.g. if we fail to reject H_0) then it is very unlikely to do better out-of-sample.

9.3.2 Using information criteria

As we noted earlier, measures of performance that do not account for the number of parameters in the model will generally always just pick the largest model. But in forecasting extra parameters can lead to increased estimation error and worsened forecast performance. Thus it is important to incorporate some sort of trade-off between increased goodness-of-fit and increased estimation error. As in Section 4.5, the AIC, HQIC and BIC may be used to accomplish this. The formulas for the AIC, HQIC and BIC given in Chapter 4 were relevant for conditional mean modelling, and below we give their more general versions, which use the value of the log-likelihood at the optimum, $\log \mathcal{L}$, the sample size, T, and the number of parameters, k. (Note that $\log \mathcal{L}$ below denotes the *sum* of the log-likelihood at each point in the sample, so $1/T \log \mathcal{L}$ represents the *mean* log-likelihood.)

$$\begin{aligned} AIC &= -\frac{2}{T}\log \mathcal{L} + \frac{2k}{T} \\ HQIC &= -\frac{2}{T}\log \mathcal{L} + \frac{2k}{T}\log\log T \\ BIC &= -\frac{2}{T}\log \mathcal{L} + \frac{2k}{T}\log \sqrt{T} \end{aligned}$$

As before, the first terms in these expressions represent the goodness-of-fit, and the second terms represent a penalty for extra parameters. We want to choose the volatility model that yields the smallest information criterion.

9.3.3 Using statistical goodness-of-fit measures

When estimating standard econometric models we would usually evaluate goodness-of-fit by the R^2, which corresponds to ranking competing models by their MSE:

$$MSE = \frac{1}{T}\sum_{t=1}^{T} e_t^2$$

where e_t is the residual from the model for the conditional mean. We would choose the model with the highest R^2 , or the lowest MSE. What would be the right way to measure the accuracy of a volatility model?

Evaluating accuracy requires some knowledge of the realised value of the variable of interest, in this case the conditional variance. But the conditional variance is not observable even *ex post*, and so we must instead rely on *volatility proxies*, which we denote $\tilde{\sigma}_t^2$. The simplest volatility proxy is the squared residual (or squared return, if we assume that returns have zero mean):

$$\tilde{\sigma}_t^2 = \varepsilon_t^2$$

A volatility proxy is a variable that is useful for estimating the value of the true volatility. The squared residual can be justified as a volatility proxy because it is a conditionally unbiased estimator of the true conditional variance:

$$E_{t-1}\left[\varepsilon_t^2\right] = E_{t-1}\left[\sigma_t^2 \nu_t^2\right] = \sigma_t^2 E_{t-1}\left[\nu_t^2\right] = \sigma_t^2$$

and so, on average, the squared residual will correctly estimate the true conditional variance. Note, importantly, that while it will be correct on average it will estimate the conditional variance with error. That is, the squared residual is a 'noisy' volatility proxy. Other, less noisy, volatility proxies have gained attention recently, see Andersen, *et al.* (2001, 2003) and Barndorff-Nielsen and Shephard (2004).

If we denote a volatility forecast by h_t (where the h stands for 'heteroskedasticity') and a volatility proxy by $\tilde{\sigma}_t^2$ then two possible goodness-of-fit measures for a volatility forecast are:

$$\begin{aligned} Squared\ error &: \ L\left(\tilde{\sigma}_t^2, h_t\right) = \left(\tilde{\sigma}_t^2 - h_t\right)^2 \\ QLIKE &: \ L\left(\tilde{\sigma}_t^2, h_t\right) = \frac{\tilde{\sigma}_t^2}{h_t} - \log\frac{\tilde{\sigma}_t^2}{h_t} - 1 \end{aligned}$$

and we would rank two forecasts according to their average loss over the forecast period:

$$\bar{L}_i = \frac{1}{T}\sum_{t=1}^{T} L\left(\tilde{\sigma}_t^2, h_t^{(i)}\right)$$

for model i. The 'QLIKE' loss function may be recognised as the central part of the normal log-likelihood for estimating a volatility model. Thus ranking by the average QLIKE loss function is equivalent to ranking by the average (normal) log-likelihood.

In the table below we look at the performance of the three volatility models considered above (the GARCH-in-mean model is a hybrid mean/volatility model and so we do not include it here), using six different metrics. (Note that we want the highest $\log \mathcal{L}$, but the lowest values for the remaining metrics.) We see that the EGARCH model is preferred, using all metrics, for both the stock index and the T-bill. For the exchange rate the GJR-GARCH model is preferred using all metrics except the BIC, which suggests the simple GARCH model.

Choosing a volatility model

	$\log \mathcal{L}$	RMSE	QLIKE	AIC	HQIC	BIC
			Euro/USD exchange rate			
GARCH	-2561.8	0.8624	0.0138	1.8538	1.8562	1.8603*
GJR-GARCH	-2559.5*	0.8618*	0.0121*	1.8529*	1.8560*	1.8615
EGARCH	-2561.2	0.8623	0.0133	1.8541	1.8572	1.8627
			S&P 500 index			
GARCH	-10314	7.0827	0.8871	2.7257	2.7267	2.7285
GJR-GARCH	-10239	7.1241	0.8674	2.7063	2.7076	2.7100
EGARCH	-10225*	7.0067*	0.8634*	2.7024*	2.7036*	2.7060*
			US 3-month T-bill rate			
GARCH	19018	0.000897	-8.1961	-6.3573	-6.3561	-6.3539
GJR-GARCH	19043	0.000894	-8.2045	-6.3653	-6.3638	-6.3609
EGARCH	19056*	0.000878*	-8.2090*	-6.3698*	-6.3683*	-6.3654*

> **Activity 9.2** Show that the optimal forecast under 'squared error' and 'QLIKE' loss is the true conditional variance, when the volatility proxy used is conditionally unbiased.

9.3.4 Using economic goodness-of-fit measures

An alternative to statistical measures of goodness-of-fit is some *economic* measure of goodness-of-fit. Good examples of this are in West, *et al.* (1993) and Fleming, *et al.* (2001), who find that the value of modelling conditional volatility to a risk-averse investor is between 5 and 200 basis points per year. If a conditional variance forecast is going to be used to make some economic decision, such as an investment decision, then the right way to compare the performance of competing models is to compare the profits that are generated by each model. Ideally, we would *always* compare forecasting models by their performance in economic decision-making. Our use of statistical measures of goodness-of-fit is motivated by the simple fact that we generally don't know the various economic uses of the forecasts, and so we instead use general measures of goodness-of-fit.

Portfolio choice

Consider an investor with utility function that is quadratic in future wealth, generating mean-variance preferences, and who can invest in a risky asset and a risk-free asset.

$$\begin{aligned}
\mathcal{U}\left(W_{t+1}\right) &= W_{t+1} - 0.5\gamma W_{t+1}^2 \\
W_{t+1} &= W_t\left(\omega_{t+1}\left(R_{t+1} - R_{t+1}^f\right) + 1\right) \\
W_t &= \text{wealth at time } t \\
\gamma &= \text{risk aversion parameter} \\
R_t &= \text{gross return on risky asset} \\
R_t^f &= \text{gross return on risk-free asset} \\
\omega_{t+1} &= \text{portfolio weight in risky asset}
\end{aligned}$$

If we set $W_t = 1$ we get the following rule for the optimal portfolio weight:

$$\omega_{t+1}^* = \frac{E_t\left[R_{t+1} - R_{t+1}^f\right]}{V_t\left[R_{t+1} - R_{t+1}^f\right] - E_t\left[R_{t+1} - R_{t+1}^f\right]^2} \cdot \frac{1+\gamma}{\gamma}$$

Thus the optimal portfolio weight is a function of a conditional mean and a conditional variance forecast. Combined with some model for the conditional mean, we can use volatility forecasts to obtain optimal portfolio weights at each point in time. From these we can then work out the portfolio returns obtained by using a particular volatility model, and finally work out the realised utility from these returns. The better volatility forecast should yield a higher expected utility over the sample period.

Option pricing*

Another application of volatility models is in option pricing. If we take the simple Black-Scholes formula for pricing a European option on a non-dividend paying stock we get:

$$\begin{aligned}
c_t &= S_t\Phi\left(d_1\right) - K\exp\left\{-r\left(T-t\right)\right\}\Phi\left(d_2\right) \\
p_t &= K\exp\left\{-r\left(T-t\right)\right\}\Phi\left(-d_2\right) - S_t\Phi\left(-d_1\right) \\
d_1 &= \frac{\log\left(S_t/K\right) + \left(r+\sigma^2/2\right)\left(T-t\right)}{\sigma\sqrt{T-t}} \\
d_2 &= d_1 - \sigma\sqrt{T-t}
\end{aligned}$$

where

$$\begin{aligned}
c_t &= \text{European call option price at time } t \\
p_t &= \text{European put option price at time } t \\
S_t &= \text{value of underlying stock at time } t \\
K &= \text{strike price of option contract} \\
r &= \text{risk-free rate (annualised)} \\
T-t &= \text{time (in years) until expiry of contract} \\
\sigma &= \text{volatility of underlying stock price}
\end{aligned}$$

The only unobservable input to the B-S option pricing formula is the stock price volatility. This is where volatility models may be used. If h_t is a forecast of the volatility of the return on the asset between time t and time T, then one way of evaluating the performance of the model (see Christoffersen and Jacobs, 2004) is to look at the pricing errors it generates when used in the B-S pricing formula:

$$\begin{aligned}
\hat{c}_{t,j} &= S_t\Phi\left(\hat{d}_{1,t,j}\right) - K_j \exp\left\{-r\left(T-t\right)\right\}\Phi\left(\hat{d}_{2,t,j}\right) \\
\hat{p}_{t,j} &= K_j \exp\left\{-r\left(T-t\right)\right\}\Phi\left(-\hat{d}_{2,t,j}\right) - S_t\Phi\left(-\hat{d}_{1,t,j}\right) \\
\hat{d}_{1,t,j} &= \frac{\log\left(S_t/K_j\right) + \left(r + h_t/2\right)\left(T-t\right)}{\sqrt{h_t\left(T-t\right)}} \\
\hat{d}_{2,t,j} &= \hat{d}_{1,t,j} - \sqrt{h_t\left(T-t\right)} \\
MSE-call &\equiv \sum_{j=1}^{n}\left(c_{t,j} - \hat{c}_{t,j}\right)^2 \\
MSE-put &\equiv \sum_{j=1}^{n}\left(p_{t,j} - \hat{p}_{t,j}\right)^2
\end{aligned}$$

where $\{c_{t,j}\}_{j=1}^{n}$ and $\{p_{t,j}\}_{j=1}^{n}$ are the option prices available at time t, for options with various strikes prices. It should be noted that there is evidence against the B-S model empirically, and so even if h_t happened to be a perfect volatility forecast it would almost certainly generate pricing errors. Furthermore, time-varying conditional variance is not consistent with the B-S assumptions, and so there is an internal inconsistency with this method of evaluating volatility forecasts. Nevertheless, evaluation of volatility forecasts through their use in option pricing is a useful measure of their goodness-of-fit.

9.4 Overview of chapter

This chapter presented two important extensions of the basic ARCH/GARCH class of models, and discussed various of ways of choosing the 'best' volatility model.

9.5 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Compare and contrast various methods for choosing one volatility model over another
- Describe some useful extensions of the basic GARCH model

9.6 Test your knowledge and understanding

1. Describe one statistical method and one economic method for deciding between two competing volatility models. Discuss the pros and cons of each method: which method would you use?
2. What is a 'news impact curve' and what does it tell us about the volatility of an asset return?
3. A GJR-GARCH model was estimated on daily returns on the FTSE index over the period 1 January 2004 to 31 December 2013, and the following parameters were obtained (standard errors are reported in parentheses):

GJR-GARCH(1,1) model for FTSE returns

	ω	α	β	δ
Parameter estimate (Standard error)	0.0168 (0.0052)	0.0006 (0.0094)	0.9096 (0.0184)	0.1445 (0.0297)

Formally test the null that the GARCH model provides as good a fit to these returns as the GJR-GARCH model. Interpret the results of this test.

4. Don't forget to check the VLE for additional practice problems for this chapter.

9.7 Solutions to activities

Activity 9.1

If a return of 0.01 was really 'no news', then that would mean that the 'news impact curve' should be lowest when $\varepsilon_t = 0.01$, rather than when $\varepsilon_t = 0$ in the standard model. One possible model for this situation would be an asymmetric GARCH model:

$$\sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \alpha\left(\varepsilon_t - 0.01\right)^2$$

If the minimum point was not known then we could generalise this model to allow that point to be estimated:

$$\sigma_{t+1}^2 = \omega + \beta\sigma_t^2 + \alpha\left(\varepsilon_t - \gamma\right)^2$$

where γ is the new parameter. This model is equivalent to the QGARCH model of Sentana (1991):

$$\begin{aligned}\sigma_{t+1}^2 &= \omega + \beta\sigma_t^2 + \alpha\left(\varepsilon_t - \gamma\right)^2 \\ &= \tilde{\omega} + \tilde{\beta}\sigma_t^2 + \tilde{\alpha}\varepsilon_t^2 + \tilde{\delta}\varepsilon_t\end{aligned}$$

where $\tilde{\omega} = \omega + \alpha\gamma^2$, $\tilde{\beta} = \beta$, $\tilde{\alpha} = \alpha$, and $\tilde{\delta} = -2\alpha\gamma$.

Activity 9.2

Optimal forecast:

$$h_t^* \equiv \arg\min_h \; E_{t-1}\left[L\left(\tilde{\sigma}_t^2, h\right)\right]$$

Under squared-error loss this is

$$h_t^* \equiv \arg\min_h \; E_{t-1}\left[\left(\tilde{\sigma}_t^2 - h\right)^2\right]$$

with first-order condition (FOC):

$$\begin{aligned} 0 &= -2E_{t-1}\left[\tilde{\sigma}_t^2 - h_t^*\right] \\ \text{so } h_t^* &= E_{t-1}\left[\tilde{\sigma}_t^2\right] \end{aligned}$$

And since $\tilde{\sigma}_t^2$ is assumed to be conditionally unbiased, we know that $E_{t-1}\left[\tilde{\sigma}_t^2\right] = \sigma_t^2$, and so

$$h_t^* = \sigma_t^2$$

Thus the optimal forecast under squared error loss, using any conditionally unbiased volatility proxy is the true conditional variance. This means that if we use this loss function to compare the true conditional variance with **any** other variance forecast we are guaranteed to find that the true conditional variance is the best, which is precisely what we would like.

We proceed in a similar fashion for the QLIKE loss function:

$$\begin{aligned} h_t^* &\equiv \arg\min_h \; E_{t-1}\left[\frac{\tilde{\sigma}_t^2}{h} - \log\frac{\tilde{\sigma}_t^2}{h} - 1\right] \\ \text{FOC} \quad 0 &= E_{t-1}\left[\frac{1}{h_t^*} - \frac{\tilde{\sigma}_t^2}{\left(h_t^*\right)^2}\right] \\ &= E_{t-1}\left[h_t^* - \tilde{\sigma}_t^2\right], \;\text{ since } h_t^* > 0 \text{ we can multiply through by } \left(h_t^*\right)^2 \\ \text{so } h_t^* &= E_{t-1}\left[\tilde{\sigma}_t^2\right] \\ &= \sigma_t^2, \;\text{ since } E_{t-1}\left[\tilde{\sigma}_t^2\right] = \sigma_t^2 \end{aligned}$$

And so again the optimal forecast is the true conditional variance.

Chapter 10
Multivariate volatility models

10.1 Introduction

Most financial decisions involve considering the risks of multiple assets, not just a single asset, and in this chapter we consider models that can help with such decisions. We will consider two popular models for the conditional covariance matrix of a vector of asset returns, and discuss how these models can be compared via an application to portfolio decisions.

10.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce the problem of modelling covariance matrices
- Discuss two widely-used covariance matrix models
- Consider an application to a portfolio decision problem

10.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Describe the two main problems in multivariate volatility modelling
- Compare and contrast two popular models for covariance matrices
- Derive the mean and variance for a portfolio return from the mean vector and covariance matrix of the underlying assets

10.1.3 Essential reading

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487]. Chapter 7, Sections 1–2.

10.1.4 Further reading

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487]. Chapter 7, Sections 3–5.

- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354]. Chapter 10. (Harder)

10.1.5 References cited

- Bollerslev, T. 'Modeling the Coherence in Short-run Nominal Exchange Rates: A Multivariate Generalized ARCH Model,' *Review of Economics and Statistics*, 1990, 72(3), pp.498–505.

10.2 Modelling covariance matrices

Univariate volatility models have received a lot of attention in the financial econometrics literature, and a wide variety of models have been proposed and tested empirically. In many (perhaps most) economic decisions, however, we require an estimate of the full conditional covariance matrix of a *collection* of assets, not just the individual variances. Examples include portfolio decisions, risk management decisions involving many risk positions (e.g., risk management at the 'group level' rather than at the 'desk level'), pricing derivatives with multiple underlying assets (e.g., basket derivatives) and statistical arbitrage trading (predicting the *relative* moves in the prices of two or more assets).

Multivariate volatility modelling is not much more difficult than univariate modelling theoretically, but it poses a number of practical problems. In this chapter we will review a few widely-used multivariate volatility models, focussing on those from the ARCH class, and discuss the main practical problems in multivariate volatility modelling.

The set up is as follows:

$$\begin{aligned}\underset{(k\times1)}{\mathbf{r}_{t+1}} &= \underset{(k\times1)}{\mu_{t+1}} + \underset{(k\times1)}{\varepsilon_{t+1}}\\ \underset{(k\times1)}{\varepsilon_{t+1}}|\mathcal{F}_t &\sim N\left(\underset{(k\times1)}{0}, \underset{(k\times k)}{H_{t+1}}\right)\end{aligned}$$

We are considering a vector of returns, $\mathbf{r}_{t+1}$, which has k elements. The conditional mean of $\mathbf{r}_{t+1}$ given $\mathcal{F}_t$ is μ_{t+1} and the conditional variance is H_{t+1}. Multivariate volatility modelling is concerned with capturing the movements in H_{t+1}.

> **Activity 10.1** Similar to the use of the squared residual as a proxy for the conditional variance, show that $\tilde{\sigma}_{ij,t} \equiv \varepsilon_{i,t}\varepsilon_{j,t}$ is a conditionally unbiased proxy for $\sigma_{ij,t}$, the conditional covariance between $r_{i,t}$ and $r_{j,t}$.

10.2.1 Recap: Means and variances of random vectors

Recall that the expectation of a vector of random variables is the vector of expectations:

$$\underbrace{E_t[\underset{(k\times1)}{\mathbf{r}_{t+1}}]}_{(k\times1)} = E_t\begin{bmatrix} r_{1,t+1}\\ r_{2,t+1}\\ \vdots\\ r_{k,t+1}\end{bmatrix} = \begin{bmatrix} E_t\left[r_{1,t+1}\right]\\ E_t\left[r_{2,t+1}\right]\\ \vdots\\ E_t\left[r_{k,t+1}\right]\end{bmatrix} \equiv \begin{bmatrix} \mu_{1,t+1}\\ \mu_{2,t+1}\\ \vdots\\ \mu_{k,t+1}\end{bmatrix} = \mu_{t+1}$$

The variance of a vector of random variables is *not* the vector of variances of the individual variables, rather it is a matrix of variances and covariances:

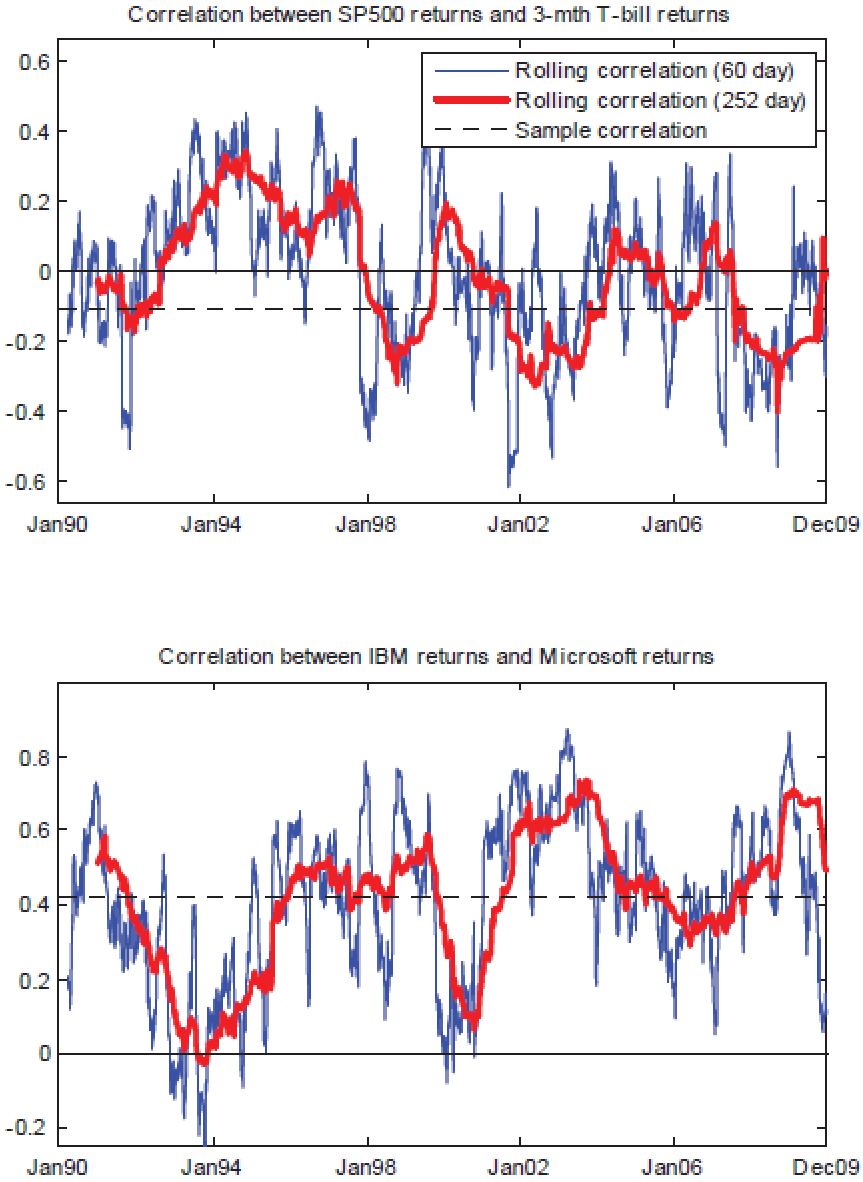
$$
\begin{aligned}
\underbrace{V_t[\underset{(k\times 1)}{\mathbf{r}_{t+1}}]}_{(k\times k)} &= E_t[\underset{(k\times 1)}{(\mathbf{r}_{t+1}-\mu_{t+1})}\underset{(1\times k)}{(\mathbf{r}_{t+1}-\mu_{t+1})'}] = E_t[\mathbf{r}_{t+1}\mathbf{r}_{t+1}'] - \mu_{t+1}\mu_{t+1}' \\
&= E_t\left[\varepsilon_{t+1}\varepsilon_{t+1}'\right] \\
&= E_t\begin{bmatrix} \varepsilon_{1,t+1}^2 & \varepsilon_{1,t+1}\varepsilon_{2,t+1} & \cdots & \varepsilon_{1,t+1}\varepsilon_{k,t+1} \\ \bullet & \varepsilon_{2,t+1}^2 & \cdots & \varepsilon_{2,t+1}\varepsilon_{k,t+1} \\ \vdots & \vdots & \ddots & \vdots \\ \bullet & \bullet & \cdots & \varepsilon_{k,t+1}^2 \end{bmatrix} \\
&= \begin{bmatrix} E_t\left[\varepsilon_{1,t+1}^2\right] & E_t\left[\varepsilon_{1,t+1}\varepsilon_{2,t+1}\right] & \cdots & E_t\left[\varepsilon_{1,t+1}\varepsilon_{k,t+1}\right] \\ \bullet & E_t\left[\varepsilon_{2,t+1}^2\right] & \cdots & E_t\left[\varepsilon_{2,t+1}\varepsilon_{k,t+1}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \bullet & \bullet & \cdots & E_t\left[\varepsilon_{k,t+1}^2\right] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_{1,t+1}^2 & \sigma_{12,t+1} & \cdots & \sigma_{1k,t+1} \\ \bullet & \sigma_{2,t+1}^2 & \cdots & \sigma_{2k,t+1} \\ \vdots & \vdots & \ddots & \vdots \\ \bullet & \bullet & \cdots & \sigma_{k,t+1}^2 \end{bmatrix} \\
&\equiv H_{t+1}
\end{aligned}
$$

The notation ' $\bullet$ ' for symmetric matrices means that a given element is equal to its corresponding element in the upper triangle.

> **Activity 10.2** Consider a portfolio of assets, with weight vector $\mathbf{w}$. If the vector of asset returns, $\mathbf{r}_t$, has conditional mean vector μ_t and conditional covariance matrix H_t, find the conditional mean and variance of the portfolio return, defined as $r_{p,t} = \mathbf{w}'\mathbf{r}_t$.

10.2.2 Empirical example

Figure 10.1 presents rolling window correlations between daily returns on the S&P 500 index and a 3-month Treasury bill, as well as correlations between daily returns on IBM and Microsoft. In the top panel we see that correlation between equity and T-bill returns can swing from positive (as high as 0.4) to negative (as low as -0.6). These large movements and changes in sign mean that the diversification benefits that are possible from holding these two assets change substantially through time. The correlation between IBM and Microsoft also varies substantially through time, from as low as zero to as high as 0.8.


Correlation between SP500 returns and 3-mth T-bill returns

Correlation between IBM returns and Microsoft returns

Figure 10.1: This figure plots correlations between daily returns on the S&P 500 index and a 3-month Treasury bill (upper panel) and between daily returns on IBM and Microsoft (lower panel) over the period January 1990 to December 2009.

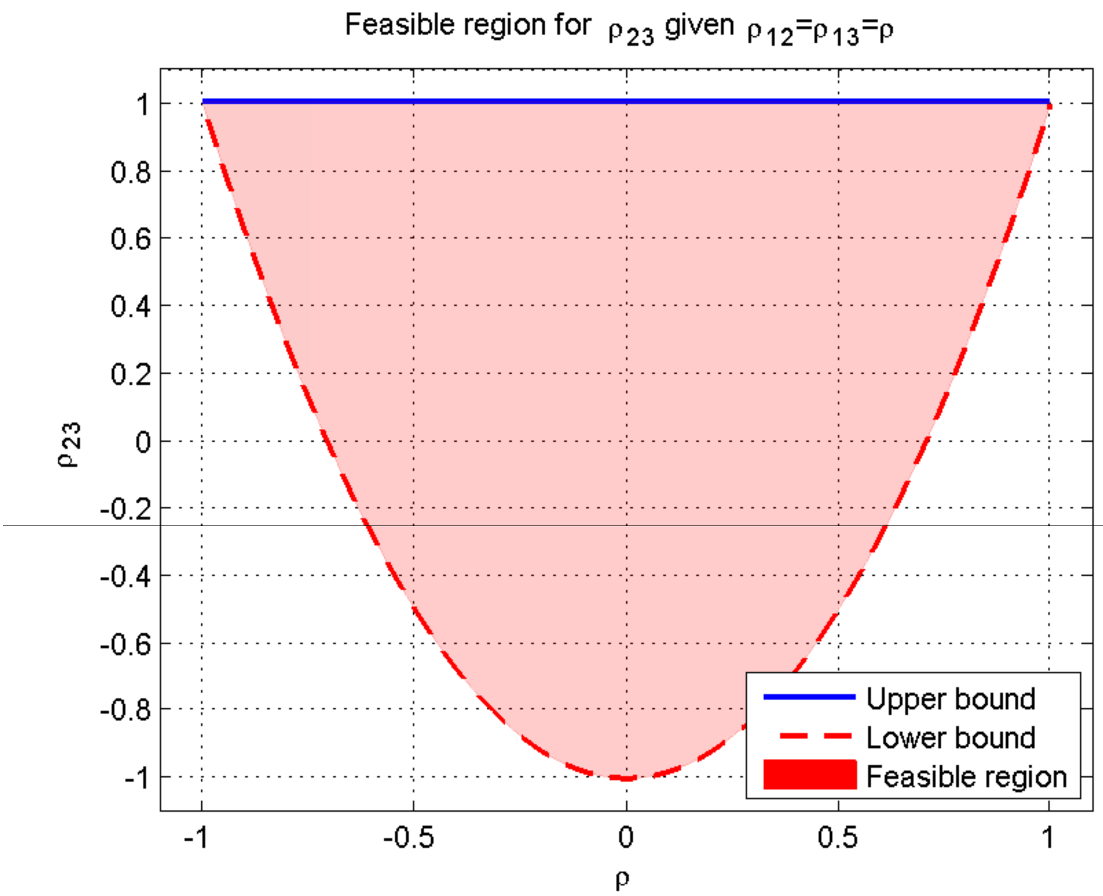


Figure 10.2: Bounds on ρ_{23} *given that* $\rho_{12} = \rho_{13} = \rho$, *for various values of* ρ.

10.2.3 The two main problems in multivariate volatility modelling

Parsimony

Models for time-varying covariance matrices tend to grow very quickly with the number of variables being considered. Recall that the covariance matrix of a collection of k variables grows with $k^2 = \mathcal{O}\left(k^2\right)$, while the number of data points only grows linearly with the number of variables (number of observations $= kT = O\left(k\right)$). It is important that the number of free parameters to be estimated be controlled somehow, and this often comes at the cost of flexibility.

Positive definiteness

Left unconstrained some multivariate volatility models do not always yield positive (semi-) definite conditional covariance matrix estimates. This is the multivariate 'equivalent' of a univariate volatility model predicting a negative variance. This is clearly undesirable. Imposing positive definiteness on some models leads to non-linear constraints on the parameters of the models which can be difficult to impose practically. Figure 10.2 shows that range of permissable values for ρ_{23} given that $\rho_{12} = \rho_{13} = \rho \in [-1, 1]$. We see that if $\rho_{12} = \rho_{13}$ and these two correlations are near -1 or +1, then the range of possible values for ρ_{23} is quite narrow.

10.3 Two popular multivariate volatility models

10.3.1 The constant conditional correlation (CCC) model

Recall that a (conditional) covariance can be decomposed into three parts

$$h_{ijt} = \sqrt{h_{it}h_{jt}} \cdot \rho_{ijt}$$

That is, the conditional covariance between r_{it} and r_{jt} is equal to the product of the two conditional standard deviations and the conditional correlation between r_{it} and r_{jt}. Bollerslev (1990) proposed assuming that the time variation we observe in conditional covariances is driven entirely by time variation in conditional variances; i.e., the conditional correlations are constant. Making this assumption greatly simplifies the specification and estimation of multivariate GARCH models.

With this assumption the conditional covariances are obtained as

$$\text{Assume} \quad h_{ijt} = \sqrt{h_{it}h_{jt}} \cdot \bar{\rho}_{ij}$$

where $\bar{\rho}_{ij}$ is the **unconditional** correlation between the standardised residuals $\varepsilon_{it}/\sqrt{h_{it}}$ and $\varepsilon_{jt}/\sqrt{h_{jt}}$, and h_{it} and h_{jt} are obtained from some univariate GARCH model.

The matrix versions of the above two equations are:

$$\begin{aligned}
H_t &= D_t R_t D_t \\
\text{where } D_t &= \begin{bmatrix} \sqrt{h_{11,t}} & 0 & ... & 0 \\ 0 & \sqrt{h_{22,t}} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sqrt{h_{kk,t}} \end{bmatrix} \equiv diag\left(\left[\sqrt{h_{11,t}}, \sqrt{h_{22,t}}, ..., \sqrt{h_{kk,t}}\right]\right) \\
R_t &= \begin{bmatrix} 1 & \rho_{12,t} & \cdots & \rho_{1k,t} \\ \rho_{12,t} & 1 & \cdots & \rho_{2k,t} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1k,t} & \rho_{2k,t} & \cdots & 1 \end{bmatrix}
\end{aligned}$$

So D_t is a matrix with the conditional standard deviations on the diagonal and zeros elsewhere, and R_t is the conditional correlation matrix.

Bollerslev's model is:

$$H_t = D_t \bar{R} D_t$$

where the elements of D_t are specified to be univariate GARCH processes, and $\bar{R}$ is simply the unconditional correlation of the standardised residuals. If we assume standard GARCH(1,1) processes for the individual volatility models then the number of parameters (P) here is:

$$\begin{aligned}
P &= 3k + k(k-1)/2 \\
&= O\left(k^2\right)
\end{aligned}$$

An important distinction needs to made here: while there are $O\left(k^2\right)$ parameters in this model, it is specified in such a way that it can be consistently (though not efficiently) estimated in *stages*: the first k stages are the univariate GARCH models, and the final stage is the correlation matrix. Since the unconditional correlation matrix is simple to compute (as are univariate GARCH models with today's technology) this is an easy model to estimate.

Bollerslev's model is both parsimonious and ensures positive definiteness (as long as the variances from the univariate GARCH models are positive). However, there exists some empirical evidence against the assumption that conditional correlations are constant. Nevertheless, Bollerslev's model is a reasonable benchmark against which to compare any alternative multivariate GARCH model.

10.3.2 The RiskMetrics exponential smoother

A very simple method of constructing a time-varying conditional covariance matrix is to use the multivariate version of the RiskMetrics model for conditional volatility. This model takes the form:

$$H_{t+1} = \lambda H_t + (1 - \lambda) \varepsilon_t \varepsilon_t'$$

where, as in the univariate case, λ is set equal to 0.94 for daily data and 0.97 for monthly data. This model has the benefits of *no* estimated parameters, and so extreme parsimony, and that H_{t+1} is guaranteed to be positive definite if a positive definite matrix (such as the unconditional covariance matrix) is used as H_0. The assumption that the 'smoothing' parameter λ is the same for all elements of H_{t+1} is restrictive and not likely to be true in practice. However allowing the smoothing parameter λ to be different across the different elements of H_{t+1} means that positive definiteness is no longer guaranteed. Despite the restrictive assumption, the RiskMetrics model often performs reasonably well in applied work.

10.4 Economic evaluation of multivariate volatility models

Multivariate volatility models can be evaluated in a variety of ways. One interesting method, which is also widely applicable, is to investigate how the various models perform in a portfolio decision problem. Consider, for example, the problem of allocating wealth between the risk-free asset, Coca Cola and Intel stocks. There are only two risky assets, and so our covariance matrix forecasts will be (2×2). Let $\mathbf{r}_t$ be the (2×1) vector of returns on Coca Cola and Intel minus the risk-free return; (i.e. the vector of *excess* returns at date t). Let us assume that the conditional mean of these returns is constant, and denoted μ. By examining excess returns, and not imposing any short-sales constraints on the investor, we will make the investment in the risk-free asset equal to the weight that would make the portfolio weights sum to one.

For simplicity, let's assume that the investor has quadratic utility, with $c > 0$, of the form:

$$\begin{aligned} \mathcal{U}(w_{1,t+1}, w_{2,t+1}) &= (w_{1,t+1} r_{1,t+1} + w_{2,t+1} r_{2,t+1}) \\ &\quad - \frac{1}{2} c (w_{1,t+1} r_{1,t+1} + w_{2,t+1} r_{2,t+1})^2 \end{aligned}$$

For this case the optimal portfolio weights can be shown to equal:

$$\begin{aligned} w^*_{1,t+1} &= \frac{1}{c} \cdot \frac{h_{22t+1}\mu_1 - h_{12t+1}\mu_2}{h_{11t+1}\mu_2^2 + h_{22t+1}\mu_1^2 - 2h_{12t+1}\mu_1\mu_2 + h_{11t+1}h_{22t+1} - h^2_{12t+1}} \\ w^*_{2,t+1} &= \frac{1}{c} \cdot \frac{h_{11t+1}\mu_2 - h_{12t+1}\mu_1}{h_{11t+1}\mu_2^2 + h_{22t+1}\mu_1^2 - 2h_{12t+1}\mu_1\mu_2 + h_{11t+1}h_{22t+1} - h^2_{12t+1}} \end{aligned}$$

A reasonable choice of c for our application, where we measure returns in percent, and so $r_{i,t} = 100 \times \log(P_{i,t}/P_{i,t-1})$, is around 0.03. (If we choose c too large, then the quadratic utility starts to *decrease* as returns get higher, clearly not a reasonable description of investor preferences.)

Let us assume for simplicity that the risk-free rate is constant at 4% per annum for the sample period. Some summary statistics on the excess returns for our portfolio decision are:

	Coca Cola	**Intel**
Mean	0.06	0.13
Std dev	1.53	2.42
Annualised mean	14.49	32.39
Annualised std dev	23.93	37.32
Skewness	0.03	-0.23
Kurtosis	5.63	5.37

So from the above table we take $\mu = (0.06, 0.13)$. We will use this portfolio decision problem to compare the RiskMetrics exponential smoother and the CCC model combined with GARCH(1,1) models for the univariate variances. From Figures 10.3 and 10.4 we see substantial variability in conditional variance, and also some variability in the conditional correlation. Comparing the RiskMetrics model with the CCC model will help us determine whether capturing time-varying correlations is helpful for asset allocation decisions between Coca Cola and Intel over our sample period.

I estimated these two multivariate volatility models on the excess returns on Coca Cola and Intel and obtained the optimal portfolio weights using the formula above. Some summary statistics on the portfolio weights and performance statistics are given below, and the optimal weights are plotted in Figures 10.5 and 10.6. For comparison, I also present the performance of a portfolio that uses weights $[0.5, 0.5, 0]$ in Coca Cola, Intel and the risk-free asset (the ‘equally-weighted portfolio’, denoted EQ).

The table below shows, as expected, that as we increase the degree of risk aversion the aggressiveness of the portfolio weights decreases: the average short position in the risk-free asset decreased from -1.56 to -0.28 in the RiskMetrics portfolio, for example.

The bottom row in the table shows that the RiskMetrics portfolio generated higher average utility than the CCC portfolio for all levels of risk aversion, while the equally-weighted portfolio generated the lowest average utility for all levels of risk aversion. We conclude that the RiskMetrics model out-performs the CCC model for portfolio decisions involving these two assets, and both RiskMetrics and the CCC model out-perform a simple equally-weighted portfolio.

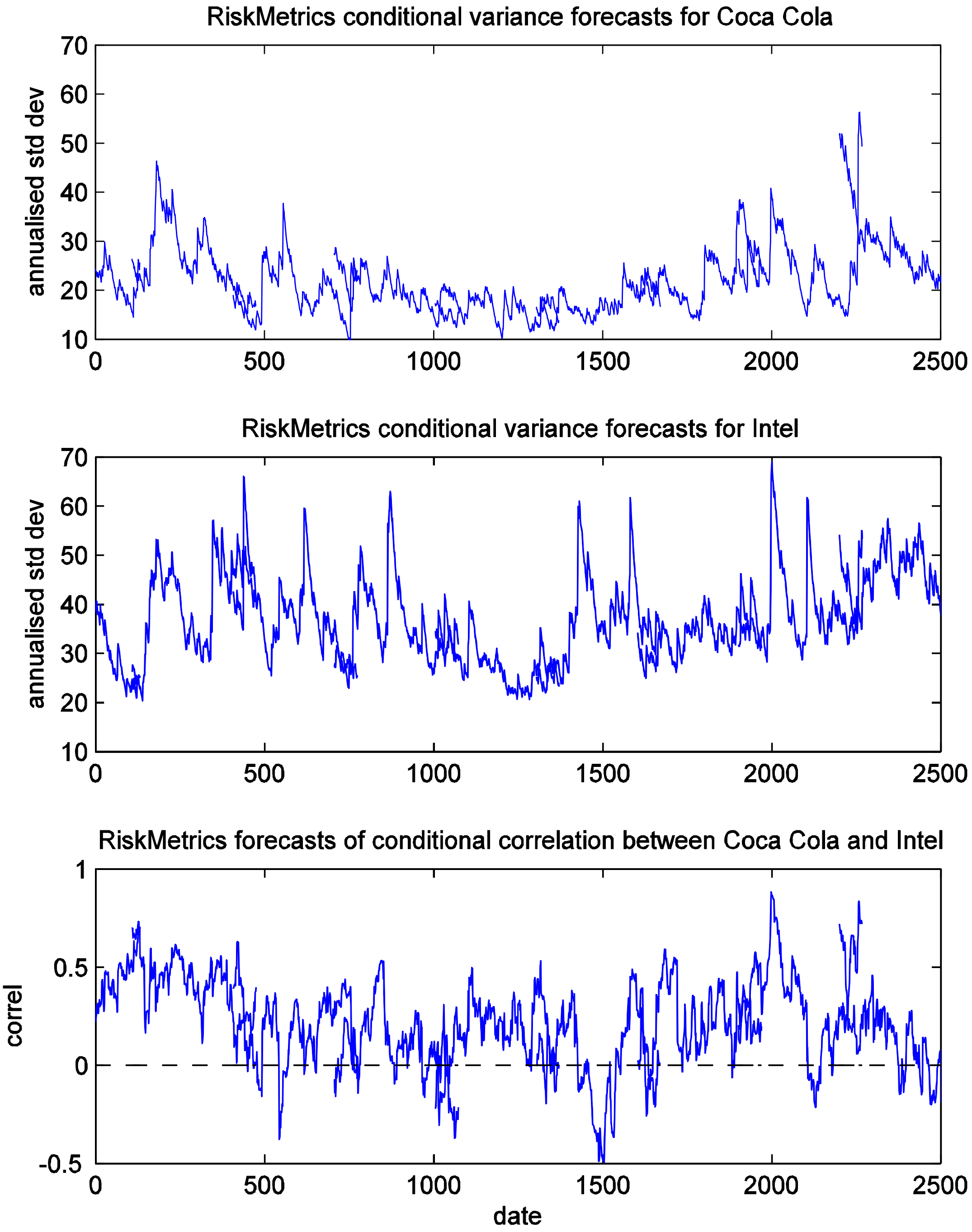


Figure 10.3: Conditional variance and correlation forecasts from the RiskMetrics model (with $\lambda = 0.94$) over the period January 1990 to December 1999.

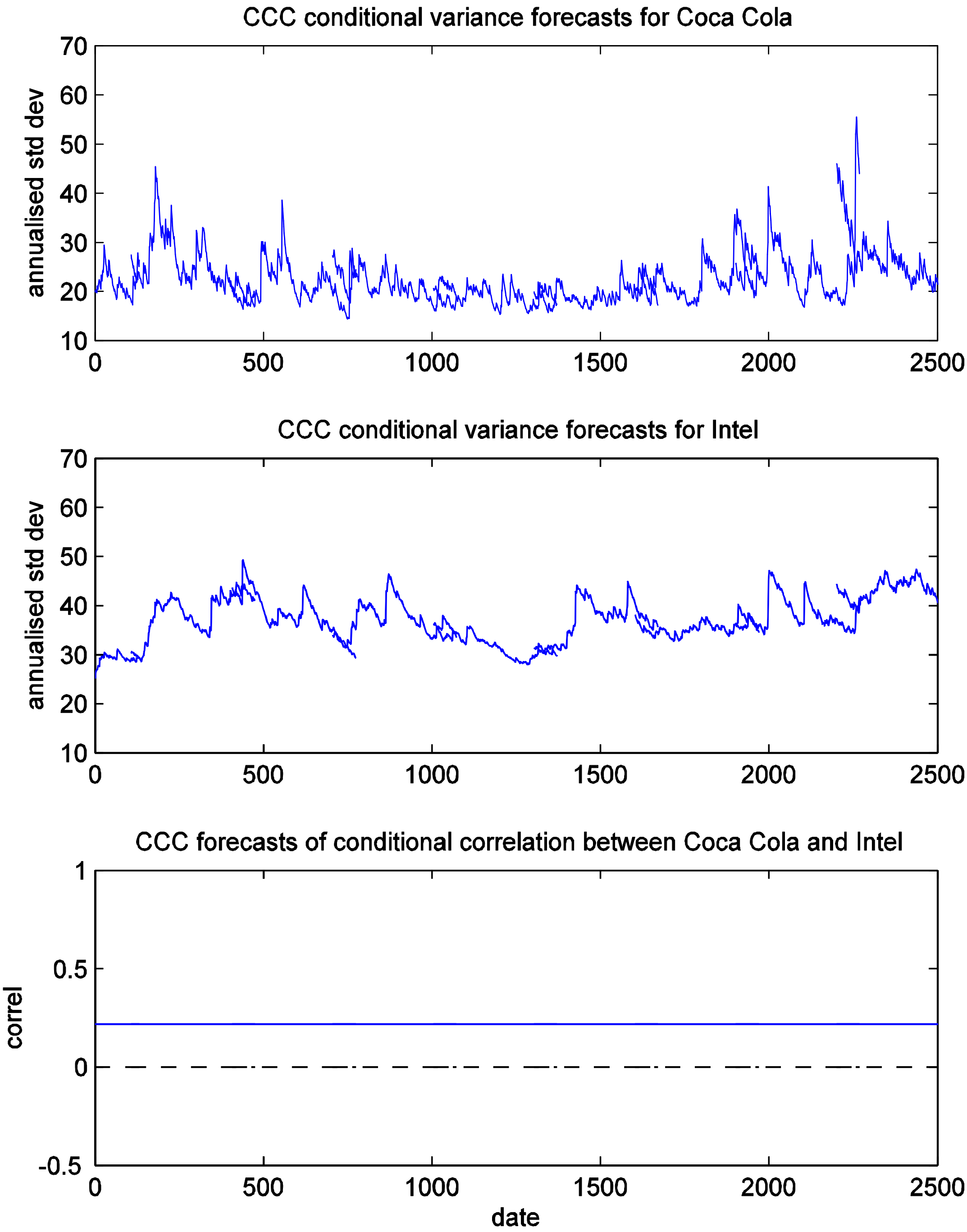


Figure 10.4: Conditional variance and correlation forecasts from the CCC model over the period January 1990 to December 1999.

	$c = 0.02$			$c = 0.03$			$c = 0.04$		
	RM	CCC	EQ	RM	CCC	EQ	RM	CCC	EQ
Avg CC weight	1.27	1.04	0.50	0.85	0.70	0.50	0.64	0.52	0.50
Avg Intel weight	1.28	1.02	0.50	0.86	0.68	0.50	0.64	0.51	0.50
Avg R$_f$ weight	-1.56	-1.06	0.00	-0.71	-0.38	0.00	-0.28	-0.03	0.00
Avg return	0.27	0.19	0.09	0.19	0.13	0.09	0.14	0.10	0.09
Std dev	3.90	3.08	1.57	2.60	2.06	1.57	1.95	1.54	1.57
Skewness	-0.01	-0.08	-0.09	-0.01	-0.08	-0.09	-0.01	-0.08	-0.09
Kurtosis	5.64	4.37	4.91	5.64	4.37	4.91	5.64	4.37	4.91
Avg utility*100	11.99	9.47	6.83	8.52	6.84	5.59	6.78	5.52	4.35

10.5 Overview of chapter

This chapter considered methods and models that relate to financial decisions that involve multiple sources of risk, such as portfolio decisions and risk management decisions. We discussed two simple and widely-used models for conditional covariance matrices, and reviewed an application of these models to a portfolio decision problem.

10.6 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Describe the two main problems in multivariate volatility modelling
- Compare and contrast two popular models for covariance matrices
- Derive the mean and variance for a portfolio return from the mean vector and covariance matrix of the underlying assets

10.7 Test your knowledge and understanding

1. Discuss the pros and cons of the CCC model of Bollerslev (1990) versus the RiskMetrics model for modelling multivariate volatility.
2. Consider an example involving three asset returns (X, Y, Z). Assume that $E[X] = E[Y] = E[Z] = 0$ and that $V[X] = V[Y] = V[Z] = 1$. A researcher uses a model that assumes $Corr[X, Y] = Corr[X, Z] = 0.8$ and $Corr[Y, Z] = 0$. Find the variance the return on a portfolio with weights $[-1, 0.5, 0.5]$ and interpret your result. (Hint: the portfolio variance is negative.)
3. Don't forget to check the VLE for additional practice problems for this chapter.

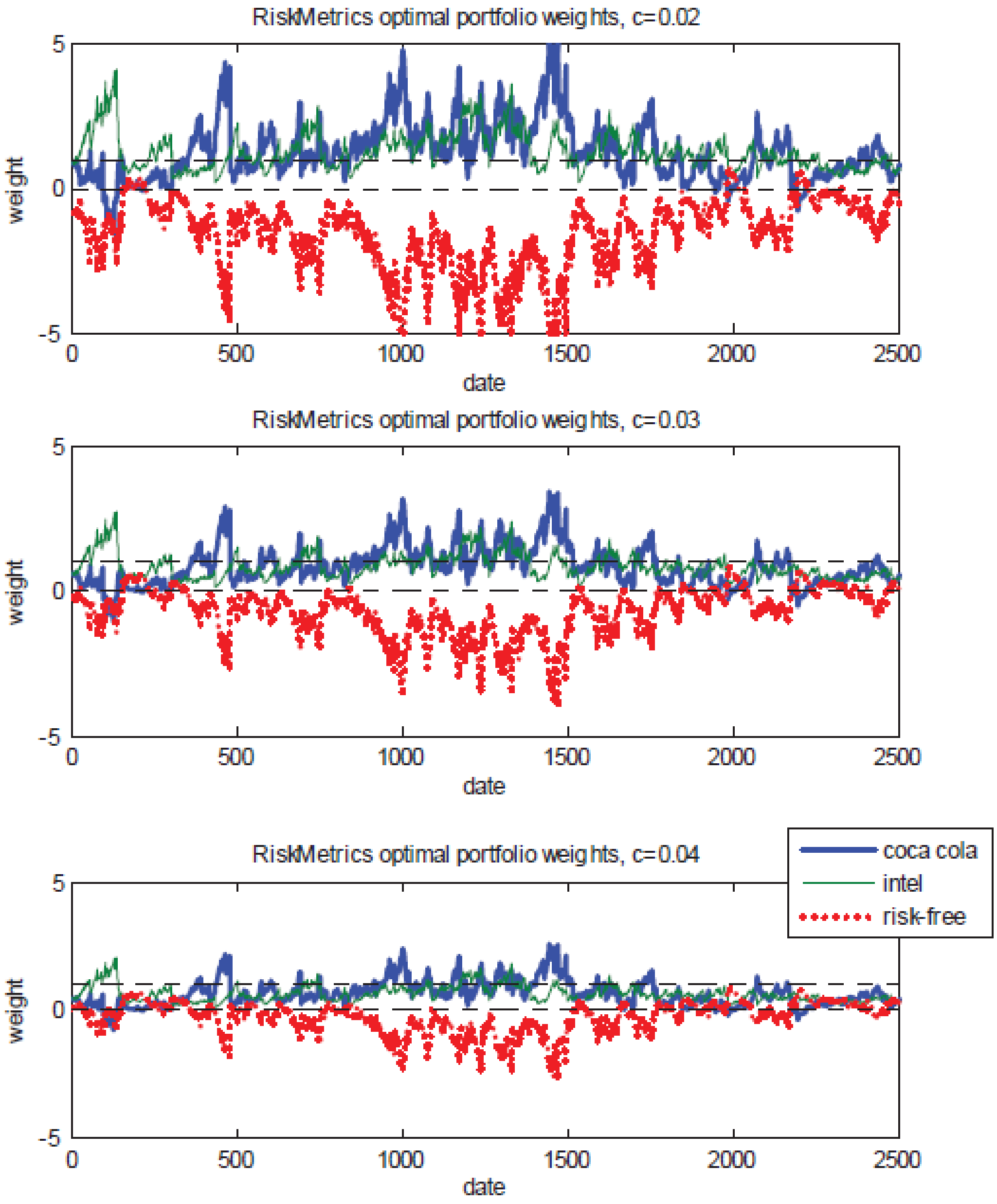


Figure 10.5: Optimal portfolio weights obtained using the RiskMetrics covariance matrix forecasts.

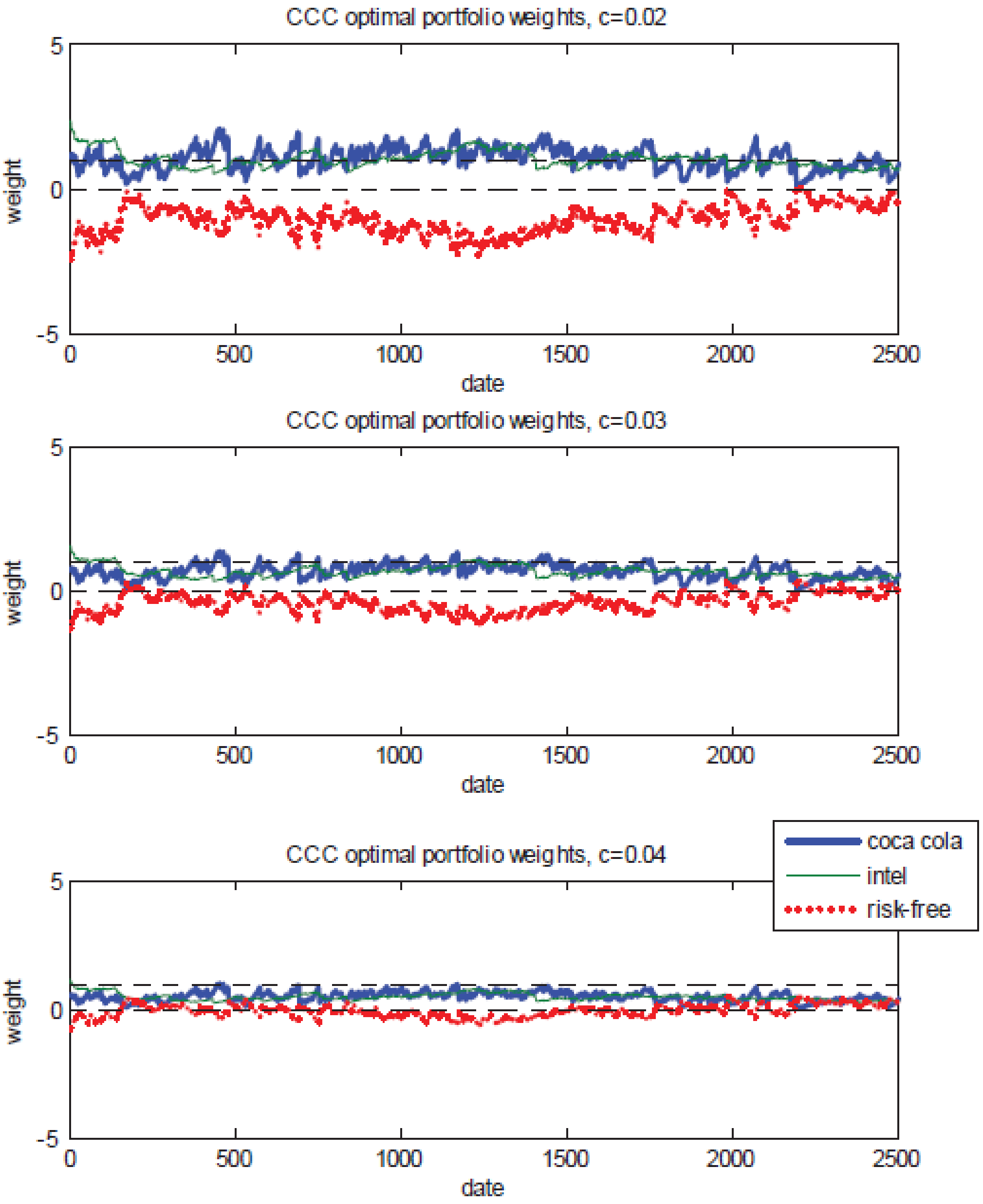


Figure 10.6: Optimal portfolio weights obtained using the CCC covariance matrix forecasts.

10.8 Solutions to activities

Activity 10.1

If we have

$$\begin{aligned}\mathbf{r}_t &= \mu_t + \varepsilon_t \\ \varepsilon_t|\mathcal{F}_{t-1} &\sim N(0, H_t)\end{aligned}$$

and we set $\tilde{\sigma}_{ij,t} = \varepsilon_{i,t}\varepsilon_{j,t}$ then

$$\begin{aligned}E_{t-1}[\tilde{\sigma}_{ij,t}] &= E_{t-1}[\varepsilon_{i,t}\varepsilon_{j,t}] \\ &= E_{t-1}[(r_{i,t} - \mu_{i,t})(r_{j,t} - \mu_{j,t})] \\ &\equiv Cov_{t-1}[r_{i,t}, r_{j,t}]\end{aligned}$$

and so $\tilde{\sigma}_{ij,t} = \varepsilon_{i,t}\varepsilon_{j,t}$ is a conditionally unbiased proxy for $\sigma_{ij,t} \equiv Cov_{t-1}[r_{i,t}, r_{j,t}]$.

Activity 10.2

We are given:

$$\begin{aligned}E_{t-1}[\mathbf{r}_t] &= \mu_t \\ V_{t-1}[\mathbf{r}_t] &= H_t \\ \text{and } r_{p,t} &= \mathbf{w}'\mathbf{r}_t\end{aligned}$$

then the mean is easily obtained:

$$E_{t-1}[r_{p,t}] = E_{t-1}[\mathbf{w}'\mathbf{r}_t] = \mathbf{w}'E_{t-1}[\mathbf{r}_t] = \mathbf{w}'\mu$$

The variance of the portfolio return can be obtained either by recalling the formula that $V[\mathbf{a}'\mathbf{X}] = \mathbf{a}'V[\mathbf{X}]\mathbf{a}$, for any random vector $\mathbf{X}$ and constant vector $\mathbf{a}$, and so:

$$V_{t-1}[r_{p,t}] = V_{t-1}[\mathbf{w}'\mathbf{r}_t] = \mathbf{w}'V_{t-1}[\mathbf{r}_t]\mathbf{w} = \mathbf{w}'H_t\mathbf{w}$$

or by deriving it explicitly using the expression for the variance of a random vector from above:

$$\begin{aligned}V_{t-1}[r_{p,t}] &= V_{t-1}[\mathbf{w}'\mathbf{r}_t] \\ &= E_{t-1}\left[(\mathbf{w}'\mathbf{r}_t - \mathbf{w}'\mu)(\mathbf{w}'\mathbf{r}_t - \mathbf{w}'\mu)'\right] \\ &= E_{t-1}[(\mathbf{w}'\mathbf{r}_t - \mathbf{w}'\mu)(\mathbf{r}_t'\mathbf{w} - \mu'\mathbf{w})]\text{, property of transpose operator} \\ &= E_{t-1}[\mathbf{w}'\mathbf{r}_t\mathbf{r}_t'\mathbf{w}] - E_{t-1}[\mathbf{w}'\mathbf{r}_t\mu'\mathbf{w}] \\ &\quad -E_{t-1}[\mathbf{w}'\mu\mathbf{r}_t'\mathbf{w}] + E_{t-1}[\mathbf{w}'\mu\mu'\mathbf{w}]\text{, expanding} \\ &= \mathbf{w}'E_{t-1}[\mathbf{r}_t\mathbf{r}_t']\mathbf{w} - \mathbf{w}'\mu\mu'\mathbf{w}\text{, simplifying} \\ &= \mathbf{w}'(E_{t-1}[\mathbf{r}_t\mathbf{r}_t'] - \mu\mu')\mathbf{w} \\ &= \mathbf{w}'H_t\mathbf{w}\end{aligned}$$

Chapter 11
Optimal forecasts and forecast evaluation

11.1 Introduction

This chapter and the next consider the general problems of forecast evaluation and comparison. In this chapter, we will formally define what is meant by a 'good' (or 'optimal') forecast and present methods for testing whether a given sequence of forecasts is optimal. In the next chapter we will consider methods for comparing competing forecasts, and for combining these forecasts into (hopefully) an even better forecast.

11.1.1 Aims of the chapter

The aims of this chapter are to:

- Present the formal definition of forecast optimality
- Present some standard methods for testing forecast optimality in data
- Discuss how these methods can be adapted for use in applications where the variable of interest is unobservable (eg, volatility forecasting).

11.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Derive the optimal forecast for a given loss function
- Interpret the results of Mincer-Zarnowitz tests of forecast optimality

11.1.3 Essential reading

- Diebold, F.X. *Elements of Forecasting.* (Thomson South-Western, Canada, 2006) fourth edition [ISBN 9780324323597]. Chapter 12, Section 12.1
- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487]. Chapter 4, Section 4.6.

11.1.4 References cited

- Mincer, J. and V. Zarnowitz, 'The evaluation of economic forecasts,' in J. Mincer (ed.) *Economic Forecasts and Expectations* (National Bureau of Economic Research, New York).

11.2 Optimal forecasts

The evaluation of a given forecast involves checking whether that forecast has the properties of the theoretically *optimal forecast.* We define this formally in the next subsection, and then derive some examples for specific loss functions.

11.2.1 Definition of an optimal forecast

Given a loss function L, an optimal forecast is obtained by minimising the conditional expectation of the future loss:

$$\hat{Y}^*_{t+h|t} \equiv \arg\min_{\hat{y}} E\left[L\left(Y_{t+h}, \hat{y}\right)|\mathcal{F}_t\right].$$

There are several important ingredients in the above definition:

1. What is the *target variable*, Y_{t+h}? Is it a return over some interval of time, the level of some index, the volatility of an asset return?

2. What is the *information set*, $\mathcal{F}_t$? Is it based only on past observations of the target variable available at time t, or can we use information from other sources? (This is related to the weak-form and semi strong-form efficient markets hypotheses.)

3. The difference in the time subscripts on the target variable and the information set defines the *forecast horizon*, h. Note that the horizon will vary with the observation frequency. If daily data is used, then a one-month-ahead forecast will have $h = 22$ (business days), while if monthly data is used a one-month-ahead forecast will have $h = 1$.

4. What is the *loss function*, L, that will be used to measure the quality of the forecast? A loss function maps the realisation of the target variable and the forecast to a loss or cost. It is usually normalised to be zero when $y = \hat{y}$. We will look at a few common loss functions in the next subsection.

A key quantity in forecast evaluation is the forecast error, which is defined as:

$$e_{t+h|t} \equiv Y_{t+h} - \hat{Y}_{t+h|t}$$

When the forecast error is exactly zero the loss incurred is normalised to be zero.

Note that the optimal forecast is *not* the 'perfect' forecast, which would be the one where $\hat{Y}_{t+h|t} = Y_t$, and so there would be zero forecast error. This is because the optimal forecast is based only on information available at time t, and Y_{t+h} is not yet known, so it is not feasible to set $\hat{Y}_{t+h|t} = Y_t$. When we conduct forecast optimality tests we will compare a given forecast, $\hat{Y}_{t+h|t}$, with the best that could have been done with the information set available (that is, with the 'optimal' forecast), not with the infeasible perfect forecast.

11.2.2 Deriving some optimal forecasts

Quadratic loss (MSE)

The most commonly-used loss function is the quadratic, or squared-error (sometimes called mean squared error or MSE) loss function:

$$L(y, \hat{y}) = (y - \hat{y})^2 .$$

Under this loss function the optimal forecast is the conditional mean:

$$\begin{aligned} \hat{Y}^*_{t+h|t} &\equiv \arg\min_{\hat{y}} E\left[(Y_{t+h} - \hat{y})^2 |\mathcal{F}_t\right] \\ \text{FOC} \quad 0 &= \frac{\partial}{\partial \hat{y}} E\left[(Y_{t+h} - \hat{y})^2 |\mathcal{F}_t\right]\Big|_{\hat{y}=\hat{Y}^*_{t+h|t}} \\ 0 &= -2E\left[Y_{t+h} - \hat{Y}^*_{t+h|t}|\mathcal{F}_t\right] \\ \text{so} \quad \hat{Y}^*_{t+h|t} &= E\left[Y_{t+h}|\mathcal{F}_t\right] \end{aligned}$$

This may help explain why the common usage of ‘forecast,’ ‘prediction,’ and ‘expecatation’ are synonyms.

Linear-exponential loss (Linex)

The quadratic loss function imposes the assumption that positive and negative forecast errors of the same size carry the same penalty. A popular *asymmetric* loss function is the linear-exponential (linex) loss function:

$$\begin{aligned} L(y, \hat{y}) &= \frac{2}{a^2}\left[\exp\{a(y - \hat{y})\} - a(y - \hat{y}) - 1\right], \quad \text{where } a \neq 0 \\ &\rightarrow (y - \hat{y})^2 \quad \text{as } a \rightarrow 0 \end{aligned}$$

This loss function is easy to manipulate when the target variable is conditionally Normally distributed, making use of the following result:

$$\text{If } X \sim N\left(\mu, \sigma^2\right), \text{ then } E\left[\exp\{X\}\right] = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}.$$

So if

$$\begin{aligned} Y_{t+h}|\mathcal{F}_t &\sim N\left(\mu_{t+h|t}, \sigma^2_{t+h|t}\right) \\ \text{then} \quad a(Y_{t+h} - \hat{y})|\mathcal{F}_t &\sim N\left(a\left(\mu_{t+h|t} - \hat{y}\right), a^2\sigma^2_{t+h|t}\right) \end{aligned}$$

Using this, we can the compute the expected linex loss as follows:

$$\begin{aligned} & E\left[\frac{2}{a^2}\left[\exp\{a(Y_{t+h} - \hat{y})\} - a(y - \hat{y}) - 1\right]\Big| \mathcal{F}_t\right] \\ =& \frac{2}{a^2}\left(E\left[\exp\{a(Y_{t+h} - \hat{y})\}|\mathcal{F}_t\right] - a\left(E\left[Y_{t+h}|\mathcal{F}_t\right] - \hat{y}\right) - 1\right) \\ =& \frac{2}{a^2}\exp\left\{a\left(\mu_{t+h|t} - \hat{y}\right) + \frac{1}{2}a^2\sigma^2_{t+h|t}\right\} - \frac{2}{a}\left(\mu_{t+h|t} - \hat{y}\right) - \frac{2}{a^2} \end{aligned}$$

The first-derivative of the expected loss with respect to $\hat{y}$ is:

$$\begin{aligned}\frac{\partial}{\partial \hat{y}} E\left[L\left(Y_{t+h},\hat{y}\right)|\mathcal{F}_t\right] &= \frac{2}{a^2}\left(-a\right)\exp\left\{a\left(\mu_{t+h|t}-\hat{y}\right)+\frac{1}{2}a^2\sigma^2_{t+h|t}\right\}+\frac{2}{a}\\ &= \frac{2}{a}\left(1-\exp\left\{a\left(\mu_{t+h|t}-\hat{y}\right)+\frac{1}{2}a^2\sigma^2_{t+h|t}\right\}\right)\end{aligned}$$

and from this we obtain the first-order condition and the optimal forecast:

$$\begin{aligned}0 &= \left.\frac{\partial}{\partial \hat{y}} E\left[L\left(Y_{t+h},\hat{y}\right)|\mathcal{F}_t\right]\right|_{\hat{y}=\hat{Y}^*_{t+h|t}}\\ &= \frac{2}{a}\left(1-\exp\left\{a\left(\mu_{t+h|t}-\hat{Y}^*_{t+h|t}\right)+\frac{1}{2}a^2\sigma^2_{t+h|t}\right\}\right)\\ \text{so } 1 &= \exp\left\{a\left(\mu_{t+h|t}-\hat{Y}^*_{t+h|t}\right)+\frac{1}{2}a^2\sigma^2_{t+h|t}\right\}\\ 0 &= a\left(\mu_{t+h|t}-\hat{Y}^*_{t+h|t}\right)+\frac{1}{2}a^2\sigma^2_{t+h|t}\\ \hat{Y}^*_{t+h|t} &= \mu_{t+h|t}+\frac{a}{2}\sigma^2_{t+h|t}\end{aligned}$$

Thus the optimal linex forecast is a function of both the conditional mean *and* the conditional variance. Notice that when $a \to 0$ the linex loss function converges to the quadratic loss function, and the optimal forecast converges to just the conditional mean.

> **Activity 11.1** Derive the optimal forecast when the loss function is $L\left(y,\hat{y}\right)=a+b\left(y-\hat{y}\right)^2$ for $b>0$.

11.3 Testing forecast optimality

As noted above, the the quadratic loss function is the most commonly-used in economic and financial forecasting. We will assume the quadratic loss function for the remainder of this chapter.

11.3.1 Testable implications of forecast optimality

We test forecast optimality by comparing properties of a given forecast with those that we know are true for the optimal forecast, $\hat{Y}_{t+h|t}$, or for the optimal forecast error, $e^*_{t+h|t} \equiv Y_{t+h}-\hat{Y}^*_{t+h|t}$. Under MSE loss the following properties are true of optimal forecasts:

1. Optimal forecasts are unbiased: $E\left[Y_{t+h}|\mathcal{F}_t\right]=\hat{Y}^*_{t+h|t}$, which implies $E\left[e^*_{t+h|t}|\mathcal{F}_t\right]=0.$

   This means that on average the forecast is correct.

2. Optimal forecast errors, $e^*_{t+h|t}$, have zero autocorrelation for lags greater than or equal to h. (Note that this implies that for $h=1$ optimal forecast errors are white noise.)

This property relates to the use of information by optimal forecasts. It implies that the forecast error should not be forecastable, otherwise the forecast itself could be improved.

3. The variance of the optimal forecast error, $V\left[e^*_{t+h|t}\right]$, is increasing in h.

   This property asserts that the distant future is harder to forecast than the near future. If it were easier to forecast next month than it is to forecast tomorrow, we should simply throw away what we've learned in the last month and forecast tomorrow using only information from a month ago. Thus this property also relates to the use of information in optimal forecasts.

Activity 11.2 Show that optimal forecast errors have *unconditional* mean zero.

11.3.2 Regression-based tests of forecast optimality

For $h = 1$ we can test forecast optimality by running a regression of the forecast errors on a constant (to check property 1) and on lagged forecast errors (to check property 2). Further, we can include as regressors any variable in the time t information set. For example:

$$\begin{aligned} e_{t+1|t} &= \alpha_0 + \alpha_1 e_{t|t-1} + u_{t+1}, \;\; or \\ e_{t+1|t} &= \alpha_0 + \alpha_1 e_{t|t-1} + \alpha_2 y_t + \alpha_3 \hat{Y}_{t+1|t} + u_{t+1} \end{aligned}$$

And then test that all estimated parameters are zero. That is, test the hypothesis:

$$\begin{aligned} H_0 &: \alpha_0 = \alpha_1 = 0 \\ \text{vs. } H_a &: \alpha_i \neq 0 \text{ for } i = 0 \text{ or } 1, \quad or \\ H_0 &: \alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ \text{vs. } H_a &: \alpha_i \neq 0 \text{ for some } i = 0, 1, 2, 3 \end{aligned}$$

If we can reject this hypothesis then we can reject the assumption that the forecast is optimal.

An alternative way of testing forecast optimality is to run a 'Mincer-Zarnowitz' (MZ) regression. A MZ regression is one where the dependent variable is regressed on the forecast:

$$Y_{t+1} = \beta_0 + \beta_1 \hat{Y}_{t+1|t} + u_{t+1}$$

And then test the hypothesis:

$$\begin{aligned} H_0 &: \beta_0 = 0 \;\cap \beta_1 = 1 \\ \text{vs. } H_a &: \beta_0 \neq 0 \;\cup \beta_1 \neq 1 \end{aligned}$$

We can generalise the above regression to include other elements of the time t information set. Such a regression is called an 'augmented MZ' regression. For example, we might estimate:

$$Y_{t+1} = \beta_0 + \beta_1 \hat{Y}_{t+1|t} + \beta_2 e_{t|t-1} + \beta_3 Y_t + u_{t+1}$$

And then test the hypothesis:

$$\begin{aligned} H_0 &: \beta_0 = \beta_2 = \beta_3 = 0 \cap \beta_1 = 1 \\ \text{vs. } H_a &: \beta_1 \neq 1 \cup \beta_i \neq 0 \text{ for } i = 0, 2, 3 \end{aligned}$$

Beware that the correlation between $\hat{Y}_{t+1|t}$ and Y_t may be quite high; if it is too high this regression may not work, and in that case Y_t should be dropped from the regression (and possibly replaced with some other variable). Testing hypotheses such as those above can be done via t-statistics and χ^2-statistics.

Under the null hypothesis we expect the residuals from an MZ regression to be serially uncorrelated, though they may still be heteroskedastic. If heteroskedasticity is a possibility (as it usually is in finance) then robust standard errors should be used to construct the χ^2 statistics.

11.3.3 Forecasting unobservable variables

In many forecasting problems the variable of interest, Y_{t+h}, is observable at time $t + h$. For example, if we forecast the return on a stock, or the temperature, or GDP growth, then at time $t + h$ we can compare our forecast with the observed value of the variable. But in other forecasting problems we can *never* observe the variable of interest, particularly in financial forecasting. For example, if the conditional variance or Value-at-Risk is the object of interest then even at time $t + h$ we still do not know what the actual value was. Depending on the particular case different approaches to overcome this problem are available.

For conditional variance forecast evaluation we can make use of the fact that the squared residual is a conditionally unbiased proxy for the (unobservable) conditional variance:

$$\begin{aligned} \text{If } \varepsilon_{t+1} &= \sigma_{t+1}\nu_{t+1}, \ \ \nu_{t+1}|\mathcal{F}_t \sim (0,1), \\ \text{then } E_t\left[\varepsilon_{t+1}^2\right] &= E_t\left[\sigma_{t+1}^2\nu_{t+1}^2\right] \\ &= \sigma_{t+1}^2 E_t\left[\nu_{t+1}^2\right] \\ &= \sigma_{t+1}^2 \end{aligned}$$

This enables us to use most forecast evaluation methods as though the squared residual was our object of interest, even though the conditional variance is our true object of interest. Importantly, we can use MZ regressions (or augmented MZ regressions) to evaluate volatility forecasts, using ε_{t+1}^2 in place of the unobservable σ_{t+1}^2.

The cost of using a conditionally unbiased proxy for the conditional variance rather than the true conditional variance is in the *power* of tests. Because the proxy is a 'noisy' (i.e., imprecise) estimate of the true conditional variance, the standard errors in the regressions will be larger, making it more difficult to reject the null hypothesis when it is false.

An alternative method for evaluating variance forecasts is a slight variation on the standard MZ regression:

$$\begin{aligned} \text{Standard MZ} \quad \varepsilon_{t+1}^2 &= \beta_0 + \beta_1 h_{t+1} + u_{t+1} \\ \text{Alternative MZ} \quad \frac{\varepsilon_{t+1}^2}{h_{t+1}} &= \beta_0 \frac{1}{h_{t+1}} + \beta_1 + u_{t+1} \end{aligned}$$

For both of these regressions we expect $\beta_0 = 0 \cap \beta_1 = 1$ if the variance forecast is optimal. Using the alternative MZ regression has the advantage that we expect the residual from the regression to be both serially uncorrelated *and* homoskedastic, so robust standard errors are not needed. Another common alternative MZ regression for variance forecasts is:

$$\begin{aligned} \frac{\varepsilon_{t+1}^2}{h_{t+1}} &= \beta_0 + \beta_1 \frac{\varepsilon_t^2}{h_t} + u_{t+1} \\ H_0 &: \beta_0 = 1 \cap \beta_1 = 0 \\ \text{vs. } H_a &: \beta_0 \neq 1 \cup \beta_1 \neq 0 \end{aligned}$$

The null hypothesis changes slightly here, but again this regression is simple to implement and does not require robust standard errors.

11.4 Numerical example

We will consider the problem for forecasting returns and volatility for the S&P 500 index, using daily data from January 1980 to December 2005 as our in-sample period ($R = 6563$ observations) and January 2006 to December 2009 as our (pseudo) out-of-sample period ($P = 1007$ observations). The data are plotted in Figure 11.1. We will consider two models for the conditional mean and two models for the conditional variance. The models for the conditional mean are a simple constant, and an AR(5):

$$Y_{t+1} = \mu + u_{t+1}, \qquad \text{Model A}$$

$$\begin{aligned} Y_{t+1} = \phi_0 + \phi_1 Y_t + \phi_2 Y_{t-1} + \phi_3 Y_{t-2} \\ +\phi_4 Y_{t-3} + \phi_5 Y_{t-4} + \varepsilon_{t+1,} \end{aligned} \qquad \text{Model B}$$

Using the residuals from the AR(5), the two models for the conditional variance are a GARCH(1,1) and a GJR-GARCH(1,1):

$$\varepsilon_{t+1} = \sigma_{t+1}\nu_{t+1}, \quad \nu_{t+1} \sim iid\ N(0,1)$$

$$\sigma_{t+1}^2 = \omega + \alpha\varepsilon_t^2 + \beta\sigma_t^2, \qquad \text{Model C}$$

$$\sigma_{t+1}^2 = \omega + \alpha\varepsilon_t^2 + \beta\sigma_t^2 + \gamma\varepsilon_t^2 \mathbf{1}\{\varepsilon_t < 0\}, \quad \text{Model D}$$

We will denote the one-step ahead forecasts from models A and B as $\hat{Y}_{t+1}^a$ and $\hat{Y}_{t+1}^b$, and will denote the volatility forecasts from models C and D as $\hat{h}_{t+1}^c$ and $\hat{h}_{t+1}^d$.

We will focus on one-step ahead forecasts. In Figure 11.2 is a plot of the conditional mean and conditional variance forecasts.

Mincer-Zarnowitz regressions for the conditional mean models and conditional variance models yield:

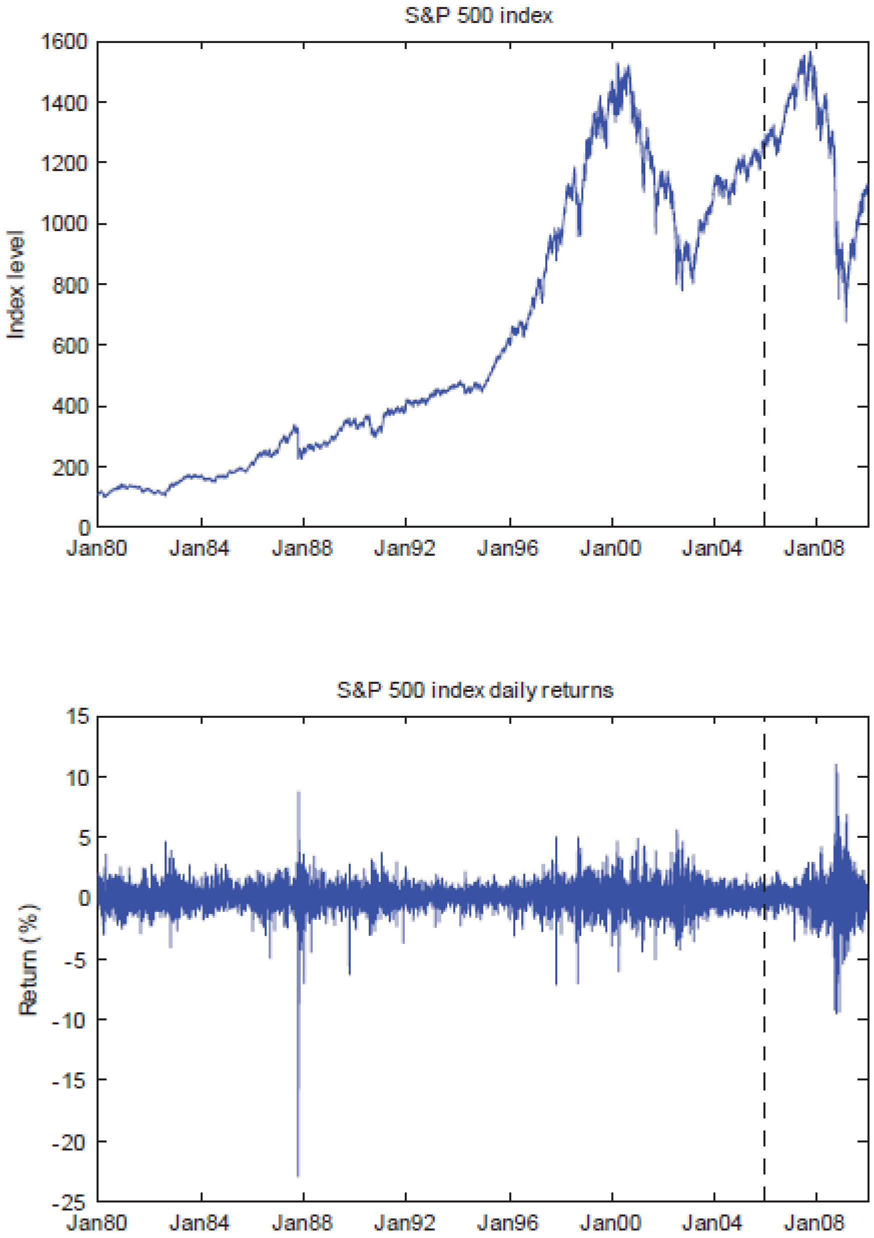


Figure 11.1: This figure plots the level (upper panel) and log-returns (lower panel) on the S&P 500 index over the period January 1980 to December 2009. A vertical dashed line at January 2006 indicates the split between the in-sample and out-of-sample periods.

Mincer-Zarnowitz regressions

	Conditional mean		*Conditional variance*	
Model	**A**	**B**	**C**	**D**
β_0	0.3788	-0.0406	0.1689	0.0078
(*std err*)	(0.5473)	(0.0517)	(0.2003)	(0.2172)
β_1	-11.2404	0.7932	0.9911	1.0185
(*std err*)	(14.9180)	(0.9932)	(0.1390)	(0.1363)
R^2	0.0007	0.0019	0.2377	0.2680
χ^2 statistic	5.1475	1.1379	1.8172	0.1217
χ^2 p-value	0.0762	0.5661	0.4031	0.9410

If we test the hypothesis that the constant equals zero and the slope equals one in these two regressions we obtain χ^2_2-statistics (p-values) of 5.1475 (0.0762) and 1.1379 (0.5661), which means that we cannot reject optimality, at the 0.05 level, for either forecast. Thus neither forecast is rejected as sub-optimal, even though their R^2's are near zero. This illustrates an important feature of forecasting: a forecast can be apparently optimal while still being quite uninformative.

Notice that the R^2 from the conditional variance regressions are much higher than in the mean forecasts, consistent with the general finding that volatility is more predictable than returns, at least at the daily frequency. Tests that the coefficients satisfy the implications of optimality lead to χ^2-statistics (p-values) of 1.8172 (0.4031) and 0.1217 (0.9410). Thus neither of these forecasts can be rejected as being sub-optimal.

Activity 11.3 Interpret the following results of two MZ regressions:

$$
\begin{aligned}
\varepsilon^2_{t+1} &= 0.049 + 0.981\hat{h}^a_{t+1} + u_{t+1},\ R^2 = 0.043 \\
\varepsilon^2_{t+1} &= 2.377 + 0.201\hat{h}^b_{t+1} + u_{t+1},\ R^2 = 0.078
\end{aligned}
$$

MZ χ^2_2-statistic (p-value) for regression 1: 0.018 (0.991). MZ χ^2_2-statistic (p-value) for regression 2: 10.232 (0.006).

11.5 Overview of chapter

This chapter presented the formal definition of 'forecast optimality' and derived the optimal forecast under quadratic and linex loss. We then examined testable implications of forecast optimality under quadratic loss, and discussed regression-based tests of these implications.

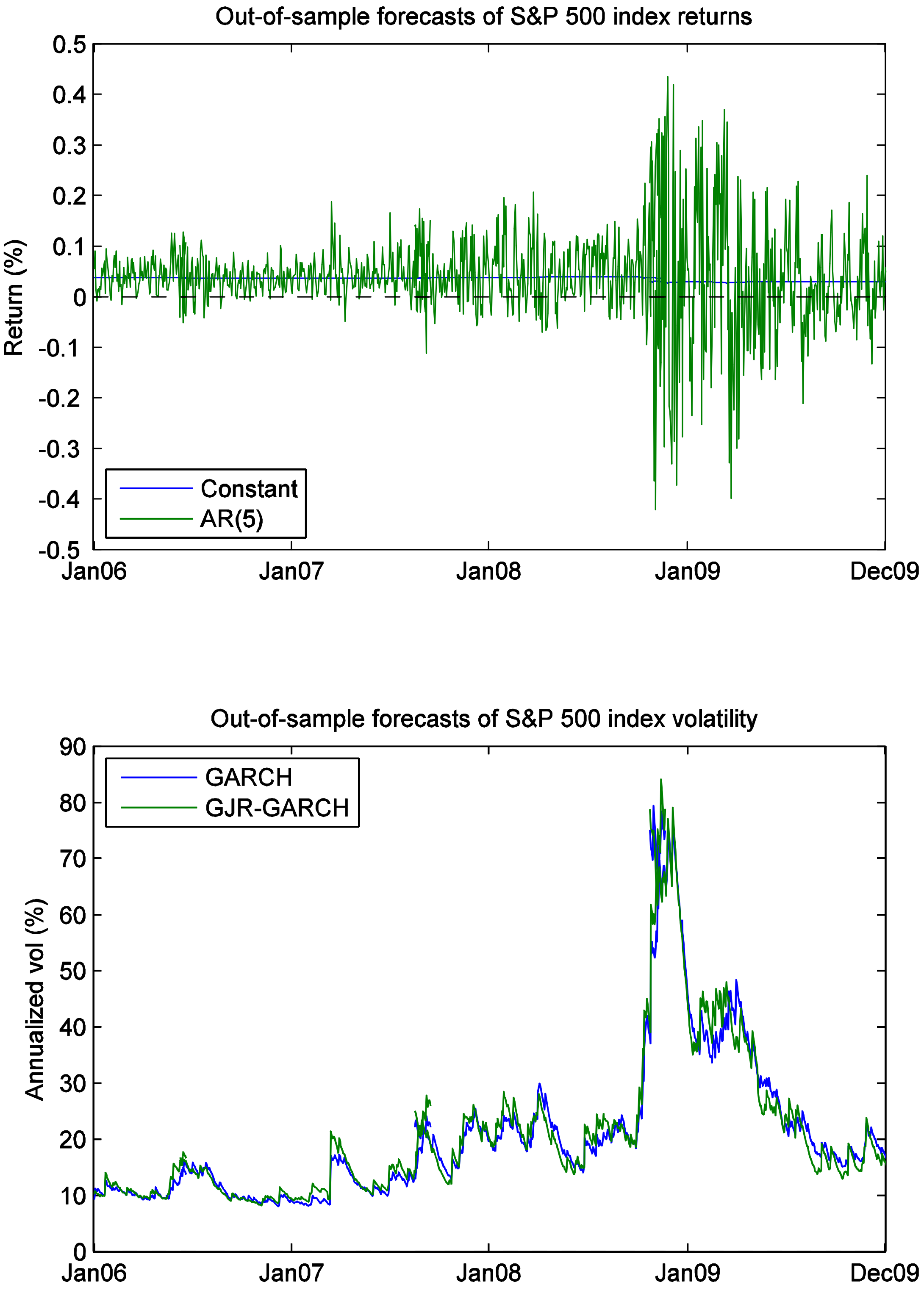


Figure 11.2: This figure plots out-of-sample forecasts of the conditional mean (upper panel) and conditional volatility (lower panel) of daily returns on the S&P 500 index over the period January 2006 to December 2009. An expanding window is used for estimation. The lower panel presents annualized volatility, computed as $\sqrt{252\hat{h}_t}$.

11.6 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Derive the optimal forecast for a given loss function
- Interpret the results of Mincer-Zarnowitz tests of forecast optimality, for both observable and unobservable variables

11.7 Test your knowledge and understanding

1. Can an economic forecast be 'optimal' and still forecast poorly (in terms of a low R^2 from a regression of the forecasted variable on a constant and the forecast)?
2. Can an economic forecast *fail to be* 'optimal' and still forecast well (in terms of a high R^2 from a regression of the forecasted variable on a constant and the forecast)?
3. Interpret the following results of two MZ regressions:

$$\begin{aligned} Y_{t+1} &= 0.094 + 0.874\hat{Y}_{t+1} + e_{t+1},\ R^2 = 0.006 \\ \varepsilon^2_{t+1} &= -0.242 + 1.301\hat{h}_{t+1} + u_{t+1},\ R^2 = 0.048 \end{aligned}$$

   MZ χ^2_2-statistic (p-value) for regression 1: 3.449 (0.178). MZ χ^2_2-statistic (p-value) for regression 2: 6.808 (0.033).

4. Don't forget to check the VLE for additional practice problems for this chapter.

11.8 Solutions to activities

Activity 11.1

If the loss function is

$$L(y, \hat{y}) = a + b(y - \hat{y})^2,\ b > 0$$

then the derivative of the expected loss with respect to $\hat{y}$ is

$$\begin{aligned} \frac{\partial}{\partial \hat{y}} E\left[L(Y_{t+h}, \hat{y})|\mathcal{F}_t\right] &= \frac{\partial}{\partial \hat{y}} E\left[a + b(Y_{t+h} - \hat{y})^2 |\mathcal{F}_t\right] \\ &= -2b\left(E\left[Y_{t+h}|\mathcal{F}_t\right] - \hat{y}\right) \end{aligned}$$

and so the first-order condition yields

$$\begin{aligned} \text{FOC} \quad 0 &= \left.\frac{\partial}{\partial \hat{y}} E\left[L(Y_{t+h}, \hat{y})|\mathcal{F}_t\right]\right|_{\hat{y}=\hat{Y}^*_{t+h|t}} \\ \text{so } \hat{Y}^*_{t+h|t} &= E\left[Y_{t+h}|\mathcal{F}_t\right] \end{aligned}$$

which is the same answer as for quadratic loss. This shows that adding a constant (a) and/or multiplying by some positive constant (b) does not affect the optimal forecast.

Activity 11.2

From Property 1 of an optimal forecast, we know that

$$E\left[e^*_{t+h|t}|\mathcal{F}_t\right] = 0.$$

By the law of iterated expectations we then have

$$E\left[E\left[e^*_{t+h|t}|\mathcal{F}_t\right]\right] = E\left[e^*_{t+h|t}\right] = 0$$

Thus optimal forecast errors have both *conditional* and *unconditional* mean zero.

Activity 11.3

The first regression has parameters close to $(0, 1)$, which is what we expect if $\hat{h}^a_t$ is optimal. The p-value from the MZ test is 0.991, which is greater than 0.05, and so we fail to reject the joint null hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$. Thus we conclude that we have no evidence that $\hat{h}^a_t$ is sub-optimal. The second regression, in contrast, as parameters far from $(0, 1)$ and the MZ p-value is 0.006. This is less than 0.05, and so we reject the joint null hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$, at the 95% confidence level, and conclude that we *do* have evidence that $\hat{h}^b_t$ is sub-optimal. The R^2s from the two regressions are both somewhat low, and reveal that although $\hat{h}^b_t$ is sub-optimal, it yielded a higher R^2 than $\hat{h}^a_t$.

Chapter 12

Forecast comparison and combination

12.1 Introduction

There are often many competing statistical models available for use in the forecasting of a particular financial variable. There are also many commercially available forecasts, issued by brokers or mutual funds. How do we determine which model or forecaster is best? Can we combine the information in multiple forecasts to obtain a better forecast than any individual forecast? These questions relate to the problems of forecast comparison and forecast combination.

12.1.1 Aims of the chapter

The aims of this chapter are to:

- Discuss the problem of forecast comparison, and the use of 'robust' standard errors for forecast comparison tests
- Introduce the notion of forecast encompassing
- Describe methods for constructing a combination forecast from multiple individual forecasts

12.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Describe the Diebold-Mariano test for comparing forecasts.
- Interpret the results of parameter estimates and t-statistics with respect to tests for forecast evaluation and comparison.
- Describe how to construct an optimal 'combination' forecast.

12.1.3 Essential reading

- Diebold, F.X. *Elements of Forecasting.* (Thomson South-Western, Canada, 2006) fourth edition [ISBN 9780324323597]. Chapter 12.

12.1.4 References cited

- Diebold, F. X. and R. Mariano, ‘Comparing Predictive Accuracy,’ *Journal of Business & Economic Statistics*, 1995, 13, pp.253–265.
- Newey, W. K., and K. D. West, ‘A Simple, Positive Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,’ *Econometrica*, 1987, 55(3), pp.703–708.

12.2 Comparing forecasts

Most economists acknowledge that *any* proposed model for forecasting will not be perfect. In spite of this, it can be difficult to reject these models (as we saw in the previous chapter). In some of the natural sciences the basic laws governing the process of random variables are invariant, and may be observed with less and less error through further experimentation. But we can only observe one history for the economy, and the agents within it are constantly changing their behavior, and so changing the behavior of economic variables. Thus in economics and finance it is important to keep in mind that just about any model will be mis-specified/sub-optimal (but that does not necessarily mean not useful).

One way of comparing two forecasts is to conduct a ‘Diebold-Mariano’ test. A Diebold-Mariano test is a statistical comparison of the accuracy of two forecasts. Given forecasts $\hat{Y}_t^a$ and $\hat{Y}_t^b$, a Diebold-Mariano test works by comparing the average loss incurred from using the two forecasts.

$$\begin{aligned} l_{a,t} &= L\left(Y_t, Y_t^a\right) \\ l_{b,t} &= L\left(Y_t, Y_t^b\right) \\ d_t &= l_{a,t} - l_{b,t} \end{aligned}$$

where the loss function L can be any loss function. For example, we could compare the forecasts by comparing the difference in the squared errors from the two forecasts:

$$d_t = (e_t^a)^2 - (e_t^b)^2$$

Or we could think about the utility or profits made when following the two forecasts, thus using economic measures of goodness-of-fit. In this case we might write:

$$\begin{aligned} l_{a,t} &= -\mathcal{U}\left(Y_t, \hat{Y}_t^a\right) \\ l_{b,t} &= -\mathcal{U}\left(Y_t, \hat{Y}_t^b\right) \end{aligned}$$

where $\mathcal{U}$ is some utility function. We used a minus sign in front of the utility function to make $l_{a,t}$ and $l_{b,t}$ ‘losses’ rather than ‘gains.’ The null hypothesis is that the two forecasts are equally good, and so we test

$$\begin{aligned} H_0 &: E\left[d_t\right] = 0 \\ \text{vs. } H_1 &: E\left[d_t\right] > 0 \\ \text{and } H_2 &: E\left[d_t\right] < 0 \end{aligned}$$

If Forecast A is better than Forecast B it will have lower expected loss and thus the mean of d_t will be significantly less than zero.

12.2.1 Estimating the variance of $\hat{E}[d_t]$: 'HAC' standard errors

Notice that the Diebold-Mariano test is simply a test that the mean of a random variable is equal to zero. These types of tests are covered in introductory statistics courses, and should be quite familiar. We will use a t-test that $E[d_t] = 0$, using the fact that

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}[\bar{d}]}} \overset{\mathcal{D}}{\to} N(0,1) \text{ as } T \to \infty$$

where the sample mean is computed in the usual way:

$$\hat{E}[d_t] \equiv \bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$$

The only complication in the application to Diebold-Mariano tests is that we need to allow for the fact that d_t might exhibit serial correlation. This means that computing the variance of $\bar{d}$ is more difficult than usual: if the variables are serially correlated then the variance of the sample mean is not simply $1/T$ times the variance of the individual variable; it must also include the covariance terms.

In the standard *iid* case we use the following derivation:

$$\begin{aligned}
V[\bar{d}] &= V\left[\frac{1}{T}\sum\nolimits_{t=1}^{T} d_t\right] \\
&= \frac{1}{T^2}V\left[\sum\nolimits_{t=1}^{T} d_t\right] \\
&= \frac{1}{T^2}\sum_{t=1}^{T} V[d_t], \text{ since } Cov[d_t, d_s] = 0 \text{ for } t \neq s \\
&= \frac{1}{T}V[d_t], \text{ since } V[d_1] = V[d_2] = ... = V[d_T] \\
\text{and } \hat{V}[\bar{d}] &= \frac{1}{T}\left(\frac{1}{T}\sum_{t=1}^{T}\left(d_t - \bar{d}\right)^2\right) \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\left(d_t - \bar{d}\right)^2 \quad \text{as an estimator, so} \\
DM &= \frac{\bar{d}}{\sqrt{\hat{V}[\bar{d}]}} = \frac{\sqrt{T}\bar{d}}{\sqrt{\hat{V}[d_t]}}
\end{aligned}$$

which is the usual form for a t-statistic.

If we want to allow for serial correlation (i.e., we cannot assume that $Cov[d_t, d_s] = 0$ for

$t \neq s$) then we must deal with more terms:

$$\begin{aligned}
V\left[\bar{d}\right] &= V\left[\frac{1}{T}\sum\nolimits_{t=1}^{T} d_t\right] \\
&= \frac{1}{T^2} V\left[\sum\nolimits_{t=1}^{T} d_t\right] \\
&= \frac{1}{T^2}\sum_{t=1}^{T} V\left[d_t\right] + \frac{2}{T^2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} Cov\left[d_t, d_s\right] \\
&= \frac{1}{T} V\left[d_t\right] + \frac{2}{T^2}\sum_{j=1}^{T-1}\left(T-j\right) Cov\left[d_t, d_{t+j}\right] \\
&= \frac{1}{T} V\left[d_t\right] + \frac{2}{T}\sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right) Cov\left[d_t, d_{t+j}\right]
\end{aligned}$$

since $V\left[d_1\right] = V\left[d_2\right] = ... = V\left[d_T\right]$ and $Cov\left[d_t, d_{t+j}\right] = Cov\left[d_s, d_{s+j}\right]$. In the *iid* derivation we simply used the sample variance in the place of the true variance, and this also works here for the variance. But the estimation of the autocovariances is made difficult by the fact that high-order autocovariances are estimated with very few observations.

Since we cannot estimate the very high-order autocovariances accurately, we usually truncate this sum at some value, M, where M is some value such as $floor\left[T^{1/3}\right]$ or $floor\left[4\left(T/100\right)^{2/9}\right]$. One estimate of the sample variance in this case is the 'Newey-West' variance estimator:[1]

$$\begin{aligned}
\hat{V}\left[\bar{d}\right] &= \frac{1}{T}\hat{V}\left[d_t\right] + \frac{2}{T}\sum_{j=1}^{M}\left(1-\frac{j}{M+1}\right)\widehat{Cov}\left[d_t, d_{t+j}\right] \\
\text{where } \hat{V}\left[d_t\right] &= \frac{1}{T}\sum_{t=1}^{T}\left(d_t-\bar{d}\right)^2 \\
\text{and } \widehat{Cov}\left[d_t, d_{t+j}\right] &= \frac{1}{T-j}\sum_{t=1}^{T-j}\left(d_t-\bar{d}\right)\left(d_{t+j}-\bar{d}\right)
\end{aligned}$$

The test statistic for the above null hypothesis is:

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}\left[\bar{d}\right]}} \sim N\left(0,1\right)$$

and so if $|DM| > 1.96$ we reject the null hypothesis that both forecasts are equally good. If $DM > 1.96$ we conclude that model A's losses are significantly larger than model B's, and so model B is better than model A. If $DM < -1.96$ we conclude that model A's losses are significantly smaller than model B's and so model A is better than model B.

[1] To compute the Newey-West variance estimate EViews uses the rule $M = floor\left[4\left(T/100\right)^{2/9}\right]$, which equals 6 when $T = 1000$, for example.

12.2.2 Numerical example, continued

Let us now conduct a DM test to compare our two models for the return on the S&P 500 index from the previous chapter. The MSE for models A and B are 2.7720 and 2.7671, thus we can see that overall model B seems better under MSE loss, though not by much. The DM test will tell us whether the difference between model A and B is significant. The DM t-statistic is 0.2949, less than the critical values of ±1.96, and so the DM test suggests that these two forecasts are equally good (or equally bad).

Now we will do the same for the volatility forecasts (GARCH and GJR-GARCH), and d_t for this case is:

$$d_t = \left(\varepsilon_t^2 - \hat{h}_t^c\right)^2 - \left(\varepsilon_t^2 - \hat{h}_t^d\right)^2$$

The MSE for models C and D are 57.5053 and 55.2132, and so Model D appears better. The DM test statistic is 1.6801, which is again less than the 95% critical value, and so the DM test cannot distinguish between these two volatility models. (Note that if we were using 90% critical values of ±1.645 we would conclude that Model D is significantly better than Model C.)

Diebold-Mariano forecast comparison tests				
Model	*A*	*B*	*C*	*D*
MSE	2.7720	2.7671	57.5053	55.2132
Difference	0.0049		2.2921	
DM t-stat	0.2949		1.6801	

Activity 12.1 Suppose you have two forecasts, $\hat{Y}_t^a$ and $\hat{Y}_t^b$, from competing banks, and the following information:

$$\begin{aligned} \bar{d} &= \frac{1}{100}\sum_{t=1}^{100}\left(\left(e_t^a\right)^2 - \left(e_t^b\right)^2\right) = 0.15 \\ \hat{\sigma}_{\bar{d}}^2 &= \frac{1}{100}\sum_{t=1}^{100}\left(\left(e_t^a\right)^2 - \left(e_t^b\right)^2 - \bar{d}\right)^2 = 0.46 \end{aligned}$$

Assuming that the data are serially independent, describe how you would use this information to test whether forecast A is significantly better/worse than forecast B. Conduct the test (recall that the 95% critical values for a N(0,1) random variable are ±1.96) and interpret the result.

12.3 Forecast encompassing and combining

One forecast is said to 'encompass' another if it contains all the information that the other does, and possibly more. If this is true, then there is no need to try combining forecasts, as one of the forecasts contains all the available information. However, if neither forecast encompasses the other then some combination of the two may be better than either of the individual forecasts on their own. A test for forecast encompassing may be done as a regression of the target variable on the two forecasts:

$$Y_{t+1} = \beta_1 \hat{Y}^a_{t+1} + \beta_2 \hat{Y}^b_{t+1} + u_{t+1}$$

We then test the two hypotheses:

$$\begin{aligned} H_0^A &: \beta_1 = 1 \cap \beta_2 = 0 \\ \text{vs. } H_a^A &: \beta_1 \neq 1 \cup \beta_2 \neq 0, \text{ and} \\ H_0^B &: \beta_1 = 0 \cap \beta_2 = 1 \\ \text{vs. } H_a^B &: \beta_1 \neq 0 \cup \beta_2 \neq 1 \end{aligned}$$

If H_0^A cannot be rejected, then we say that forecast A encompasses forecast B. If H_0^B cannot be rejected then we say that forecast B encompasses forecast A. Quite often both of these hypotheses can be rejected, and in this case we conclude that neither forecast encompasses the other. When neither forecast encompasses the other we may create a better forecast by combining the two.

12.3.1 Numerical example, continued

Let us run a simple test for forecast encompassing for the two sets of models we considered above. The results from these regressions are in the table below.

Forecast encompassing regressions

	Conditional mean	*Conditional variance*
β_1	-1.3168	-1.7449
(std err)	(1.3857)	(0.8571)
β_2	0.8136	2.6916
(std err)	(0.9980)	(0.8772)
R^2	0.0020	0.2878
Test $\beta_1 = 1 \cap \beta_2 = 0$	2.8230	11.5780
p-value	0.2438	0.0031
Test $\beta_1 = 0 \cap \beta_2 = 1$	1.6871	5.0433
p-value	0.4302	0.0803

The test that $\beta_1 = 1 \cap \beta_2 = 0$ in the first regression has a χ^2-statistic (p-value) of 2.8230 (0.2438), which means that we cannot reject this null. The test that $\beta_1 = 0 \cap \beta_2 = 1$ has a χ^2-statistic (p-value) of 1.6871 (0.4302), which means that we cannot reject this null either. When both nulls cannot be rejected it is generally because both models are so poor that the parameters are estimated with low precision. The appropriate conclusion here is that we cannot determine whether one model encompasses the other or not.

For the second regression, using the volatility forecasts, the test that $\beta_1 = 1 \cap \beta_2 = 0$ has a χ^2-statistic (p-value) of 11.5780 (0.0031), which means that we can reject this null at the 0.05 level. This tells us that Model C does *not* encompass Model D. The test that $\beta_1 = 0 \cap \beta_2 = 1$ has a χ^2-statistic (p-value) of 5.0433 (0.0803), which means that we cannot reject this null at the 0.05 level (though we could at the 0.10 level). Thus in this case we would conclude that Model D encompasses Model C. This suggests that there are no gains to be had by combining these two volatility forecasts; we should instead stick with $\hat{h}_t^d$. Below we will still combine these forecasts, though in practice we would not.

> **Activity 12.2** What null hypotheses would you test to test forecast encompassing when there are 3 competing forecasts rather than 2?

12.3.2 Forecast combination

Suppose we want to create a combination forecast that is a convex combination of the two forecasts:

$$\hat{Y}_{t+1}^{combo} = \omega \hat{Y}_{t+1}^{a} + (1-\omega)\,\hat{Y}_{t+1}^{b}$$

How should we pick ω? Since we want to minimise the mean squared error, we want to choose ω to minimise

$$\frac{1}{n}\sum_{t=1}^{n}\left(Y_{t+1} - \hat{Y}_{t+1}^{combo}\right)^2 = \frac{1}{n}\sum_{t=1}^{n}\left(Y_{t+1} - \omega\hat{Y}_{t+1}^{a} - (1-\omega)\,\hat{Y}_{t+1}^{b}\right)^2$$

Notice that this looks like (and *is*) equivalent to running a regression:

$$Y_{t+1} = \omega\hat{Y}_{t+1}^{a} + (1-\omega)\,\hat{Y}_{t+1}^{b} + u_{t+1}$$

This is a restricted multiple regression, where the constant is forced to be zero and the two coefficients are forced to sum to 1. We can relax this assumption and run instead a more flexible regression:

$$Y_{t+1} = \beta_0 + \beta_1\hat{Y}_{t+1}^{a} + \beta_2\hat{Y}_{t+1}^{b} + u_{t+1}$$

Running this more flexible regression allows us to combine biased forecasts (whereas when we exclude the constant we should only combine unbiased forecasts). If we denote the OLS estimates of the coefficients in the above regressions as $\hat{\beta}$, then the optimal combination forecast is:

$$\hat{Y}_{t+1}^{combo*} = \hat{\beta}_0 + \hat{\beta}_1\hat{Y}_{t+1}^{a} + \hat{\beta}_2\hat{Y}_{t+1}^{b}$$

This forecast will be unbiased, and will incorporate the information contained in *both* forecasts.

12.3.3 Numerical example, concluded

We conclude with an attempt to combine the forecasts in our numerical example. The combination regression results are in the table below.

Forecast combination regressions	*Conditional mean*	*Conditional variance*
β_0	0.3612	0.0681
(std err)	(0.5712)	(0.1907)
β_1	−11.5910	−1.7484
(std err)	(15.7769)	(0.8544)
β_2	0.8021	2.6880
(std err)	(0.9988)	(0.8822)
R^2	0.0026	0.2878
Test $\beta_1 = 0 \cap \beta_1 = \beta_2 = 0.5$	5.1516	8.6807
p-value	0.1610	0.0339

Including both forecasts in the combination model leads to higher R^2's in both cases, as expected, though the gain for the forecast of the returns is only slight. A test that the optimal combination forecast is just an equally weighted average of the two forecasts can be conducted by testing that $\beta_0 = 0 \cap \beta_1 = \beta_2 = 0.5$. In the first regression this test has a χ^2_3-statistic (p-value) of 5.1516 (0.1610) , indicating that we cannot reject this hypothesis at the 0.05 level. For the volatility regression the χ^2_3-statistic (p-value) is 8.6807 (0.0339) , indicating that we *can* reject the null that the best combination forecast is just an equally-weighted average.

In both of these regressions we see that the first forecast in the model (forecasts A and C) have large negative coefficients, while the second, better, forecasts have positive coefficients. Be careful in attempting to interpret these parameter estimates, particularly in the volatility regression: when the regressors in a multiple regression are correlated the effects of any individual regressor is *not* solely described by its coefficient; it is a function of the coefficient on the other variables too. As Figure 11.2 showed, the volatility forecasts are highly correlated (the correlation is 0.9865) and so a large negative coefficient on $\hat{h}^c_{t+1}$ does not necessarily mean that it is a bad forecast. A better way to determine whether $\hat{h}^c_{t+1}$ is needed in this regression is to look at its t-statistic, which is -2.05, and thus significant at the 0.05 level.

12.4 Overview of chapter

This chapter addressed the problems of forecast comparison and forecast combination. These problems arise because we often have many forecasts to choose from, without a clear indication *ex ante* which individual forecast is best.

12.5 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Describe the Diebold-Mariano test for comparing forecasts.
- Interpret the results of parameter estimates and t-statistics with respect to tests for forecast evaluation and comparison.
- Interpret tests of forecast 'encompassing'
- Describe how to construct an optimal 'combination' forecast.

12.6 Test your knowledge and understanding

1. Say you run the following encompassing regression:

$$Y_{t+1} = \beta_1 \hat{Y}^a_{t+1} + \beta_2 \hat{Y}^b_{t+1} + u_{t+1}$$

There are two relevant hypotheses here:

$$\begin{aligned} H_0^A &: \beta_1 = 1 \cap \beta_2 = 0 \\ H_0^B &: \beta_1 = 0 \cap \beta_2 = 1 \end{aligned}$$

You run the regression and perform tests of the above two hypotheses. Interpret the following outcomes:

(a) Fail to reject H_0^A, reject H_0^B.

(b) Reject H_0^A, fail to reject H_0^B.

(c) Reject H_0^A, reject H_0^B.

(d) Fail to reject H_0^A, fail to reject H_0^B.

2. How would you optimally combine 3 forecasts (assuming that none encompasses the other(s)), rather than just combining 2?

3. Don't forget to check the VLE for additional practice problems for this chapter.

12.7 Solutions to activities

Activity 12.1

Given the above information we can conduct a 'Diebold-Mariano' test. This test will tell us whether forecast A has a different expected squared error than forecast B. We compute the DM test statistic as:

$$\begin{aligned} DM &= \frac{\sqrt{n}\bar{d}}{\hat{\sigma}_d} \\ &= \frac{\sqrt{100}\,(0.15)}{\sqrt{0.46}} \\ &= 2.2116 \end{aligned}$$

The test statistic is greater than 1.96, and so we reject the null hypothesis that the two forecasts are equally accurate. Because $\bar{d}$ is greater than zero we conclude that forecast A has a higher expected squared error, and so forecast B is better than forecast A.

Activity 12.2

If we had three forecasts rather than two, the encompassing regression would be

$$Y_{t+1} = \beta_1 \hat{Y}^a_{t+1} + \beta_2 \hat{Y}^b_{t+1} + \beta_3 \hat{Y}^c_{t+1} + u_{t+1}$$

and the relevant null hypotheses would be

$$\begin{aligned} H_0^A &: \beta_1 = 1 \cap \beta_2 = \beta_3 = 0 \\ H_0^B &: \beta_2 = 1 \cap \beta_1 = \beta_3 = 0 \\ H_0^C &: \beta_3 = 1 \cap \beta_1 = \beta_2 = 0 \end{aligned}$$

Chapter 13

Risk management and Value-at-Risk: Models

13.1 Introduction

Measuring and managing the exposure to risk generated by a trading desk, a structured product, or a traditional portfolio is one of the most important and interesting parts of quantiative finance. Modern risk management focusses heavily on a meaure of risk known as 'Value-at-Risk', or VaR. This is partly due to some advantages of this measure over variance, and partly due to regulation (the Basel Accords are based on VaR as a measure of risk). In this chapter we will introduce VaR formally and discuss some of the most common models for measuring VaR.

13.1.1 Aims of the chapter

The aims of this chapter are to:

- Present the formal definition of Value-at-Risk as a quantile of the distribution of returns
- Present and discuss a variety of methods for estimating VaR

13.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- State formally the definition of Value-at-Risk as a property of the distribution of returns
- Explain why VaR may be a better measure of risk than variance
- Compare and contrast some of the methods used for estimating VaR

13.1.3 Essential reading

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487]. Chapters 2 and 6 (not Sections 6.5 and 6.8).

13.1.4 Further reading

- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354]. Chapter 7.

13.1.5 References cited

- Hansen, B. E., 'Autoregressive conditional density estimation,' *International Economic Review*, 1994, 35, pp.705–730.

13.2 An alternative measure of risk: VaR

The fundamental problem in finance is trading off 'risk' against return. Returns are observable and there is little variation across types of investors (ie, investors with different utility functions) in the treatment of returns or expected returns. The main variation across investors is in their evaluation of 'risk'. For investors with quadratic utility, or investors facing multivariate normally distributed returns, risk is easily summarised as *variance*. For other investors or other distributions there may be no single number that can summarise the risk of an investment decision. Nevertheless, there is a strong desire for a single summary measure of risk. Value-at-Risk (or VaR) is one such measure.

In words, the $\alpha\%$ VaR is the cut-off such that there is only an $\alpha\%$ probability that we will see a return as low or lower. For the formal definition it is helpful to first define the inverse of a cdf:

Definition 13.1 (Quasi-inverse of a distribution function) The quasi-inverse, $F^{(-1)}$, of a distribution function, F, is defined as:

$$F^{(-1)}(\alpha) = \inf\{x : F(x) \geq \alpha\}, \text{ for } \alpha \in [0,1]$$

Note that $F^{(-1)} = F^{-1}$, the usual inverse of F, when F is strictly increasing.

The value of $F^{-1}(\alpha)$ is the 'α^{th} quantile of the distribution F'. The formal definition of VaR can now be given as:

$$\begin{aligned} \text{Unconditional} &: \quad \Pr\left[r_{t+1} \leq VaR^{\alpha}\right] = \alpha, \\ \text{so } F\left(VaR^{\alpha}\right) &= \alpha, \text{ or} \\ VaR^{\alpha} &= F^{-1}(\alpha) \\ \text{Conditional:} & \quad \Pr\left[r_{t+1} \leq VaR^{\alpha}_{t+1} | \mathcal{F}_t\right] = \alpha \\ VaR^{\alpha}_{t+1} &= F^{-1}_{t+1}(\alpha) \end{aligned}$$

Illustrations of how VaR can be extracted from a *cdf* or *pdf* are given in Figures 13.1 and 13.2.

Variance is a *moment* of the distribution of future returns, as it can be expressed in the form $E\left[(r_{t+1} - \mu)^m\right]$ or $E\left[(r_{t+1} - \mu_{t+1})^m | \mathcal{F}_t\right]$, $m = 1, 2, , ..$ VaR is a *quantile* of the disitribution of future returns, because it can be expressed as $F^{-1}_{t+1}(\alpha)$, $\alpha \in [0,1]$.

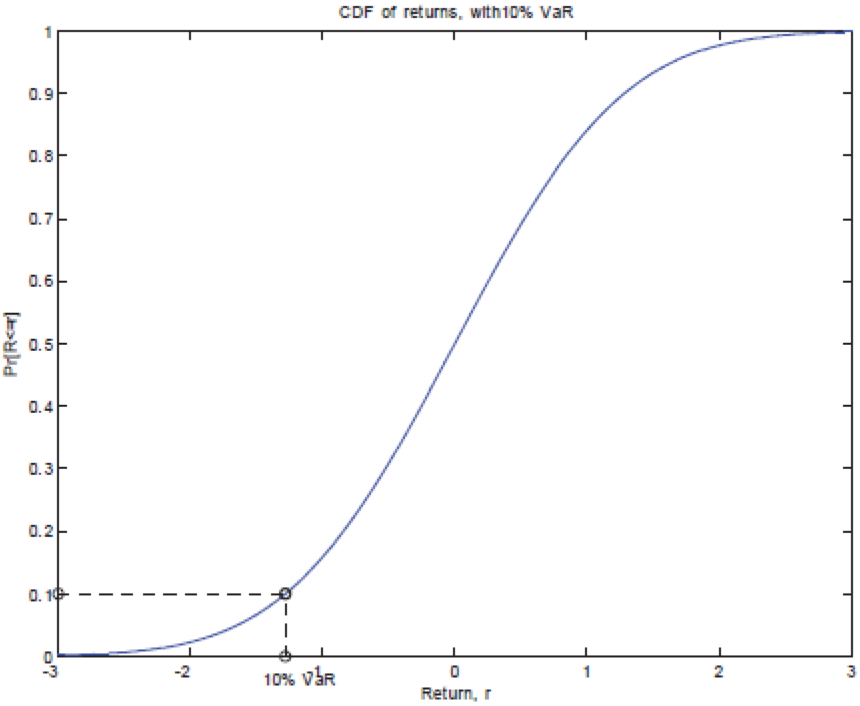


Figure 13.1: Using the CDF of returns to obtain the Value-at-Risk.

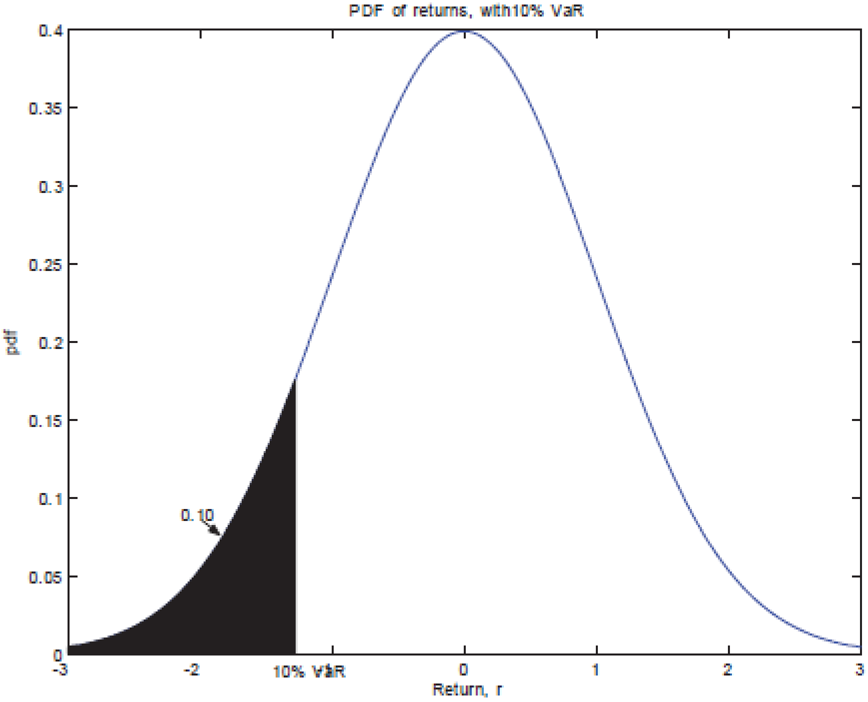


Figure 13.2: Using the PDF of returns to obtain the Value-at-Risk.

Variance as a measure of risk has the drawback that it ‘penalises’ (by taking a higher value) large positive returns in the same way as large negative returns. As investors, however, we would generally only consider ‘risk’ to be the possibility of the value of our portfolio falling, not rising. VaR overcomes this by focussing on the lower tail of the distribution of returns only.

There are two required inputs to the VaR: the probability of interest, α, and the time horizon, h. The probability level used is usually equal to 0.10, 0.05 or 0.01. The Basel Accord requires banks to use a 10-day horizon, while some internal risk management methods require the institution to forecast just the one-day-ahead VaR. In some special cases the h-step-ahead VaR is proportional to the 1-step-ahead VaR, but this is not generally true. We will focus on one-step-ahead VaR.

A variety of methods have been proposed to estimate/forecast the VaR of an asset or portfolio. Some of these are based on very unrealistic assumptions, and others are based on sophisticated modelling of the returns process. Empirical work has shown that even the simplest methods often compare well with more sophisticated methods in certain situations, in out-of-sample studies. We will review both the simple and the advanced methods for forecasting VaR.

13.3 Historical simulation

‘Historical simulation’ is possibly the simplest of all methods for estimating VaR. In using this method to estimate VaR we assume that the returns over the past m periods were *iid* from some unknown distribution F. If this is true, then we can use the empirical distribution of these returns to estimate F :

$$\hat{F}_{t+1}(r) = \frac{1}{m}\sum_{j=0}^{m-1} \mathbf{1}\{r_{t-j} \leq r\}$$

We then obtain our forecast VaR by inverting the empirical cdf:

$$\widehat{VaR}^{\alpha}_{HS,t+1} = \hat{F}^{-1}_{t+1}(\alpha)$$

In practical terms, ‘inverting the empirical cdf’ corresponds to taking the original returns, $r_1, r_2, ..., r_m$, and putting them into ascending order: $r_{(1)}, r_{(2)}, ..., r_{(m)}$, where $r_{(1)}$ is the smallest return in the sample and $r_{(m)}$ is the largest. Let $a_1 = \lfloor \alpha m \rfloor$ be αm rounded *down* to the nearest integer, and $a_2 = \lceil \alpha m \rceil$ be αm rounded *up* to the nearest integer. For example, if $m = 252$ and $\alpha = 0.05$ then $\alpha m = 12.6$, $a_1 = 12$ and $a_2 = 13$.

If αm is an integer, then $a_1 = a_2 = \alpha m$, and to find the α% VaR we look at the αm^{th} largest return, $r_{(\alpha m)}$. If αm is *not* an integer we have three options: we can either interpolate between $r_{(a_1)}$ and $r_{(a_2)}$, or we can just use $r_{(a_1)}$ or $r_{(a_2)}$.

$$\begin{aligned}
\text{Conservative} &: \widehat{VaR}^{\alpha}_{HS,t+1} = r_{(a_1)} \\
\text{Formal definition} &: \widehat{VaR}^{\alpha}_{HS,t+1} = r_{(a_2)} \\
\text{Interpolation} &: \widehat{VaR}^{\alpha}_{HS,t+1} = (a_2 - \alpha m)\, r_{(a_1)} + (\alpha m - a_1)\, r_{(a_2)}
\end{aligned}$$

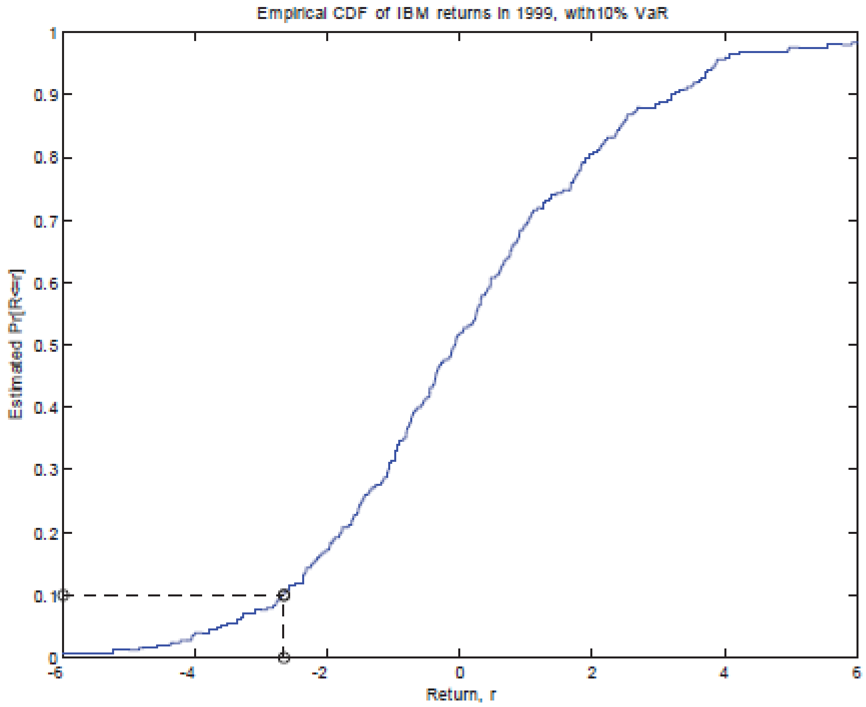


Figure 13.3: Estimating VaR using 'historical simulation'.

Historical simulation has the benefits that it is simple to implement, and it does not require us to specify the distribution of returns - we estimate this distribution nonparametrically using the empirical *cdf*, rather than estimating parameters of some specified distribution via maximum likelihood (which we will cover in the next section). However it requires the strong, and unrealistic, assumption that returns are *iid* through time, thus ruling out widely observed empirical regularities such as volatility clustering. Further, it requires an arbitrary decision on the number of observations, m, to use in estimating the *cdf*. If m is too large, then the most recent observations will get as much weight as very old observations. If m is too small then it is difficult to estimate quantiles in the tails (as is required for VaR analysis) with precision. For daily data, m is typically chosen to be between 250 and 1000. In Figure 13.3 I used $m = 250$.

13.3.1 Empirical example using IBM returns

We will use 2500 daily returns on IBM between January 1990 and December 1999 to illustrate the methods and tests in this section. In the figure above we showed how to obtain the VaR by historical simulation. In Figure 13.4 we compare the VaR estimate obtained using the full sample versus using a 'rolling window' of 250 observations. We will focus on 1% VaR for the purpose of this example.

Activity 13.1 Given the following (short) sample of returns on a hypothetical asset, compute the 10% and 20% Value-at-Risk using 'historical simulation'. Report the 'conservative', 'formal' and 'interpolation' VaRs. (Computing the 5% and 1%

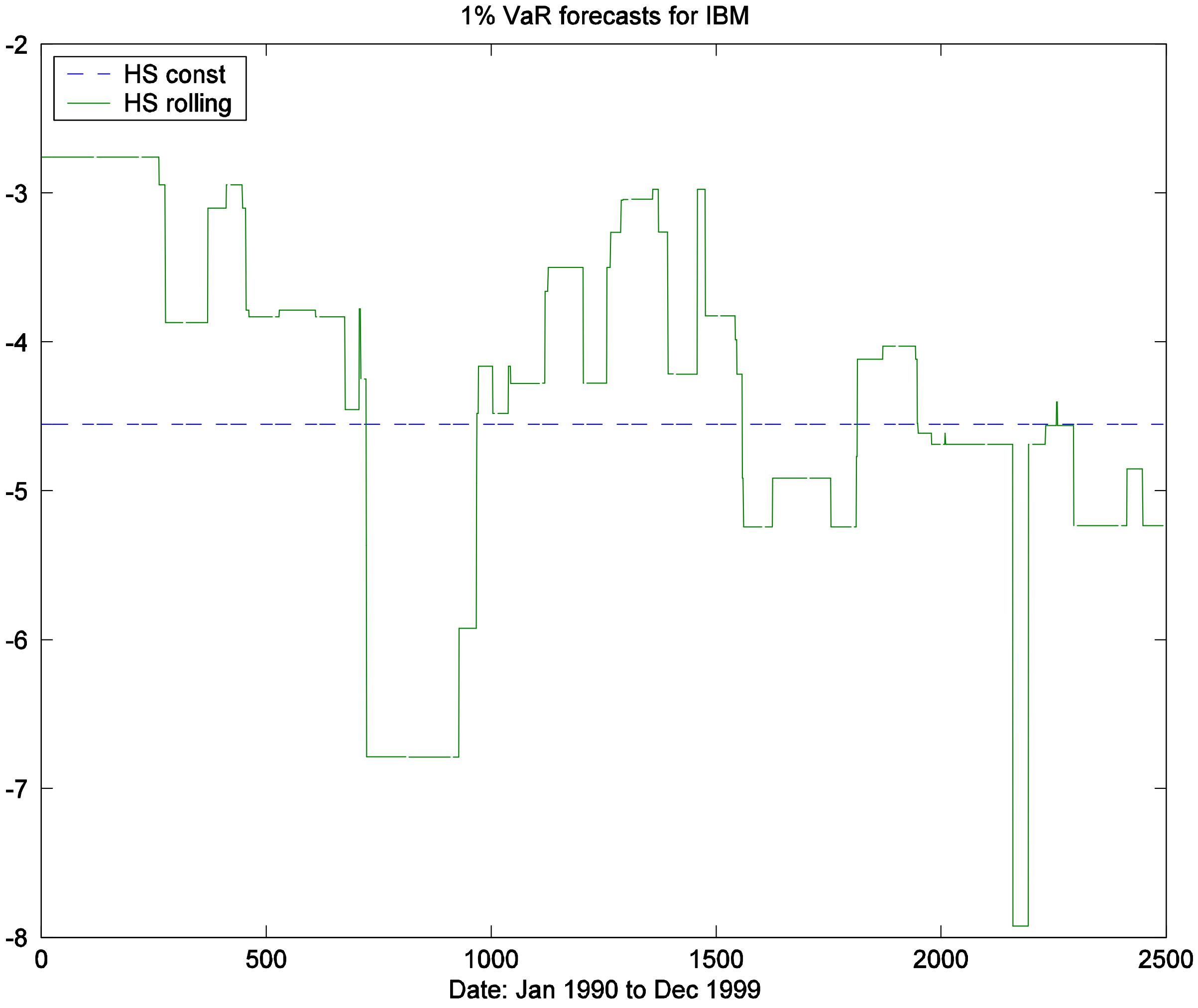


Figure 13.4: Daily 1% VaR forecasts for IBM over 1990-1999, full sample and rolling 250-day window estimates.

VaR requires more data, and so I focus on the 10% and 20% instead.)

time	return
1	1.92
2	0.53
3	-0.40
4	0.50
5	-0.53
6	0.08
7	-0.02
8	-2.77
9	2.02
10	-0.15
11	-0.78
12	0.85

13.4 Weighted historical simulation

The use of historical simulation requires the choice of m, the number of observations to use estimating the VaR. All observations older than m get zero weight, and all observations more recent than m get equal weight. This is an extreme choice for a 'weighting function' for the observations we have available. An alternative might assign higher weight to more recent observations, and a lower weight to older observations, with the weights smoothly declining in the age of the observation. Such an approach is called 'weighted historical simulation'.

For example, let m be the sample of observations to use and let λ be a smoothing parameter inside $(0, 1)$. Then we may use an exponentially declining weighting function:

$$\omega_j = \begin{cases} \lambda^j (1-\lambda) / (1-\lambda^m) & \text{if } 0 \le j < m \\ 0 & \text{else} \end{cases}$$

This function is such that the weights decline exponentially as j increases, and that $\sum_{j=0}^{m-1} \omega_j = 1$. Standard historical simulation would instead use

$$\omega_j = \begin{cases} 1/m & \text{if } 0 \le j < m \\ 0 & \text{else} \end{cases}$$

By weighting recent observations more heavily than older observations we capture more of the time-varying nature of the conditional distribution of returns. Further, since observations near m have low weight, the choice of m becomes less critical, however the choice of λ can be very important. Values of $\lambda = 0.99$ or $\lambda = 0.95$ have been used in past studies. The weighted empirical CDF is

$$\hat{F}^w_{t+1}(r) = \sum_{j=0}^{m-1} \omega_j \mathbf{1}\{r_{t-j} \le r\}$$

and the VaR forecast based on the weighted empirical CDF is again obtained by inverting the function

$$\widehat{VaR}^{\alpha}_{WHS,t+1} = \hat{F}^{w(-1)}_{t+1}(\alpha)$$

which is obtained in practice by assigning weights, ω_j, to each observation in the sample, r_{t-j}, then sorting these observations, and then finding the observation such that the sum of the weights assigned to returns less than or equal that observation is equal to α. The same three methods used for standard historical simulation (conservative/formal/interpolation) can be employed if the sum of the weights never exactly equals α.

13.5 Models based on the normal distribution

A number of VaR forecasting models are based on the normal distribution. They all generally have the following form:

$$\begin{aligned}
r_{t+1} &= \mu_{t+1} + \sigma_{t+1}\nu_{t+1} \\
\nu_{t+1}|\mathcal{F}_t &\sim iid\ N\left(0,1\right), \text{ so} \\
VaR_{t+1}^{\alpha} &= \mu_{t+1} + \sigma_{t+1}\Phi^{-1}\left(\alpha\right) \\
&= \begin{cases} \mu_{t+1} - 1.28\sigma_{t+1} & \text{for } \alpha = 0.10 \\ \mu_{t+1} - 1.65\sigma_{t+1} & \text{for } \alpha = 0.05 \\ \mu_{t+1} - 2.33\sigma_{t+1} & \text{for } \alpha = 0.01 \end{cases}
\end{aligned}$$

where μ_{t+1} and σ_{t+1} are models for the conditional mean and standard deviation, and Φ^{-1} is the inverse of the standard Normal *cdf*. For daily returns many people set $\mu_{t+1} = 0$ or $\mu_{t+1} = \mu$.

The simplest possible model based on the normal distribution is one that assumes that returns are *iid* $N\left(\mu, \sigma^2\right)$. Then the last m observations are used to estimate μ and σ^2 and the forecast VaR is given as

$$\widehat{VaR}_{N1,t+1}^{\alpha} = \hat{\mu} + \hat{\sigma}\Phi^{-1}\left(\alpha\right)$$

This method is very restrictive: it assumes not only *iid* returns, but also normally distributed returns, making it more restrictive than historical simulation which requires only the first of these assumptions. In Figure 13.5 we compare the VaR forecasts obtained using historical simulation with the constant normal model.

More sophisticated methods allow the conditional variance, and possibly the conditional mean, to vary over time. Two common choices for the model for σ_{t+1} are GARCH-type models (which we covered in an earlier chapter) or the RiskMetrics model, which assumes:

$$\begin{aligned}
\sigma_{t+1}^2 &= \lambda\sigma_t^2 + \left(1-\lambda\right)\sigma_t^2\nu_t^2 \\
\lambda &= \begin{cases} 0.94 & \text{for daily returns} \\ 0.97 & \text{for monthly returns} \end{cases}
\end{aligned}$$

This is clearly a restricted form of the more general GARCH(1,1) model. The parameter $\lambda = 0.94$ was found by JP Morgan, the developers of RiskMetrics, to work well for a variety of assets returns at the daily frequency. Its advantage over GARCH-type models is that it does not require any parameters to be estimated, at the cost of reduced flexibility.

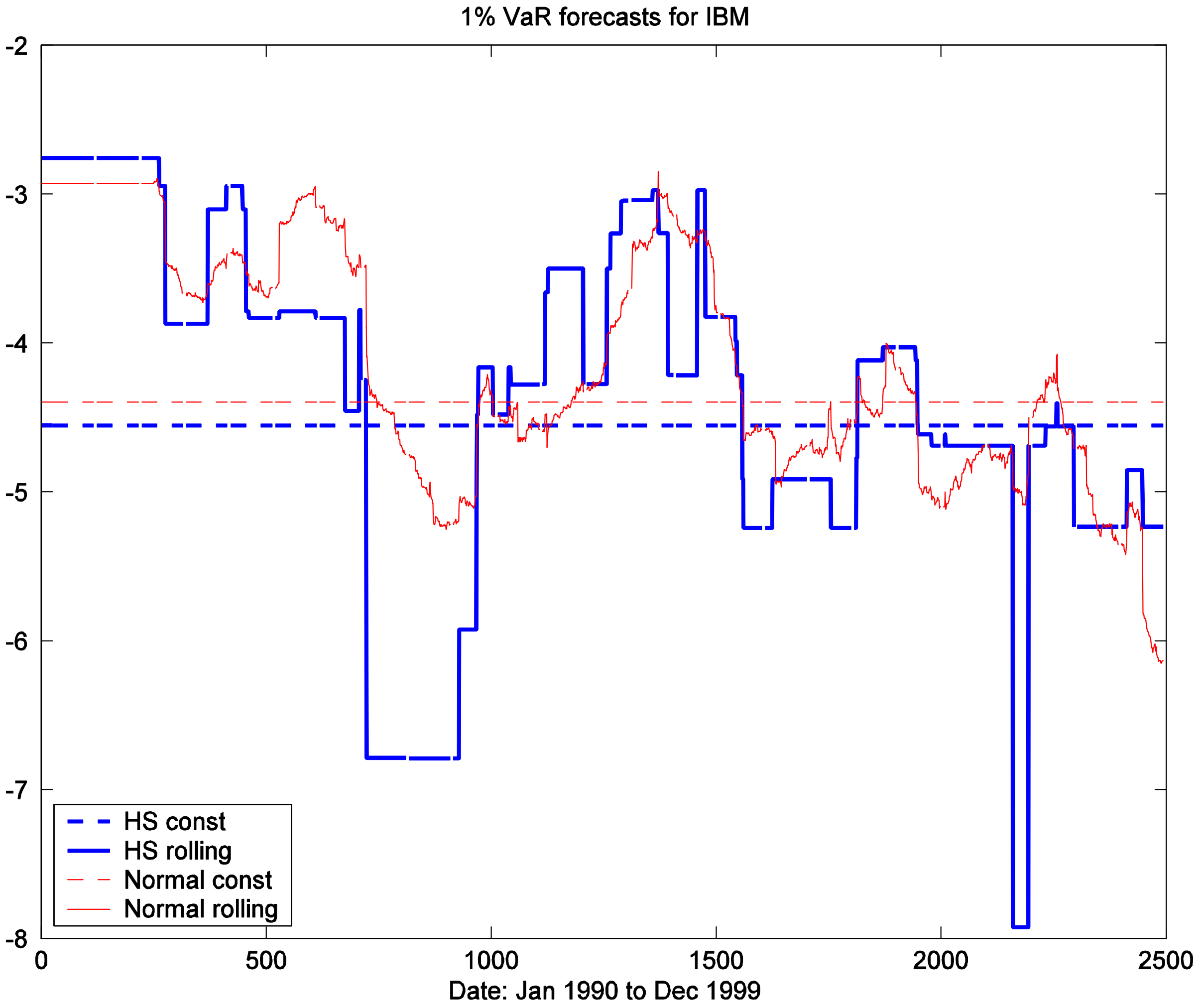


Figure 13.5: Daily 1% VaR forecasts for IBM using historical simulation and the constant normal models.

The normal distribution provides a poor fit to the unconditional distribution of returns, but generally a reasonable (though not perfect) fit to the *conditional* distribution of asset returns. (We will come back to this in a later chapter.) Thus models based on conditional normality, such as RiskMetrics and Normal-GARCH models, provide a decent fit to most asset returns. The presence of skewness and/or kurtosis in the conditional distribution of asset returns has prompted some people to instead use more flexible distributions.

13.5.1 Empirical example, continued

To illustrate, I now present VaR forecasts based on a Normal-GARCH model, using the full sample for estimation against using an 'expanding window' of data, starting with 250 observations and increasing by one each day. For the historical simulation and constant normal models I instead used a rolling window of data. This was done to capture, in an approximate way, time variation in the conditional distribution of returns. The GARCH model is designed to explicitly model this time variation, and so we use an expanding window rather than a rolling window. The only reason to use a rolling window for a GARCH-based model would be if you suspected that the parameters of the GARCH model were changing over time. This may be true over longer periods of time, but over periods of 10 years or so it is usually safe to assume that these parameters are constant.

Ideally I would re-estimate the GARCH model with each new observation, but this is quite computationally expensive, and so I re-estimate the model every 100 days. From Figure 13.6 we can see that the GARCH model is imprecisely estimated when the number of observations is less than 750, but beyond that the expanding window VaR forecast yields similar results to the full-sample model. The use of full-sample 'forecasts' is not realistic, as real forecasts must be made using *only* data available at the time the forecast is made. Wherever possible, it is always preferable to generate and analyse out-of-sample forecasts.

13.6 Models based on flexible distributions

These models are simple extensions of those based on conditional normality:

$$\begin{aligned}
r_{t+1} &= \mu_{t+1} + \sigma_{t+1}\nu_{t+1} \\
\nu_{t+1}|\mathcal{F}_t &\sim iid\ F(\eta)\text{, such that} \\
E_t[\nu_{t+1}] &= 0 \text{ and } V_t[\nu_{t+1}] = 1 \\
VaR^{\alpha}_{t+1} &= \mu_{t+1} + \sigma_{t+1}F^{-1}(\alpha;\eta)
\end{aligned}$$

And so for these types of models we simply substitute a more flexible distribution for the innovations, ν_{t+1}, for the standard normal. These flexible distributions generally have 'shape' parameters, which are used to capture conditional kurtosis and/or conditional skewness, and these parameters are either chosen by the researcher or estimated using maximum likelihood. The most widely used such distribution is the Student's t, and so the shape parameter in that case is the degrees of freedom parameter. Other common flexible distributions for the innovation term are the 'generalised error distribution' , or GED, and the skewed Student's t distribution of Hansen (1994) for example.

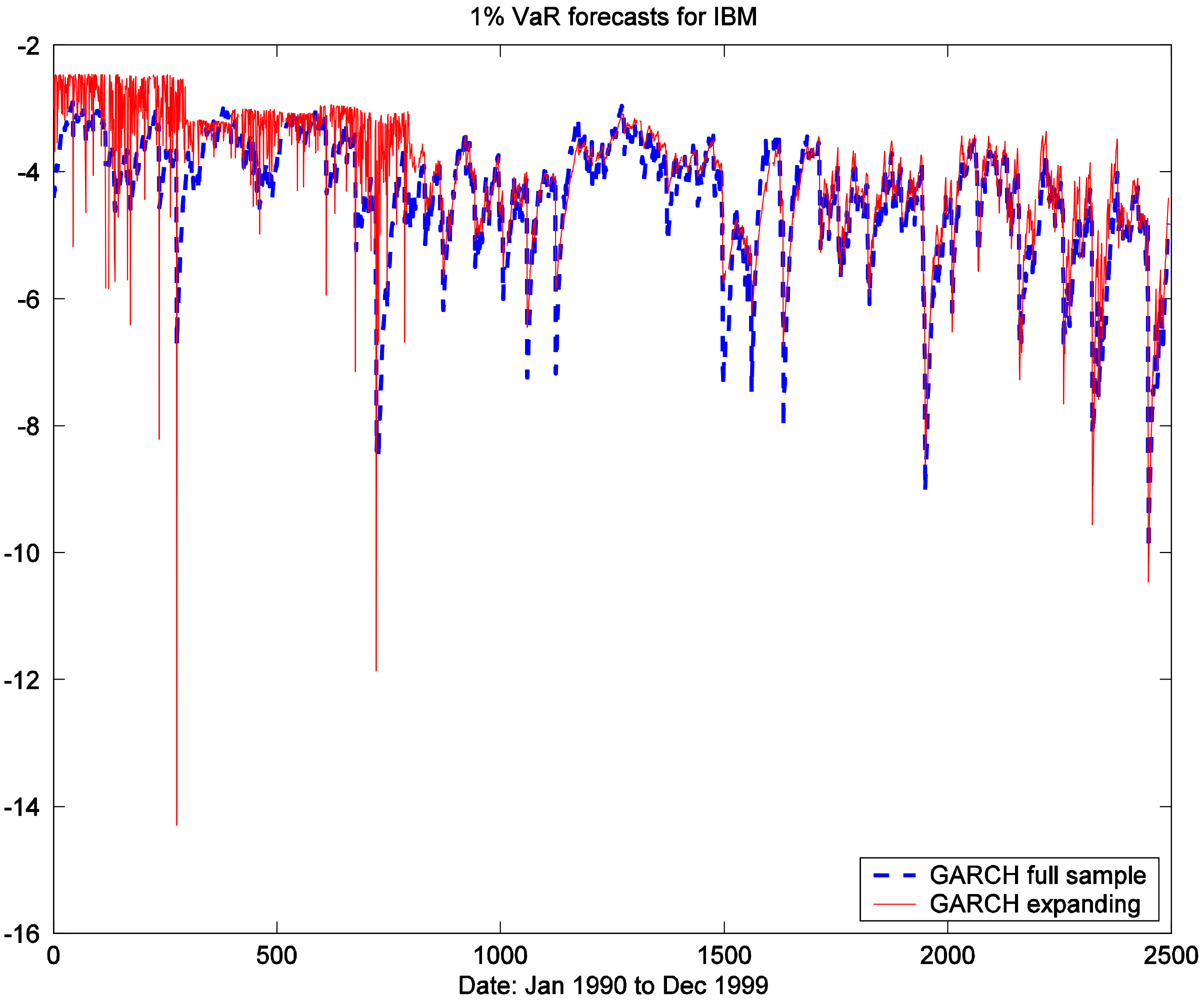


Figure 13.6: Daily 1% VaR forecasts for IBM over 1990 - 1999 using a Normal GARCH model.

To give an idea of the impact of excess kurtosis on the VaR consider the following:

Value-at-Risk using the Student's t distribution

Degrees of freedom, η	*Kurtosis*	$F^{-1}(0.10;\eta)$	$F^{-1}(0.05;\eta)$	$F^{-1}(0.01;\eta)$
∞	3	-1.28	-1.65	-2.33
30	3.23	-1.27	-1.64	-2.38
10	4.00	-1.23	-1.63	-2.48
8	4.50	-1.21	-1.62	-2.51
6	6.00	-1.18	-1.59	-2.57
5	9.00	-1.15	-1.57	-2.61
4.75	11.00	-1.14	-1.56	-2.62
4.50	15.00	-1.12	-1.54	-2.63
4.25	27.00	-1.11	-1.53	-2.64

The above table shows that, under the Student's t distribution, increasing kurtosis can either increase or decrease the VaR. For 10% and 5% VaR the level fell as kurtosis increased from 3 to 27, whereas for the 1% VaR the level increased.

The benefit of these types of models is that their flexibility, hopefully, yields a better fit to the data than the simpler models. By explicitly modelling the serial dependence in returns and volatility we hopefully capture most of the time variation in the conditional distribution. The main cost of these models is that their increased complexity may require numerical optimisation methods, and may make them less intuitive. Also, the selected parametric distribution may be mis-specified.

13.7 Semiparametric models

These types of models aim to combine the benefits of parametric modelling of the conditional mean and conditional variance, with the benefits of nonparametric density estimation. The general framework of these models is:

$$\begin{aligned} r_{t+1} &= \mu_{t+1} + \sigma_{t+1}\nu_{t+1} \\ \nu_{t+1}|\mathcal{F}_t &\sim iid\ F \end{aligned}$$

Parametric models for the conditional mean and variance are estimated, and the time series of estimated standardised innovations are obtained:

$$\hat{\nu}_{t+1} \equiv \frac{r_{t+1} - \hat{\mu}_{t+1}}{\hat{\sigma}_{t+1}}$$

This method is sometimes called 'filtered historical simulation', because the standardised innovations are returns that have had their means and variances 'filtered' away. These innovations are assumed to be *iid* according to some unknown distribution F, which we estimate via the empirical cdf:

$$\hat{F}_{\nu,t+1}(\nu) = \frac{1}{t}\sum_{j=0}^{t-1} \mathbf{1}\{\hat{\nu}_{t-j} \leq \nu\}$$

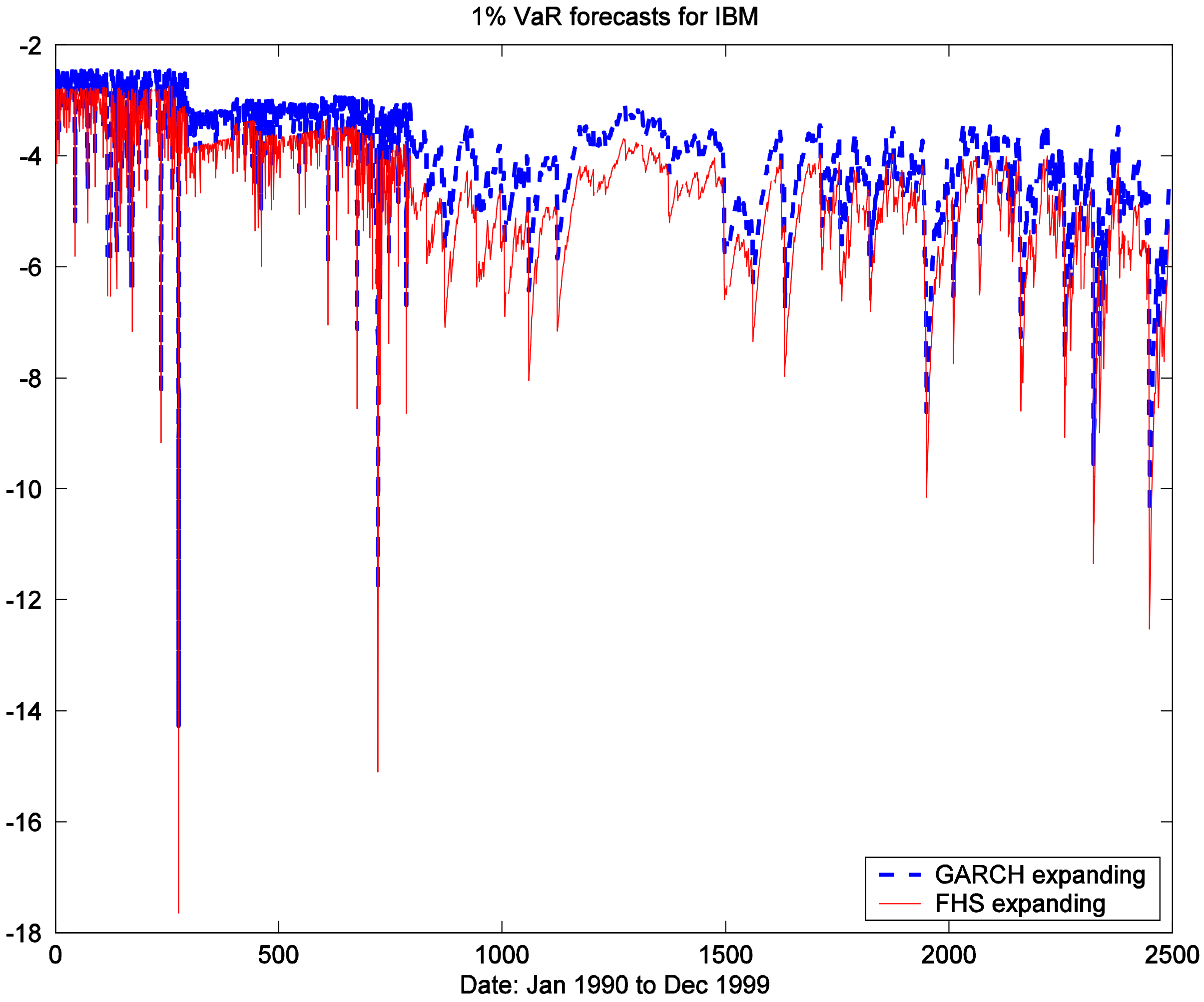


Figure 13.7: Daily 1% VaR forecasts for IBM over Jan 1990 to Dec 1999, using expanding window filtered historical simulation and Normal-GARCH models.

And the $\alpha\%$ VaR is:

$$\widehat{VaR}^{\alpha}_{FHS,t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\hat{F}^{-1}_{\nu,t+1}(\alpha)$$

In Figure 13.7 I contrast the filtered historical simulation VaR forecast, where a constant mean and a GARCH(1,1) model is used for the filtering, with the Normal-GARCH VaR forecast. The figure clearly shows that the empirical distribution of the standardised residuals has a fatter left tail than the normal distribution.

Activity 13.2 Given the following (short) sample of returns on a hypothetical asset, compute the 10% and 20% Value-at-Risk using 'filtered historical simulation'. Report the 'conservative', 'formal' and 'interpolation' VaRs for the return at time $t = 12$.

time	return	μ_t	σ_t
1	-0.05	0.12	1.41
2	2.49	0.09	1.37
3	1.81	0.28	1.40
4	0.45	0.21	1.39
5	-0.70	0.12	1.35
6	-0.15	0.04	1.33
7	-1.16	0.09	1.30
8	1.14	0.00	1.29
9	-0.40	0.19	1.28
10	-3.61	0.05	1.26
11	0.47	-0.19	1.40
12	-0.30	0.15	1.37

13.8 Overview of chapter

This chapter introduced Value-at-Risk as a measure of the risk of an asset. We covered its formal definition as a quantile of the predictive distribution of an asset return, and discussed a variety of models and methods for estimating VaR.

13.9 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- State formally the definition of Value-at-Risk as a property of the distribution of returns.
- Explain why VaR may be a better measure of risk than variance.
- Compare and contrast some of the methods used for estimating VaR.

13.10 Test your knowledge and understanding

1. Your firm trades in equities and is considering implementing a new VaR model for risk management. Write a brief proposal for **one** VaR model covered in this course. Describe the model's benefits and its shortcomings.
2. Compare and contrast 'historical simulation' with one other model for VaR. Describe in detail the assumptions required for each model, and which model you would recommend for use with equity returns.
3. Don't forget to check the VLE for additional practice problems for this chapter.

13.11 Solutions to activities

Activity 13.1

This problem can be solved with the following information

1.

α	10%	20%
αm	1.2	2.4
$a_1 = \lfloor \alpha m \rfloor$	1	2
$a_2 = \lceil \alpha m \rceil$	2	3
$r_{(1)}$	-2.77	
$r_{(2)}$	-0.78	
$r_{(3)}$	-0.53	
Conservative VaR	-2.77	-0.78
Formal VaR	-0.78	-0.53
Interpolation VaR	-2.37	-0.68

Activity 13.2

This problem can be solved by first computing the values for $\nu_t = (r_t - \mu_t) / \sigma_t$ and then with the following information:

1.

α	10%	20%
αm	1.2	2.4
$a_1 = \lfloor \alpha m \rfloor$	1	2
$a_2 = \lceil \alpha m \rceil$	2	3
$\nu_{(1)}$	-2.90	
$\nu_{(2)}$	-0.96	
$\nu_{(3)}$	-0.61	
Conservative VaR for ν_t	-2.90	-0.96
Formal VaR for ν_t	-0.96	-0.61
Interpolation VaR for ν_t	-2.52	-0.82
Conservative VaR for r_{12}	-3.83	-1.17
Formal VaR for r_{12}	-1.17	-0.68
Interpolation VaR for r_{12}	-3.30	-0.97

Noting that the VaR for returns at time t is equal to the VaR for ν_t (which is assumed to be constant in filtered historical simuation), multiplied by σ_t and added to μ_t.

Chapter 14

Risk management and Value-at-Risk: Backtesting

14.1 Introduction

In the previous chapter we covered several models for measuring and forecasting Value-at-Risk (VaR). An important part of managing risk is testing how well your risk models are performing, a task known as 'backtesting' in the risk management literature (it is called model or forecast evaluation in other applications). These tests can also be useful for indicating ways to improve risk models. This chapter will cover some simple and some more advanced methods for backtesting VaR models.

14.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce the 'Hit' variable as a key method for evaluating (backtesting) VaR models
- Present tests of 'unconditional' and 'conditional' coverage based on the Hit variable
- Discuss the use of Diebold-Mariano tests to compare VaR forecasts, and to highlight the correct loss function to use for these tests

14.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Describe methods for evaluating ('backtesting') VaR forecasts.
- Describe how 'unconditional coverage' and 'conditional coverage' probabilities can be used to backtest VaR estimates.
- Describe how to formally test that one VaR method is better than another.

14.1.3 Essential reading

- Christoffersen, P.F. *Elements of Financial Risk Management.* (Academic Press, London, 2011) second edition [ISBN 9780123744487]. Chapter 13, Sections 13.1–13.3.

14.1.4 References cited

- Christoffersen, P. F., 'Evaluating interval forecasts,' *International Economic Review*, 1998, 39, pp.841–862.
- Christoffersen, P. F., J. Hahn and A. Inoue, 'Testing and Comparing Value-at-Risk Measures,' *Journal of Empirical Finance*, 8, 2001, pp.325–342.
- Engle, R. F., and S. Manganelli, 'CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles,' *Journal of Business & Economic Statistics*, 22(4), 2004, pp.367–381

14.2 Evaluating VaR forecasts

In the previous chapter we covered an array of methods for forecasting VaR. The next step is to evaluate how our forecasts performed out-of-sample. (This is sometimes referred to as 'backtesting' in the practitioner literature.) For VaR forecasting, we can do this by focussing on the 'hit' sequence, defined as:

$$Hit_{t+1} = \mathbf{1}\left\{r_{t+1} \leq VaR_{t+1}\right\} = \begin{cases} 1, & \text{if } r_{t+1} \leq VaR_{t+1} \\ 0, & \text{else} \end{cases}$$

If the VaR forecast is optimal, then the sequence of hits will be:

$$Hit_{t+1} \sim iid\ Bernoulli\left(\alpha\right)$$

That is, the probability of getting a hit at each point in time is independent of time t information, and will equal α.

14.2.1 Unconditional coverage tests

If a $\alpha\%$ VaR forecast is correct, then approximately $\alpha\%$ of observations should have exceeded the VaR. Note that it is optimal to have exceedences of the VaR - zero exceedences is *not* a property of an optimal VaR forecast. One way of testing whether we had the right number of exceedences on average is to test the hypothesis:

$$\begin{aligned} H_0 &: E\left[Hit_t\right] = \alpha \\ \text{vs. } H_a &: E\left[Hit_t\right] \neq \alpha \end{aligned}$$

This is a test of the unconditional mean of a random variable, and we can use a simple $t - test$:

$$tstat = \frac{\frac{1}{T}\sum_{t=1}^{T} Hit_t - \alpha}{\sqrt{\alpha\left(1-\alpha\right)/T}} \to N\left(0,1\right) \text{ as } T \to \infty$$

where in the denominator we have used the fact that under the null of the VaR forecast being optimal we know that the hits are *iid* and that $V\left[Hit_t\right] = \alpha\left(1-\alpha\right)$. A widely-quoted related metric is the so-called 'violation ratio', which is

$$\begin{aligned} \widehat{VR} &\equiv \frac{1}{\alpha}\frac{1}{T}\sum_{t=1}^{T} Hit_t \\ VR_0 &\equiv E\left[\widehat{VR}\right] = \frac{E\left[Hit_t\right]}{\alpha} \end{aligned}$$

if the VaR forecast is optimal, then $E[Hit_t] = \alpha$, and so the observed violation ratio, $\widehat{VR}$, should be near one. Given that the violation ratio is a simple re-scaling of the number of hits over the sample, testing $VR_0 = 1$ is exactly equivalent to testing $E[Hit_t] = \alpha$. We will focus on tests based on the hits themselves, rather than the violation ratio.

The above test relies on the fact that the mean of any variable, subject to basic conditions, is asymptotically normal. However we may be able to obtain a more powerful test by using the stronger implication stated above, that if the VaR forecasts are perfect then the sequence of hits are *iid Bernolli* (α). Let us now follow Christoffersen (1998) and consider testing:

$$\begin{array}{rrl} & H_0 & : \; Hit_t \sim iid\ Bernoulli\,(\alpha) \\ \text{vs.} & H_a & : \; Hit_t \sim iid\ Bernoulli\,(\pi) \end{array}$$

The pmf of a Bernoulli(π) random variable is:

$$f(Hit;\pi) = (1-\pi)^{1-Hit}\,\pi^{Hit}$$

and so the log-likelihood for a sample of T observations is:

$$\log L(\pi) = \sum_{t=1}^{T}\left\{(1-Hit_t)\log(1-\pi) + Hit_t\log\pi\right\}$$

and the maximum likelihood estimator of π is easily derived to be:

$$\hat{\pi} = \frac{1}{T}\sum_{t=1}^{T} Hit_t$$

We can then conduct a likelihood ratio test that $\pi = \alpha$. The LR test statistic is χ^2_1 under the null hypothesis, and is computed as:

$$LR = 2\left(\log L(\hat{\pi}) - \log L(\alpha)\right)$$

and we reject if $LR > \chi_1^{2^{-1}}(0.95) = 3.84$.

> **Activity 14.1** Define the 'population violation ratio' as $VR_0 \equiv E\left[\widehat{VR}\right]$. Show formally that testing $VR_0 = 1$ is exactly equivalent to testing $E[Hit_t] = \alpha$.

14.2.2 Conditional coverage tests

In the previous section we merely tested whether the unconditional mean of the hits was equal to α, the desired proportion of hits. This ignored the fact that we have an even stronger implication of optimality, that of *iid*-ness. If the VaR_{t+1} forecast is optimal, then nothing in the time t information set should help to predict whether there will be a hit at time $t+1$; all of that information should already be in VaR_{t+1}. We can thus test whether the unconditional probability of a hit is α, but also whether anything in the time t information set is correlated with the hit sequence.

> **Activity 14.2** Show that if a VaR forecast as correct *conditional* coverage then it must have correct *unconditional* coverage.

One way of conducting such a test is by a type of generalised Mincer-Zarnowitz regression, where we regress the hit variable on a constant, and anything in the time t information set, such as lagged hits or current and lagged VaR forecasts. For example:

$$Hit_{t+1} = \beta_0 + \beta_1 Hit_t + \beta_2 \widehat{VaR}_{t+1} + u_{t+1}$$

and then test:

$$\begin{aligned} H_0 &: \beta_0 = \alpha \cap \beta_1 = \beta_2 = 0 \\ \text{vs. } H_a &: \beta_0 \neq \alpha \cup \beta_1 \neq 0 \cup \beta_2 \neq 0 \end{aligned}$$

This type of test was proposed by Engle and Manganelli (2004). As for the t-test discussed above, this regression-based test uses the fact that OLS regression can be applied to both binary variables (such as the hit variable) and continuous random variables. We could also consider 'logit' or 'probit' regressions here, given the binary nature of the dependent variable.

An alternative test is that of Christoffersen (1998), who again considered capturing the fact that the hit variable is *iid Bernoulli* (α) under the null hypothesis. He suggested modelling the hit series as a 'first-order Markov chain'. In this context, this means allowing for serial correlation in the hits, by letting the probability that $Hit_{t+1} = 1$ depend on Hit_t. A first-order Markov chain is described by its 'transition matrix':

$$\begin{aligned} \Pi &= \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}, \text{ where} \\ \pi_{01} &= \Pr\left[Hit_{t+1} = 1 | Hit_t = 0\right] \\ \pi_{11} &= \Pr\left[Hit_{t+1} = 1 | Hit_t = 1\right] \end{aligned}$$

If there is no serial correlation in the hits, then $\pi_{01} = \pi_{11}$. Further, if the proportion of hits is correct, then $\pi_{01} = \pi_{11} = \alpha$.

We can estimate π_{01} and π_{11} by maximum likelihood:

$$\begin{aligned} f\left(Hit_t | Hit_{t-1}, \pi_{01}, \pi_{11}\right) &= \left(1-\pi_{01}\right)^{(1-Hit_{t-1})(1-Hit_t)} \times \pi_{01}^{(1-Hit_{t-1})Hit_t} \\ &\quad \times \left(1-\pi_{11}\right)^{Hit_{t-1}(1-Hit_t)} \times \pi_{11}^{Hit_{t-1}Hit_t} \\ \log L(\Pi) &= \sum_{t=2}^{T} \log f\left(Hit_t | Hit_{t-1}, \pi_{01}, \pi_{11}\right) \\ &= T_{00} \log\left(1-\pi_{01}\right) + T_{01} \log \pi_{01} + T_{10} \log\left(1-\pi_{11}\right) + T_{11} \log \pi_{11} \\ \text{where } T_{ij} &= \sum_{t=2}^{T} \mathbf{1}\left\{Hit_{t-1} = i \cap Hit_t = j\right\} \\ \hat{\Pi} &= \begin{bmatrix} 1-\hat{\pi}_{01} & \hat{\pi}_{01} \\ 1-\hat{\pi}_{11} & \hat{\pi}_{11} \end{bmatrix} \\ \hat{\pi}_{01} &= \frac{T_{01}}{T_{00}+T_{01}}, \ \hat{\pi}_{11} = \frac{T_{11}}{T_{10}+T_{11}} \\ \text{Let } \Pi_0 &= \begin{bmatrix} 1-\alpha & \alpha \\ 1-\alpha & \alpha \end{bmatrix} \end{aligned}$$

then we can test whether $\hat{\Pi}$ is significantly different from Π_0 by a likelihood ratio test:

$$LR = 2\left(\log L\left(\hat{\Pi}\right) - \log L\left(\Pi_0\right)\right) \sim \chi_2^2$$

and we reject if $LR > \chi_2^{2^{-1}}(0.95) = 5.99$.

> **Activity 14.3** Give an example of a sequence of 'hits' for a 10% VaR model, which has correct unconditional coverage but incorrect conditional coverage.

14.2.3 Empirical example, continued

Here I present the results of the tests presented in this section, applied to each of the out-of-sample VaR forecasts we constructed for IBM returns above. These out-of-sample forecasts were all based on an initial sample of 250 returns, and so I only evaluated the forecasts using the last 2250 observations.

	Historical simulation	Constant Normal	Normal GARCH	Filtered HS
Unc t-test	0.11	0.01	0.00	0.75
Unc Chris. test	0.13	0.02	0.00	0.75
Cond regression test	0.00	0.02	0.01	0.00
Cond Chris. test	0.23	0.06	0.00	0.74

The above table shows that the constant Normal and Normal GARCH VaR forecasts failed the unconditional tests, suggesting that normality may not be a reasonable assumption for this data. Further, all four models failed the regression-based test of conditional coverage. This is somewhat surprising, and for the FHS model it could be due to the GARCH model used being insufficient to remove all conditional variance dynamics. Alternatively, it could signal the presence of other time-varying components of the conditional distribution, such as conditional skewness or kurtosis.

14.3 Comparing VaR forecasts

The above tests of VaR forecast optimality may help eliminate some VaR forecasts as being clearly sub-optimal. For our empirical application involving IBM these tests consistently reject the 'constant Normal' and 'Normal GARCH' forecasts, while the results for the remaining forecasts are mixed. If we relied solely on the regression-based test of correct conditional coverage then all six forecasting methods would be rejected. This leads us naturally to the alternative question of *relative* performance. The availability of so many forecasts also leads to the question of VaR forecast combination. We consider these two problems in the following sections.

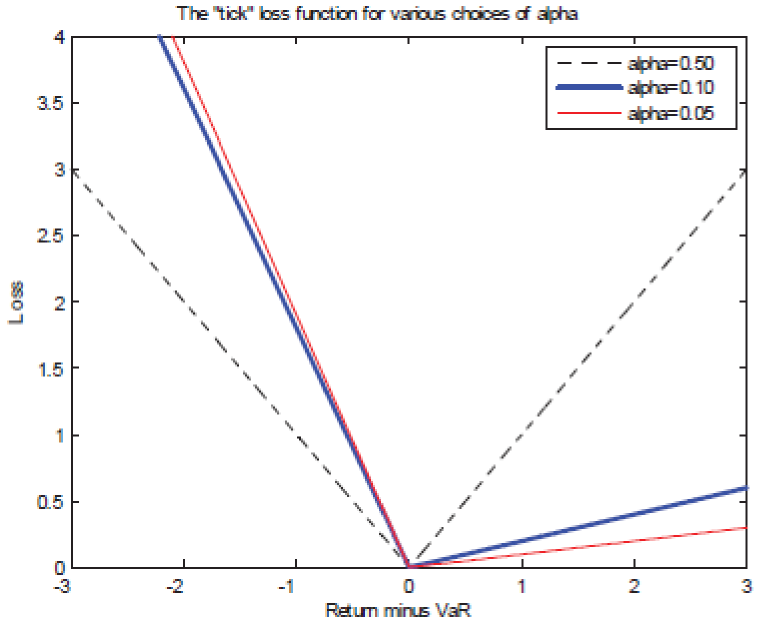


Figure 14.1: The 'tick' loss function for various values of α.

14.3.1 Diebold-Mariano tests for VaR forecasts

As with other types of forecasts, we can employ a Diebold-Mariano (DM) test to compare VaR forecasts. Care needs to be taken when considering the appropriate loss function to use for VaR forecast comparison. Some papers in the literature (both practitioner and academic) have suggested loss functions such as

$$L\left(r_t, \widehat{VaR}_t\right) = \begin{cases} \left(r_t - \widehat{VaR}_t\right)^2, & \text{if } \; r_t \leq \widehat{VaR}_t \\ 0, & \text{else} \end{cases}$$

with the motivation that only exceedences of the VaR should be 'punished'; events when the VaR was not violated are given zero penalty. This may at first seem reasonable, but upon closer inspection it becomes clear that the optimal VaR forecast under such a loss function is one that is equal to $-\infty$ every day. That forecast would never be violated, and so would result in exactly zero loss, which is the lowest possible loss. Of course, setting $\widehat{VaR}_t = -\infty \; \forall \; t$ is not an optimal VaR forecast.

The appropriate loss function to use for VaR comparison is the 'tick' loss function, illustrated in Figure 14.1, and defined as:

$$L\left(r_t, \widehat{VaR}_t\right) = 2\left(\alpha - \mathbf{1}\left\{r_t \leq \widehat{VaR}_t\right\}\right) \cdot \left(r_t - \widehat{VaR}_t\right)$$

To compare two VaR forecasts we then construct

$$d_t = L\left(r_t, \widehat{VaR}_{1,t}\right) - L\left(r_t, \widehat{VaR}_{2,t}\right)$$

and test

$$\begin{aligned} H_0 &: E\left[d_t\right] = 0 \\ \text{vs. } H_a &: E\left[d_t\right] \neq 0 \end{aligned}$$

using a standard DM test. Christoffersen, Hahn and Inoue (2001) provide further details on tests for comparing VaR forecasts.

14.3.2 Empirical example, continued

I conducted DM tests for all 6 possible pairs of VaR forecasts, and I present the t-statistics from these tests in the table below. The $(i,j)^{th}$ element of this table is based on the loss difference constructed as $d_t = L\left(r_t, \widehat{VaR}_{i,t}\right) - L\left(r_t, \widehat{VaR}_{j,t}\right)$, and so a negative t-statistic indicates that the i^{th} model is better than the j^{th} model.

	Historical simulation	Constant Normal	Normal GARCH	Filtered HS
Historical sim	–	0.27	0.51	0.82
Constant normal	-0.27	–	0.27	0.51
Normal GARCH	-0.51	-0.27	–	0.52
Filtered HS	-0.82	-0.51	-0.52	–
Average Loss × 100	15.39	15.30	15.17	15.01

In the bottom row of the table I present the average loss incurred, using the 'tick' loss function, from each of these forecasts. This can be interpreted as the quantile equivalent of the mean squared forecast error for standard types of forecasts. These average loss figures reveal that the best forecast was the filtered historical simulation forecast. Standard historical simulation was the worst-performing forecast. The DM test is then used to determine whether any of these differences in performance are statistically significant.

The table shows that none of the t-statistics are greater than 1.96 in absolute value, and so in no case can we reject the null hypothesis that two competing forecasts are equally accurate. Thus, while the VaR forecast based on filtered historical simulation performs the best, the out-performance is not statistically significant for this application.

Activity 14.4 You work in the risk management division of a large bank, and one of your colleagues is discussing the penalties incurred when the VaR is violated. He suggests that the following is a reasonable assessment of the loss from a VaR forecast:

$$L\left(r_t, \widehat{VaR}_t\right) = \begin{cases} 10 \times \left| r_t - \widehat{VaR}_t \right|, & \text{if } r_t \leq \widehat{VaR}_t \text{ and } r_{t-1} \leq \widehat{VaR}_{t-1} \\ \left| r_t - \widehat{VaR}_t \right| & \text{if } r_t \leq \widehat{VaR}_t \text{ but } r_{t-1} > \widehat{VaR}_{t-1} \\ 0, & \text{else} \end{cases}$$

Derive the optimal 'VaR' forecast under this loss function. How you would explain to your colleague that this is not a reasonable loss function for Value-at-Risk?

14.4 Overview of chapter

This chapter presented methods for evaluating and comparing Value-at-Risk forecasts, tasks known as 'backtesting' in the VaR literature. We considered both 'unconditional' and 'conditional' tests, as well as Diebold-Mariano tests to compare VaR forecasts.

14.5 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Describe methods for evaluating ('backtesting') VaR forecasts.
- Describe how 'unconditional coverage' and 'conditional coverage' probabilities can be used to backtest VaR estimates.
- Describe how to formally test that one VaR method is better than another.

14.6 Test your knowledge and understanding

1. Describe two different tests to evaluate Value-at-Risk forecasts, and discuss their pros and cons.
2. Let $\widehat{VaR}_t$ denote a forecast of the (conditional) 5% Value-at-Risk of an asset return, r_t. Show that if $\widehat{VaR}_t$ is equal to the **true** Value-at-Risk then the indicator variable
$$Hit_t \equiv \mathbf{1}\left\{r_t \leq \widehat{VaR}_t\right\}$$
will have conditional mean equal to 0.05.
3. Don't forget to check the VLE for additional practice problems for this chapter.

14.7 Solutions to activities

Activity 14.1

$$\begin{aligned}
VR_0 &\equiv E\left[\widehat{VR}\right] \\
&= E\left[\frac{\frac{1}{T}\sum_{t=1}^{T} Hit_t}{\alpha}\right] \\
&= \frac{1}{\alpha T}\sum_{t=1}^{T} E\left[Hit_t\right] \\
&= \frac{E\left[Hit_t\right]}{\alpha}
\end{aligned}$$

Thus $VR_0 = 1$ if and only if $E\left[Hit_t\right] = \alpha$. This implies that any test of the hypothesis that $E\left[Hit_t\right] = \alpha$ can be interpreted as a test that the population violation ratio is equal to one.

Activity 14.2

Correct conditional coverage implies $E\left[Hit_t|\mathcal{F}_{t-1}\right] = \alpha$. If we take unconditional expectations of both sides and use the law of iterated expectations we obtain:

$$\begin{aligned} \alpha &= E\left[Hit_t|\mathcal{F}_{t-1}\right] \\ \text{so } \alpha &= E\left[E\left[Hit_t|\mathcal{F}_{t-1}\right]\right] \\ &= E\left[Hit_t\right] \text{ by the LIE} \end{aligned}$$

And so any VaR forecast that has correct conditional coverage is guaranteed to have correct unconditional coverage.

Activity 14.3

There are many possible answers to this question. A simple answer is a series of hits where the first 10% of observations equal one and the remaining 90% of observations equal zero. In that case the series will have the right **proportion** of hits (and so it will have correct **unconditional** coverage) but the hits will be serially correlated and so they will not have correct conditional coverage. A second example would be a series of hits which equals zero everywhere except for every 10^{th} observation, when it equals one. In that case there will again be the right **proportion** of hits, but there will be serial correlation at the 10^{th} lag (in fact, you could predict perfectly when the next hit will occur in this series) and so the conditional coverage will not be correct.

Activity 14.4

The colleague presumably believes that it is costly to have a VaR violation, and even more costly to have two VaR violations in a row, but it is not costly at all when there is no violation. If the loss function truly was:

$$L\left(r_t, \widehat{VaR}_t\right) = \begin{cases} 10 \times \left|r_t - \widehat{VaR}_t\right|, & \text{if } r_t \leq \widehat{VaR}_t \text{ and } r_{t-1} \leq \widehat{VaR}_{t-1} \\ \left|r_t - \widehat{VaR}_t\right|, & \text{if } r_t \leq \widehat{VaR}_t \text{ but } r_{t-1} > \widehat{VaR}_{t-1} \\ 0, & \text{else} \end{cases}$$

then the optimal 'VaR' forecast is $\widehat{VaR}_t = -\infty \; \forall \; t$. With such a forecast the VaR would never be violated, and so the loss each day would exactly equal zero, which is the lowest possible value for this loss function. Thus $\widehat{VaR}_t = -\infty$ is the optimal forecast under this loss function.

This loss function is not reasonable for two main reasons. Firstly, we know the true VaR is a quantile of the distribution of returns, we need to choose a loss function that yields this quantile as the optimal estimate. The appropriate loss function is the 'tick' loss function. Secondly, the loss function suggested by the colleague ignores the fact that there is indeed a cost when the VaR is not violated. Some institutions have to put aside more capital when their VaR is larger (in absolute terms) and so setting $\widehat{VaR}_t = -\infty$ would lead to too much capital being put aside (and thus forfeiting income on this capital). There may also be a cost in terms of credibility: setting

$\widehat{VaR}_t = -\infty$ every day does not sound like a very reasonable forecast, and so there may be a reputation cost to setting VaR so high that it is **never** violated.

Chapter 15
Modelling high frequency financial data: Diurnality

15.1 Introduction

Many of the models for forecasting financial data are based on prices observed at the daily or monthly frequencies. However some market participants face prices at much higher frequencies: hedge funds, proprietary trading desks, market makers, algorithmic traders, and regulators all have an interest in high frequency *intra-daily* prices. Many of the methods developed for lower frequency data are applicable to high frequency data, but there are three key places where differences exist, and we will study these in the next three chapters.

The first problem that arises in some analyses of high frequency data is seasonality. Seasonality is a well-studied problem in macro- and micro-econometrics, but is not usually a concern for financial econometricians. Intra-daily patterns (called '*diurnality*' rather than 'seasonality') in certain variables are significant and must be considered. Diurnality in high frequency returns has been found to be prominent in the conditional variance, in bid-ask spreads (the difference between the bid and ask prices quoted for a stock), and in trade durations. We will study these in this chapter.

15.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce you to high frequency, intra-daily, financial data
- Discuss the issue of 'diurnality' that arises when studying high frequency data
- Consider two regression-based approaches for capturing diurnality

15.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Describe two methods for capturing seasonality/diurnality in returns and volatility
- Interpret empirical tests for the presence of seasonality/diurnality

15.1.3 Essential reading

- Diebold, F.X. *Elements of Forecasting.* (Thomson South-Western, Canada, 2006) fourth edition [ISBN 9780324323597]. Chapter 6 and 13

15.1.4 Further reading

- Taylor, Stephen J. *Asset Price Dynamics, Volatility and Prediction.* (Princeton University Press, Oxford, 2005) [ISBN 9780691134796]. Chapter 12

- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354]. Chapter 5

15.2 Diurnality in asset returns

When studying asset returns at very high frequencies, it becomes apparent that returns at certain times of the day are quite different from those at other times of the day. If these features are a pure function of the time of day, then they are said to be 'diurnal' features. This is the time-of-day equivalent of well-known seasonality in some monthly data (eg, sales of ice cream, number of miles driven by individuals, etc.)

To analyse the problem of diurnality in intra-daily asset returns we will use fixed-interval returns, such as 5-minute returns, for simplicity. In Figure 15.1 I plot the average 5 minute return and the average log squared difference between 5 minute returns and their expectation, using high frequency data on IBM for the calendar year 2009. These figures suggest that there is a strong diurnal pattern for volatility, but no such pattern for average returns.[1] We now turn to two simple methods for modeling diurnality.

15.3 Dummy variable models

The simplest way to deal with seasonality/diurnality is to use 'dummy variables', otherwise known as 'indicator' or 'binary' variables. These are variables that take the value 1 in certain situations and 0 elsewhere. In modelling seasonality we might use one indicator variable for summer months, one for spring, one for winter and one for autumn. For diurnality we could use one for the first half-hour of trade, one for the second half-hour, and so on.

Say we have a variable called $TIME_t$, which contains the time that trade t took place

[1] Note that the plot for volatility is in logs. The min and max values are -2.2 and -0.06, which correspond to 5-min standard deviations of 0.11% and 0.94%, and to annualised volatility of 15.5% to 132%.

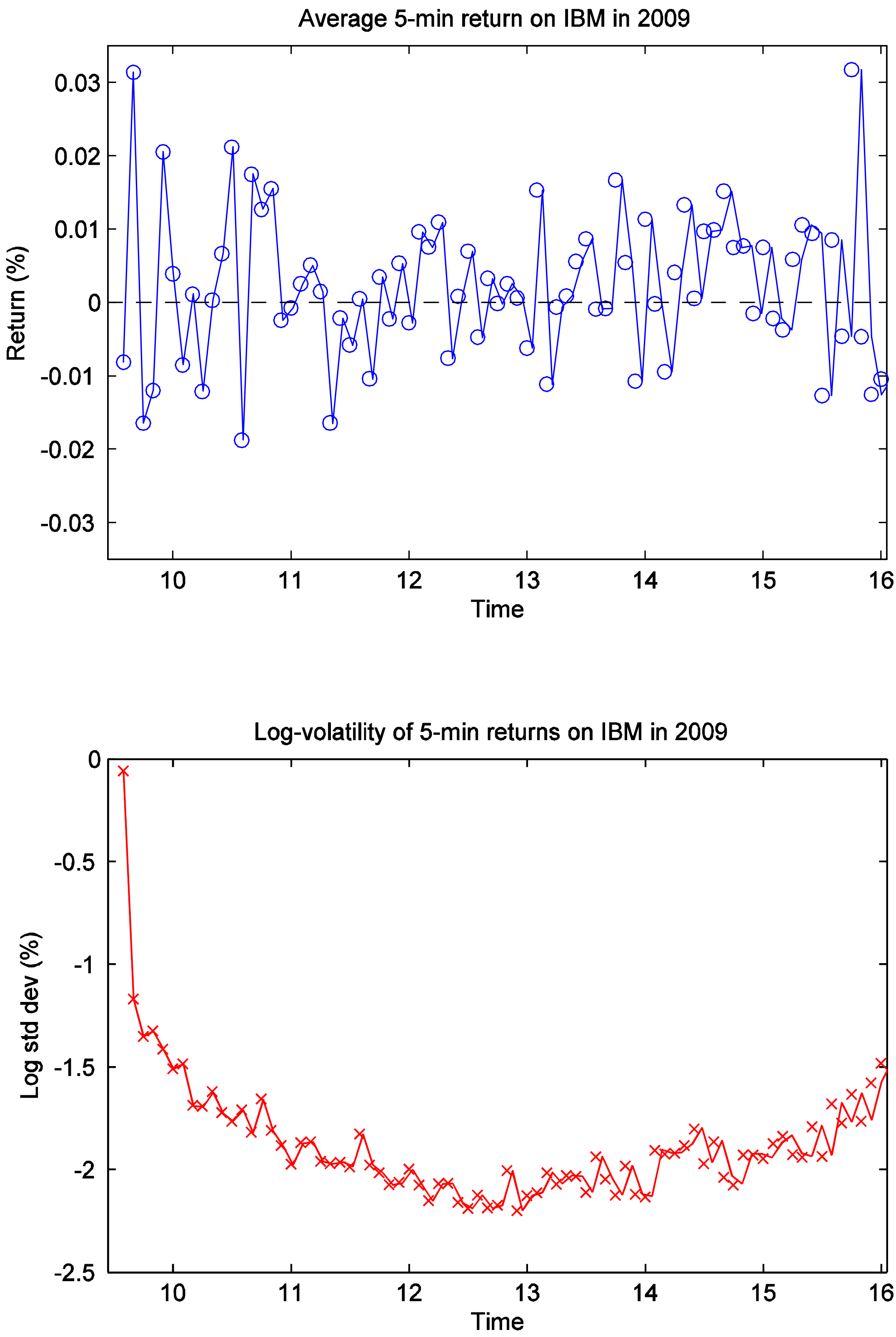


Figure 15.1: Diurnality in 5-minute returns on IBM (upper panel), and in log standard deviation of 5-minute retuns (lower panel), Jan 2009-Dec 2009. (The lower and upper bounds in the lower panel correspond to a range of 11.5% to 140% annualized volatility.

(say 9:30, 14:23, etc.). Then we could define diurnal dummy variables as:

$$
\begin{aligned}
D_{1,t} &= \begin{cases} 1 & TIME_t \leq 10:00 \\ 0 & else \end{cases} \\
D_{2,t} &= \begin{cases} 1 & 10:00 < TIME_t \leq 10:30 \\ 0 & else \end{cases} \\
&\vdots \\
D_{13,t} &= \begin{cases} 1 & 15:30 < TIME_t \leq 16:00 \\ 0 & else \end{cases}
\end{aligned}
$$

Let's firstly consider a model for diurnality in returns. The pure diurnal dummy model for returns is:

$$r_t = \beta_1 + \sum_{i=2}^{13} \beta_i D_{i,t} + e_t$$

This model will capture any pattern in returns over the trade day. Note that since we have included a constant, we only include dummy variables for *all but one* possible times of the day. (It does not matter which dummy is omitted). Including both an intercept and all dummy variables would lead to perfect multicollinearity of the regressors and the regression would fail. Note that we expect intra-daily returns to be heteroskedastic, and so it is important to use 'robust' standard errors, such as Newey-West standard errors.

There are other types of 'calendar effects' that can be captured with dummy variables. For example, we might think that the first trading day following a holiday weekend will be more volatile than other days, and so we could include a dummy to capture this effect. We can also use them to capture day-of-the-week (or 'hebdomadal') effects; for example, there has been some evidence reported that returns are more volatile on Mondays than on other days of the week.

We test for the importance of diurnality by running the regression and then testing whether all coefficients are equal. In the above example we would test:

$$
\begin{aligned}
H_0 &: \beta_2 = ... = \beta_{13} = 0 \\
H_a &: \beta_i \neq 0 \text{ for } j = 2, ...13
\end{aligned}
$$

This test can be simply conducted using a χ^2_{12} test. Using this model on the IBM data set introduced above we I obtained χ^2_{12} test statistic of 7.37, and a corresponding p-value of 0.83, thus we fail to reject the null, and conclude that we have *no* evidence of a diurnal pattern in these returns.

To test for diurnality in volatility we first control for the diurnality in the mean (this is similar to the need to estimate a model for the mean before considering a GARCH model for conditional variance), and then estimate:[2]

$$\log e_t^2 = \gamma_1 + \sum_{i=2}^{13} \gamma_i D_{i,t} + u_t$$

[2] It is common to refer to a variable with the diurnal component removed as being "de-seasonalised" rather than the uglier phrase "de-diurnalised". We will follow this convention here.

and implement a χ^2_{12} test as for returns. The test statistic and p-value in this case were 460.15 and 0.00, and thus we strongly reject the null (the 95% critical value for a χ^2_{12} distribution is 21.03) and conclude that there is significant evidence of diurnality in the volatility of these returns. I ran similar regressions for log-volume[3] and the bid-ask spread and found the following results:

Dummy models for diurnality

	Mean	*Log std dev*	*Log volume*	*Spread*
χ^2_{12} stat	7.3743	460.15	1203.3	1179.0
χ^2_{12} p-val	0.8319	0.0000	0.0000	0.0000

Activity 15.1 The following tables and figures are based on monthly returns on a U.S. stock market index July 1926 to March 2014 (a total of 1053 observations). See also Figure 15.2 below.

1. The table below presents parameter estimates, t-statistics, and p-values from a few different hypothesis tests based on the following regression:

$$r_t = \sum\nolimits_{s=1}^{4} \beta_s \mathbf{1}\left\{season_t = s\right\} + e_t$$

$$\text{where } season_t = \begin{cases} 1, & \text{if month } t \text{ is March, April or May} \\ 2, & \text{if month } t \text{ is June, July or August} \\ 3, & \text{if month } t \text{ is September, October or November} \\ 4, & \text{if month } t \text{ is December, January or February} \end{cases}$$

	β_1	β_2	β_3	β_4	
Estimate	0.573	0.908	0.083	1.031	
t-statistic	1.375	2.388	0.208	3.837	
p-value on test that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$					0.000
p-value on test that $\beta_1 = \beta_2 = \beta_3 = \beta_4$					0.226
p-value on test that $\beta_1 = \beta_2 = \beta_3 = 0$					0.028

What, if anything, does the above table tell us about average returns on the U.S. stock index?

2. The residuals from the regression in part (1) were used in the following regression

$$e_t^2 = \sum\nolimits_{s=1}^{4} \gamma_s \mathbf{1}\left\{season_t = s\right\} + u_t$$

	γ_1	γ_2	γ_3	γ_4	
Estimate	32.293	32.412	34.143	17.558	
t-statistic	3.954	3.676	6.432	9.274	
p-value on test that $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0$					0.000
p-value on test that $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4$					0.002
p-value on test that $\gamma_1 = \gamma_2 = \gamma_3 = 0$					0.000

[3] Note that when the volume in a given 5-minute interval is exactly zero we cannot take the logarithm. The smallest non-zero volume for this stock is 100. To overcome the "log of zero" problem I added 1 to each volume.

What, if anything, does the above table tell us about the volatility of returns on the U.S. stock index?

15.4 Polynomial trend models

In the above method we used 13 dummy variables to capture any patterns in intra-daily returns. This is a flexible method, but involves estimating many parameters. In applications involving high frequency data we usually do not need to worry about estimating many parameters. In some other applications, however, we may want to try to capture seasonal patterns in a more parsimonious manner. One way of doing so is through some specification of a trend. For example, we could specify a linear trend:

$$r_t = \beta_0 + \beta_1 HRS_t + e_t$$

where HRS_t counts the number of hours between midnight and the time of the return (e.g., 9.5 for a 9:30 return, 9.5547 for a 9:33:17 return, etc.). This model allows expected returns to be increasing or decreasing (or constant) in a linear fashion throughout the trade day. We could alternatively use a quadratic trend model:

$$r_t = \beta_0 + \beta_1 HRS_t + \beta_2 HRS_t^2 + e_t$$

which allows expected returns to be constant, linear increasing/decreasing, or to follow a U or inverted U shape throughout the trade day. Notice that it allows for a smaller set of possible shapes than the diurnal dummy model, but has the benefit of only requiring 3 parameters to be estimated, rather than 13. The diurnal pattern in bid-ask spreads suggested a cubic polynomial, and so I present results for that model only for the spread. We obtain a test for diurnality from a polynomial trend model by testing (jointly) that all parameters other than the constant are equal to zero.

Figures 15.3 and 15.4 present the diurnal patterns implied by each of these models, and the estimation results are presented below.

Trend models for diurnality

	Mean	*Log std dev*	*Log volume*	*Spread*
Constant	-0.0073	20.3824^*	33.827	1.4392^*
	(0.1133)	(1.4832)	(0.85507)	(0.0822)
HRS	0.0012	-3.9833^*	-3.8267^*	-0.3183^*
	(0.0172)	(0.2370)	(0.1387)	(0.0194)
HRS^2 × 10^{-4}	-0.3571	1497.7^*	1487.6^*	239.04^*
	(6.4459)	(93.006)	(55.269)	(15.078)
HRS^3 × 10^{-4}	–	–	–	-5.9507^*
				(0.3864)
χ^2 stat	0.1242	399.41	834.92	868.11
χ^2 p-val	0.9398	0.0000	0.0000	0.0000

The test results from the polynomial trend models are consistent with those from the dummy variable models: we find no evidence of diurnality in average returns, but strong evidence in volatility, volumes and spreads.

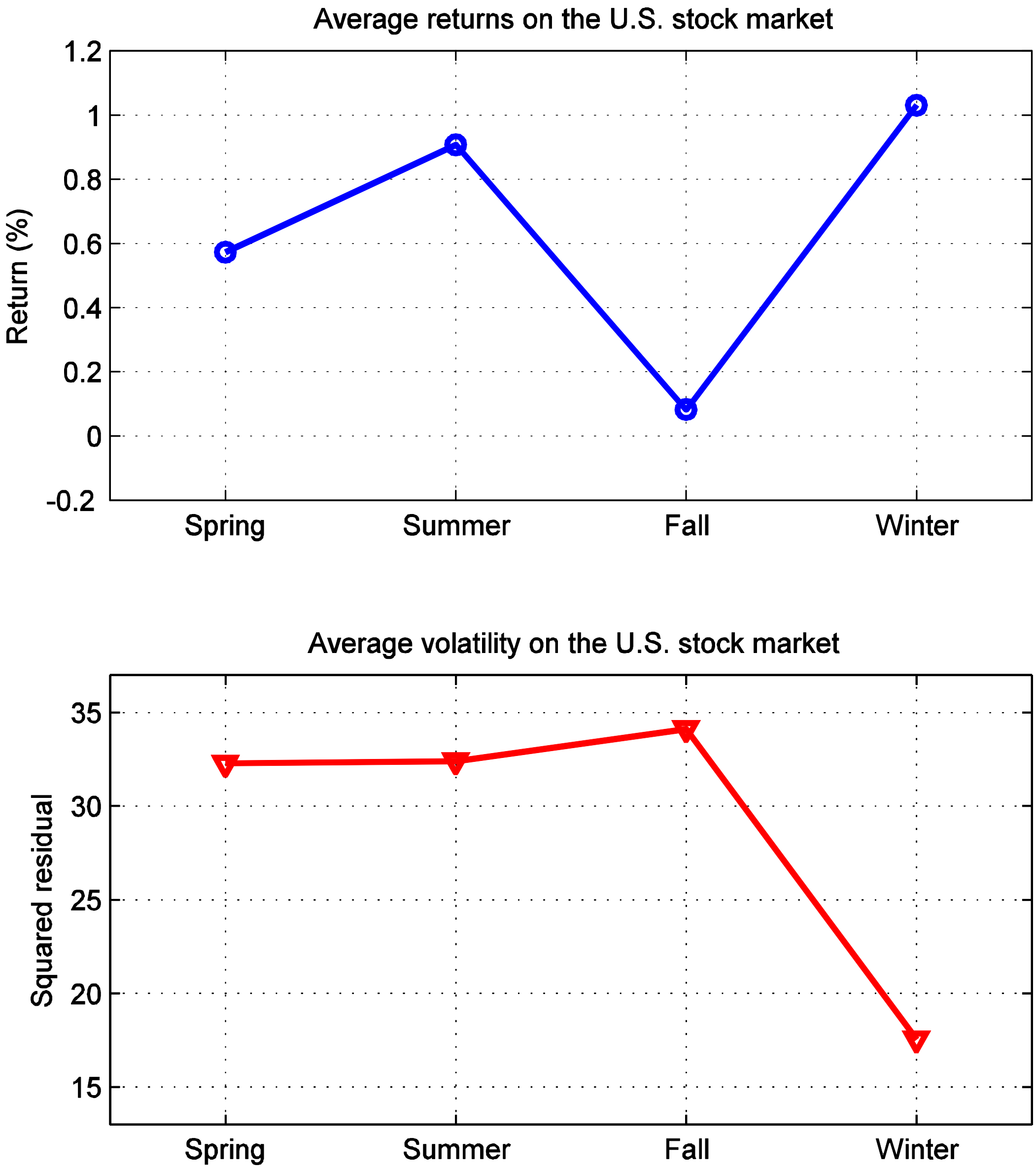


Figure 15.2: Figure for Activity 15.1.

> **Activity 15.2** Derive analytically the time of day that yields the lowest volatility, according to the quadratic trend model.

15.5 Modelling conditional volatility in the presence of diurnality

Standard conditional volatility models assume that all of the time variation in the conditional variance is non-deterministic. The presence of diurnality suggests we need to decompose the variation in volatility into deterministic and stochastic components. Let:

$$\begin{aligned} r_t &= \mu_t + e_t \\ e_t &= s_t e_t^* \\ e_t^* &= \sigma_t Z_t, \quad Z_t \sim iid \;\; (0,1) \\ \text{where} \quad \log s_t^2 &= \beta_0 + \beta_1 HRS_t + \beta_2 HRS_t^2 \\ \sigma_t^2 &= \omega + \beta\sigma_{t-1}^2 + \alpha\sigma_{t-1}^2 Z_{t-1}^2 \end{aligned}$$

That is, returns possibly have a diurnal conditional mean, and diurnality in log volatility described using a diurnal dummy model, GARCH in the de-seasonalised residuals, and an innovation term, Z_t, which is *iid* with mean zero and variance one. The de-seasonalised residuals can be extracted from the observed returns as:

$$e_t^* = \frac{r_t - \mu_t}{s_t}$$

In practice, we need to estimate μ_t and s_t to do this transformation, and the s_t component is a little more complicated than usual as we estimate it in logs.

$$\begin{aligned} e_t &= r_t - \mu_t \\ &= \exp\{\log s_t\} e_t^* \\ \text{so} \quad \log e_t^2 &= \log s_t^2 + u_t, \quad \text{where } u_t = \log\left(e_t^{*2}\right) \\ &= \widehat{\log s_t^2} + \hat{u}_t, \text{ from the diurnal dummy regression} \\ e_t^2 &= \exp\left\{\widehat{\log s_t^2}\right\} \exp\{\hat{u}_t\} \\ \text{so} \;\; E\left[e_t^2|s_t\right] &= \exp\left\{\widehat{\log s_t^2}\right\} E\left[\exp\{\hat{u}_t\}|s_t\right] \\ &= \exp\left\{\widehat{\log s_t^2}\right\} E\left[\exp\{\hat{u}_t\}\right], \text{ if } u_t \text{ is independent of } s_t \\ &\approx \exp\left\{\widehat{\log s_t^2}\right\} \frac{1}{T}\sum_{j=1}^{T} \exp\{\hat{u}_j\} \\ &\equiv \hat{s}_t^2 \\ \text{and so} \;\; \hat{e}_t^* &= \frac{r_t - \hat{\mu}_t}{\hat{s}_t} \end{aligned}$$

We need to deal with the additional term, $\frac{1}{T}\sum_{j=1}^{T}\exp\{\hat{u}_j\}$, because we are interested in $E\left[e_t^2|s_t\right]$ rather than $E\left[\log e_t^2|s_t\right]$. This non-linear transformation of the dependent

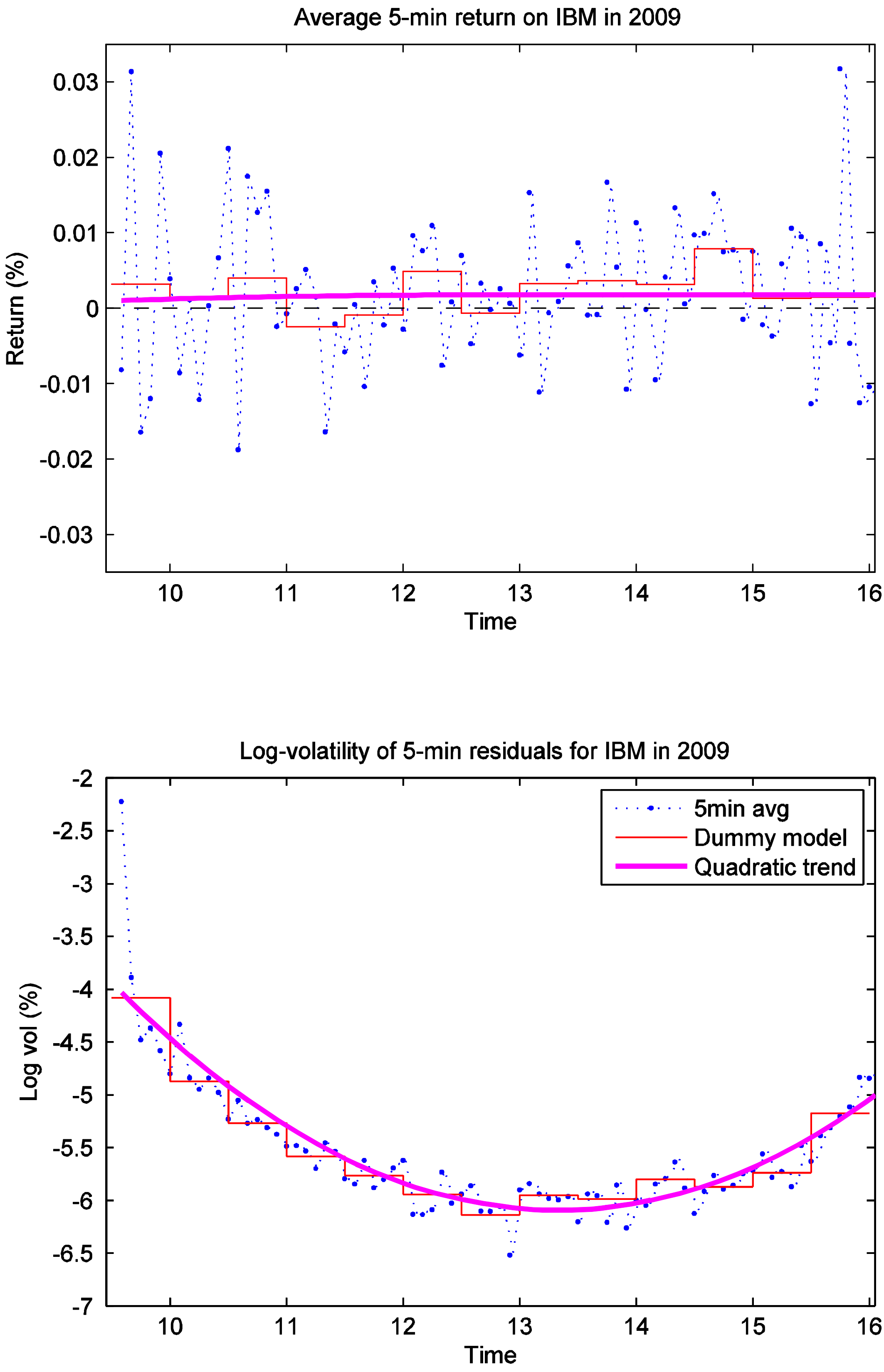


Figure 15.3: Diurnality in 5-minute returns on IBM (upper panel), and in log standard deviation of 5-minute retuns (lower panel), Jan 2009-Dec 2009, with estimated values from a dummy variable model and a quadratic trend model.

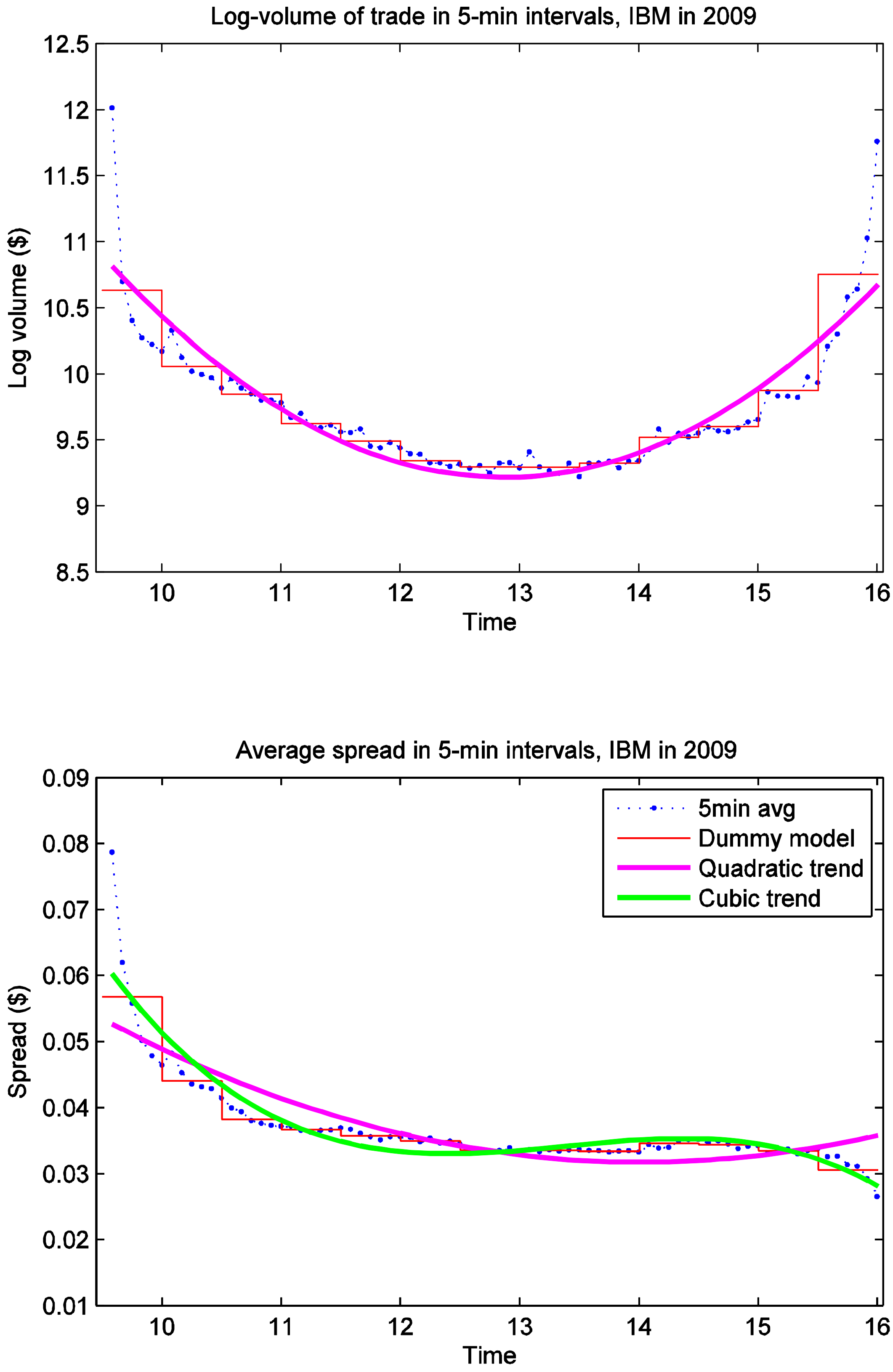


Figure 15.4: Diurnality in 5-minute log-volume (upper panel), and bid-ask spreads (lower panel), for IBM, Jan 2009-Dec 2009, with estimated values from a dummy variable model and a polynomial trend model.

variable in the regression leads to the need for the correction term $\frac{1}{T}\sum_{j=1}^{T}\exp\{\hat{u}_j\}$. Failing to include this in the standardisation will mean that the standardised returns will be mean zero, with no diurnal patterns in either mean or variance, but will *not* have variance one. In some cases this will not matter, but in other cases it will.

We estimate the diurnal models in the first stage, de-seasonalise the returns, and then estimate a GARCH model on the de-seasonalised returns. To compare the GARCH parameter estimates obtained with and without de-seasonalising the returns, I present both below:

	Raw returns	De-seasonalised returns
ω	0.0013	0.0019
β	0.7383	0.9594
α	0.2617	0.0339

The constant terms in the GARCH models are not directly comparable, as the de-seasonalised returns have a different unconditional variance to the raw returns. The parameters defining the dynamics (α and β) are directly comparable, and we see that the GARCH model on raw returns places more weight on the most recent innovation (α is higher) than the GARCH model on de-seasonalised returns.

The variance forecasts obtained from the model estimated on de-seasonalised returns need to be scaled up by $\exp\left\{\widehat{\log s_t^2}\right\}\frac{1}{T}\sum_{j=1}^{T}\exp\{\hat{u}_j\}$, to reflect both the GARCH and the diurnal components of the model:

$$\begin{aligned} V_t\left[e_{t+1}^*\right] &= \sigma_{t+1}^2 \\ V_t\left[e_{t+1}\right] &= s_{t+1}^2\sigma_{t+1}^2 \\ &\approx \exp\left\{\widehat{\log s_{t+1}^2}\right\}\left(\frac{1}{T}\sum_{j=1}^{T}\exp\{\hat{u}_j\}\right)\hat{\sigma}_{t+1}^2 \end{aligned}$$

Figure 15.5 presents the two components of this volatility model, along with the combined conditional variance estimate, for the first 5 days of the sample. The diurnal component in the top panel reveals the smooth time-of-day effect in volatility, and of course this pattern repeats identically from day to day. The middle panel shows the GARCH variations in the volatility of the de-seasonalised data. It is centered on one, as it scales the diurnal component up or down depending on the history of returns. The lower panel shows the combined model, which is dominated by the diurnal component, but with non-trivial variations due to the GARCH component.

Figure 15.6 compares the volatility forecasts from a GARCH model with and without a diurnal component, again for the first five days of the sample. We see that the model with a seasonality component is able to capture the volatility spike we see in the first half-hour of trading, through the use of a dummy variable. The GARCH model on raw returns is not able to do this explicitly, and so the model attempts to replicate this pattern by adjusting the α and β parameters of the GARCH model. Capturing the diurnal pattern explicitly is clearly preferable to ignoring it.

We can test formally whether the volatility model with a seasonal component is better than one that ignores it using Mincer-Zarnowtiz regressions and a Diebold-Mariano test.

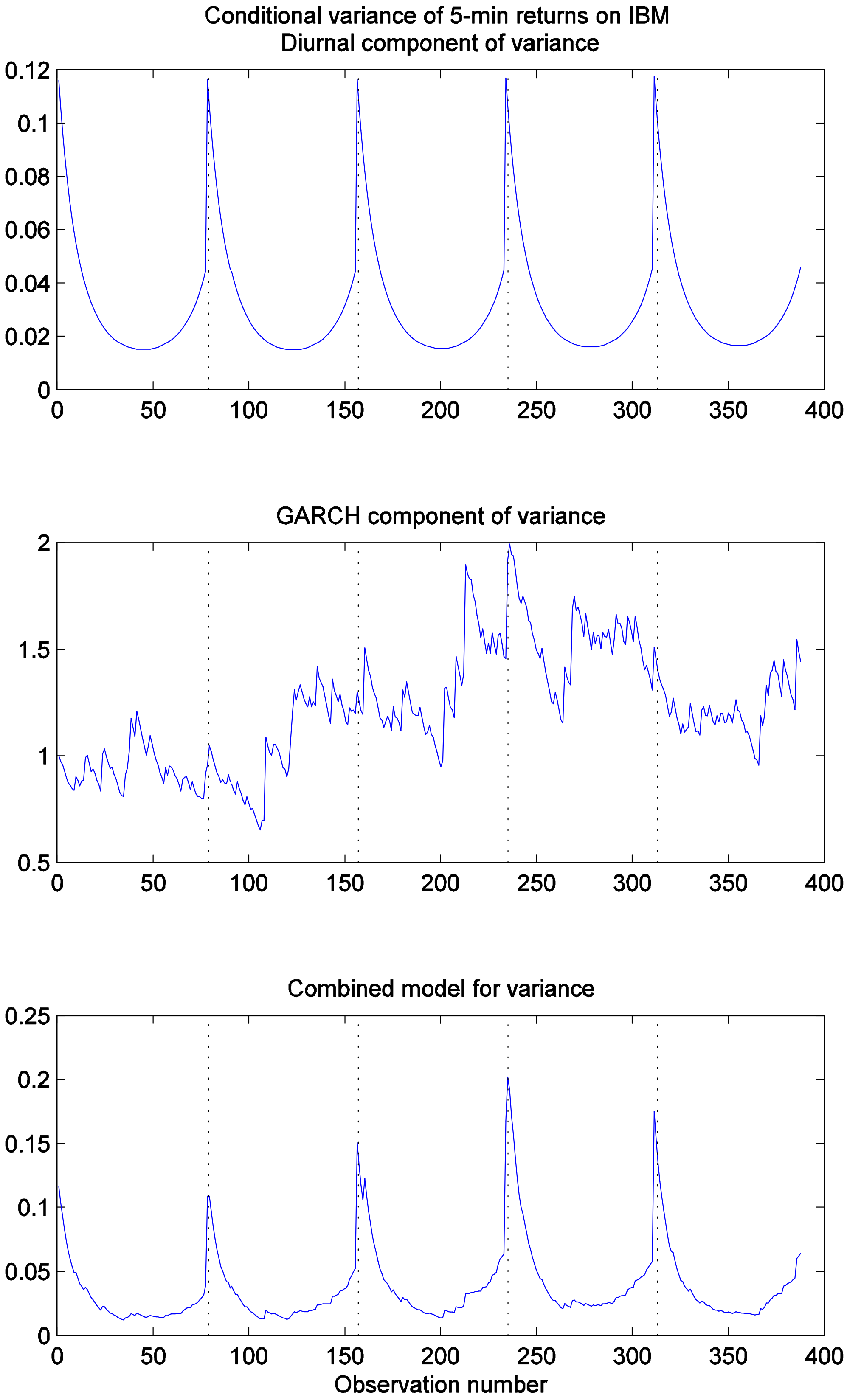


Figure 15.5: The components of the volatility model for IBM returns from a GARCH model with diurnality.

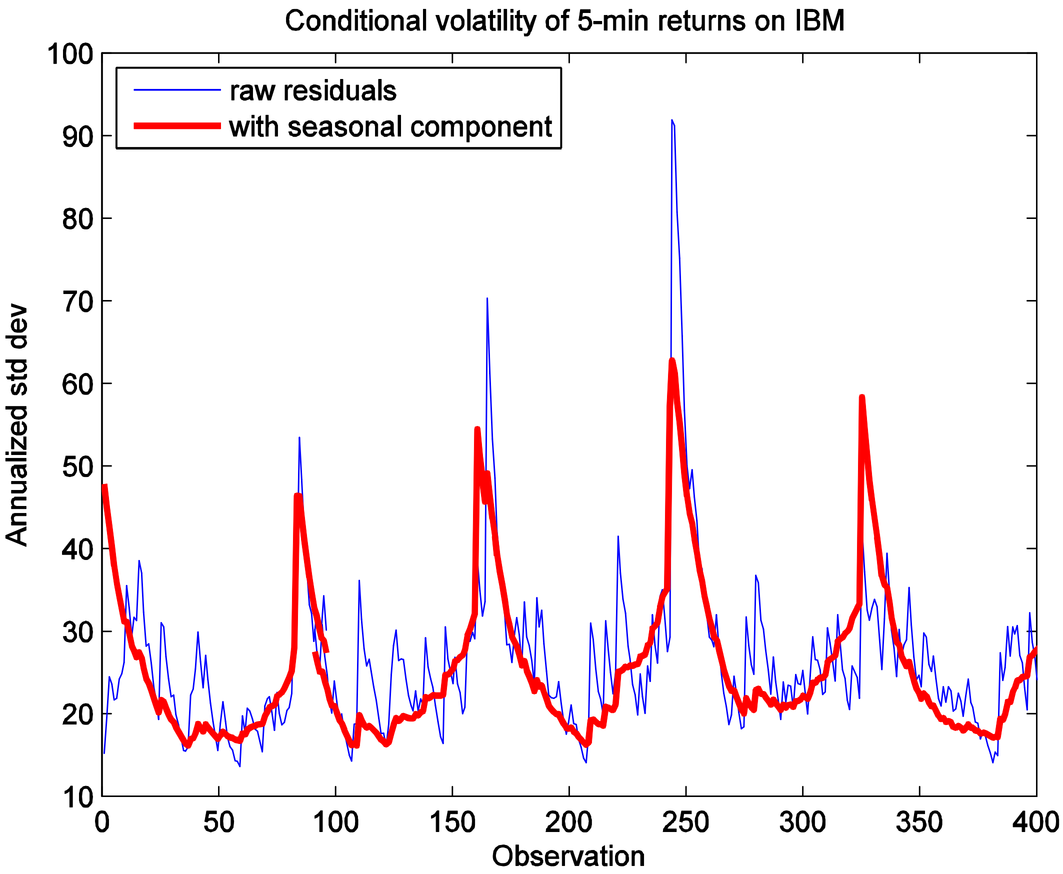


Figure 15.6: Conditional volatility of IBM returns from GARCH models with and without volatility diurnality models.

I used the squared residuals (*not* de-seasonalised) as the volatility proxy and estimated:

$$\begin{aligned} e_t^2 &= \beta_0 + \beta_1 h_t^{(j)} + u_t \\ H_0 &: \beta_0 = 0 \cap \beta_1 = 1 \\ \text{vs.} \quad H_a &: \beta_0 \neq 0 \cup \beta_1 \neq 1 \end{aligned}$$

where $h_t^{(j)}$ is the volatility forecast from the GARCH model with or without a seasonal component. I also considered Diebold-Mariano tests using the MSE and QLIKE loss functions:

$$\begin{aligned} d_t^{MSE} &\equiv \left(e_t^2 - h_t^{(raw)}\right)^2 - \left(e_t^2 - h_t^{(seasonal)}\right)^2 \\ d_t^{QLIKE} &\equiv \frac{e_t^2}{h_t^{(raw)}} + \log h_t^{(raw)} - \frac{e_t^2}{h_t^{(seasonal)}} - \log h_t^{(seasonal)} \\ H_0 &: E\left[d_t^{(L)}\right] = 0 \\ \text{vs.} \quad H_1 &: E\left[d_t^{(L)}\right] > 0 \\ \text{or} \quad H_2 &: E\left[d_t^{(L)}\right] < 0 \end{aligned}$$

where $L = MSE$ or $L = QLIKE$. The results clearly indicate that the model with a seasonal component is preferred: the MZ test rejects the model with no such component, but does not reject the model with a seasonal component, and the DM tests favor the model with a seasonal component.

	Raw resids	With seasonal
$\hat{\beta}_0$ $(s.e.)$	0.0325* (0.0042)	0.0042 (0.0064)
$\hat{\beta}_1$ $(s.e.)$	0.0930* (0.0238)	0.9620 (0.0246)
MZ p-val	0.0000	0.9878
DM^{MSE} t-stat	1.26	
DM^{QLIKE} t-stat	30.62*	

15.6 Overview of chapter

This chapter introduced two models for capturing time-of-day features (known as 'diurnal' features) in asset returns. The first model is based on dummy variables for different periods of time (eg, each hour of the day) and the second model is based on a polynomial trend model. We also discussed how to use these models to formally test for the presence of a diurnal pattern.

15.7 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Describe two methods for capturing seasonality or diurnality in returns and volatility
- Interpret empirical tests for the presence of seasonality and diurnality

15.8 Test your knowledge and understanding

1. You are studying the returns on a particular stock, and would like to determine whether there is a day-of-the-week pattern (known as a 'hebdomadal' pattern) in these returns. Describe two models for testing whether such a pattern is present.

2. What are the pros and cons of the 'diurnal dummy' model versus the 'quadratic trend' model?

3. A researcher estimates a quadratic trend model for five-minute returns on IBM:

$$r_t = \beta_0 + \beta_1 HRS_t + \beta_2 HRS_t^2 + e_t$$

where HRS_t is the number of hours between midnight and the time of the trade at time t. Interpret the following table of results in terms of the evidence for/against diurnality in these returns.

Constant (standard error)	0.1767 (0.0469)
HRS (standard error)	0.0242 (0.0071)
$HRS^2 \times 10^{-4}$ (standard error)	8.4096 (2.7455)
p-value on test that $\beta_0 = \beta_1 = \beta_2$	0.000
p-value on test that $\beta_0 = \beta_1 = \beta_2 = 0$	0.000
p-value on test that $\beta_1 = \beta_2 = 0$	0.000

4. Don't forget to check the VLE for additional practice problems for this chapter.

15.9 Solutions to activities

Activity 15.1

This question relates to testing for seasonality in stock returns. Part (1) presents a dummy variable model to capture predictability in returns as a function of the season. Since a dummy is included for all possible seasons, the null of *no* seasonality is given by:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \beta_3 = \beta_4 \\ \text{vs. } H_a &: \beta_i \neq \beta_j \text{ for some } i, j \end{aligned}$$

That is, we want to test that the average return in all seasons is the same. This corresponds to the second p-value in the table, which is equal to 0.226. This is greater than 0.05, and so we fail to reject the null that there is no seasonality in average stock returns. The first p-value in the table corresponds to a test that average returns in all seasons are equal to zero. This p-value is 0.000, indicating a strong rejection of that hypothesis, i.e., average returns are significantly different from zero. The third p-value in the table is not really interesting: it tests whether average returns are zero in spring, summer and fall, but leaves average returns in winter untested.

The second table presents the corresponding results for a model of seasonality in volatility, using the squared residual as a volatility proxy in the regression. The null of no seasonality again appears in the second p-value, and we observe that it is 0.002. This is smaller than 0.05, and so we reject the null that volatility is the same across all four seasons. That is, we have significant evidence of seasonality in volatility. (From Figure 15.2 it appears this result might be driven by average volatility in winter, which appears much lower than the other seasons.) The other two p-values in this table are not really interesting: the first tests whether all coefficients are zero, but since the dependent variable in this regression is a squared residual, which is always positive, we know that this cannot be true, and so we are not surprised to see a small p-value (0.000) indicating a rejection of this null. A similar comment applies to the third p-value in this table.

Activity 15.2

Recall that the extremum of the parabola $y = ax^2 + bx + c$ occurs at $x = -b/(2a)$. Since the estimated coefficient on HRS^2 is positive for volatility, we know that the extremum will be a minimum (i.e., the parabola opens ‘upwards’), and it will occur at $3.9833/\left(2 \times 1497.7 \times 10^{-4}\right) = 13.30$, corresponding to 1:18pm. The minimum for the volume model is 12.86, which is 12:52pm.

Chapter 16

Modelling high frequency financial data: Irregularly-spaced data

16.1 Introduction

A second difference between high frequency data and daily or lower frequency data arises when deciding how to treat the raw 'tick' data, i.e., the data for each and every trade or quote. In some analyses, the problem to be studied can be addressed by aggregating the 'tick' data up to a certain frequency, say 5 or 10 minutes, and then analyse the sequence of 5-minute returns. Doing so makes the data evenly spaced, and thus more similar to well-studied, low frequency data. Many questions, however, are best addressed using tick data, meaning that we must find ways of dealing with the *irregularly-spaced observations*. This chapter will examine methods to handle this type of data.

16.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce you to irregularly-spaced time series, in contrast with more common regularly-spaced time series
- Introduce the notion of a trade 'duration' as a way to model irregular spacings
- Discuss the ACD model for durations, and outline its similarity to the GARCH model for volatility

16.1.2 Learning outcomes

By the end of this chapter, and having completed the activities, you should be able to:

- Define a trade 'duration,' and discuss how this variable changes, on average, through the trade day
- Describe the 'autoregressive conditional duration' (ACD) model for trade durations

16.1.3 Essential reading

This chapter serves as the essential reading for this topic.

16.1.4 Further reading

- Taylor, Stephen J. *Asset Price Dynamics, Volatility and Prediction.* (Princeton University Press, Oxford, 2005) [ISBN 9780691134796]. Chapter 12
- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354]. Chapter 5

16.1.5 References cited

- Engle, R. F. and J. R. Russell, ‘Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data,’ *Econometrica*, 1998, 66(5), pp.1127–1162.

16.2 The market microstructure of IBM stock prices

To illustrate some of the issues that arise in the study of high frequency prices, let us consider the price of IBM on the New York Stock Exchange (NYSE) on a single day, 31 December 2009. The NYSE is open from 9:30am to 4pm, and the TAQ database contains transaction and quote prices for all securities listed on the New York Stock Exchange, American Stock Exchange, Nasdaq National Market System, stamped with a time measured in seconds (e.g., 9:31:28). Thus we have a possible total of 23,401 observations per trade day, and on this day there were 2,616 trades, corresponding to one trade every 9 seconds. One-second trade prices for this day are presented in the upper panel of Figure 16.1, and bid-ask spreads are presented in the lower panel.

We see that the price of IBM fell slowly from around \$132.50 at the start of the trade day to around \$131 at 11am, where it stayed until about 2:30pm. The afternoon saw a small rally, with the price rising to around \$132 at 3pm, before closing at \$130.84. Perhaps more interesting is the variation in the bid-ask spread over this day: the spread started at 19 cents at 9:30am, before falling to around 2-5 cents at around 10am. It remained roughly in this range for the remainder of the day, falling slightly towards the close. This pattern is consistent with the diurnal pattern in bid-ask spreads that we modelled in the previous chapter.

Looking at trade prices for IBM over a whole day is still too low a frequency to see some of the market microstructure features of this security. In Figure 16.2 I present trade and quote prices (the ask price is the upper solid line and the bid price is the lower solid line) around the opening and closing few minutes of the trade day on December 31, 2009.

The upper panel reveals that there were relatively few trades during the first few minutes of the trade day: just 52 trades in the first 6 minutes, compared with 103 changes in one of the quote prices. Indeed, in the opening few minutes the quote prices adjust very frequently even in the absence of trades. In contrast with later in the day, when quote prices adapt to information revealed by trades, this presumably reflects market participants’ adjusting their bid and ask prices in response to others’ adjustments of bid/ask prices, both for this security and for other securities (e.g., stocks in a similar industry, changes in interest rates, etc.). The lower panel, covering the last 0.01 hours (36 seconds) of the trade day shows the opposite pattern: quote prices adjusting less frequently than trades occurring. Notice that several of the trade prices in

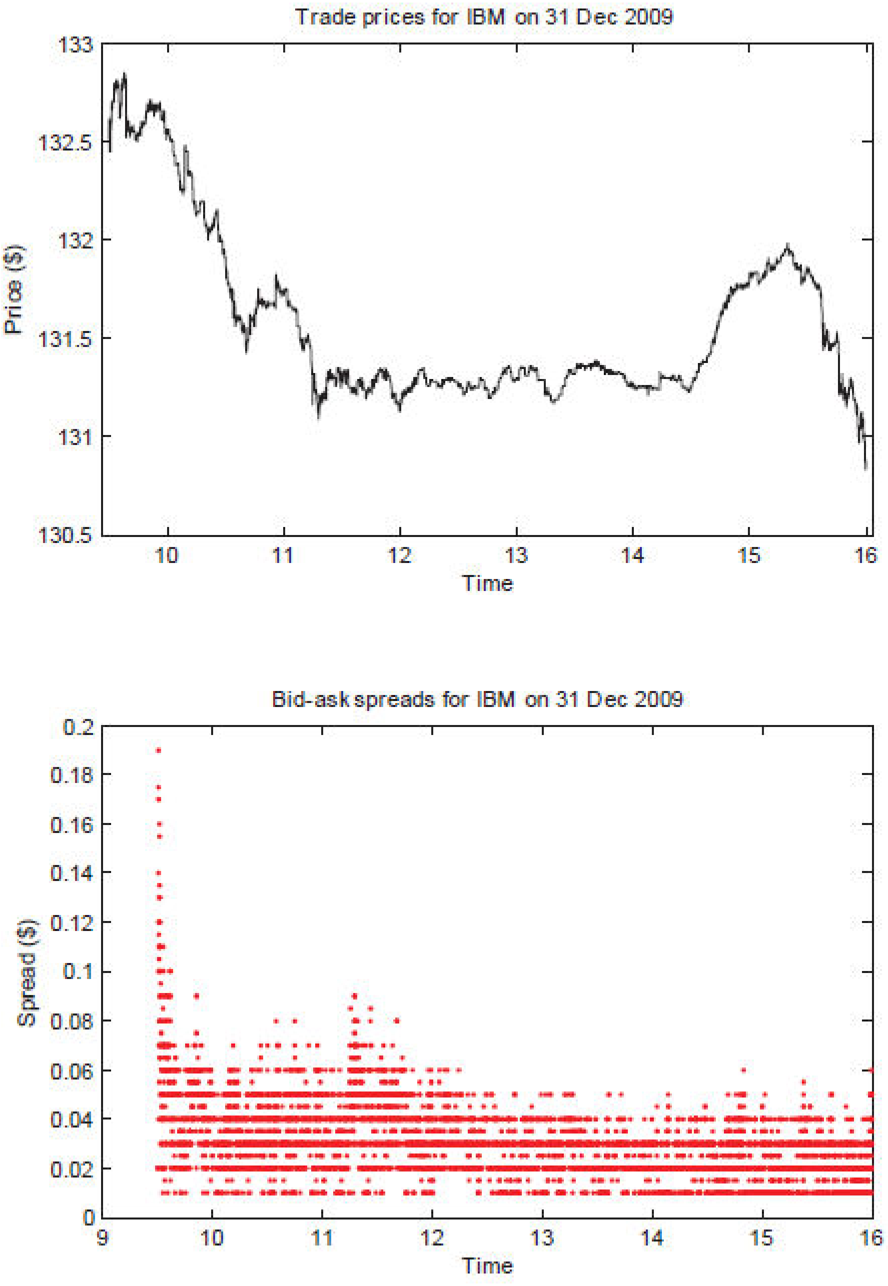


Figure 16.1: Trade prices and bid-ask spreads for IBM, 31 Dec 2009.

this interval appear inside the quote prices. This is an artifact of having several trades occurring with the same time stamp, i.e., several trades occurring in the same second. I used the standard rule of recording the average of all trade prices with the same time stamp, and so seeing a trade price just slightly above the bid price means that at least two prices were recorded in that second, and the average of those was close to the bid price. Both of these panels reveal the fact that trades and quotes arrive with *irregular* spacings in time.

16.3 Modelling durations

Although it is sometimes convenient to analyze returns that have been aggregated up to a certain frequency, e.g. 5-minute returns, market participants face data that does *not* arrive at neatly spaced 5-minute intervals. Real market participants face data that arrives at random times, sometimes close together and sometimes far apart. In standard forecasting situations we know the point on the x-axis (time) where the next observation will lie but not the point on the y-axis. In the forecasting of microstructure data we do not know the x- *or* the y-axis coordinates.

In this section we will focus on forecasting the arrival time of the next observation. We will do so by employing a model for the *durations* between trades. A 'trade duration' is simply the time between two trades, usually measured in seconds. Durations are economically interesting because they tell us something about liquidity: periods of intense trading are generally periods of greater market liquidity than periods of sparse trading, though this is not always the case. Furthermore, durations relate directly to news arrivals and the adjustment of prices to news, and so have some use in discussions of market efficiency.

Let $\{r_{t_1}, r_{t_2}, ..., r_{t_k}, ...\}$ be a sequence of high frequency returns, and $\{TIME_1, TIME_2, ..., TIME_k, ...\}$ be the sequence of associated trade times. Define the duration variable as

$$x_j = TIME_j - TIME_{j-1}$$

We will use high frequency transaction data on IBM for the calendar year 2009. Over this period we have 1,254,972 observations with a unique time stamp, implying an average of 4,980 trades per day, or 1 trade every 4.7 seconds that the market is open. (Note that the data we are using in this chapter are only stamped to the nearest second, and there are many times when we observe more than one trade within a second. Thus this figure is a lower bound on the actual number of trades in 2009.) Figure 16.4 plots the first few transactions on 31 December 2009, and 16.5 plots the first 160 trade durations of 2009.

16.3.1 Diurnality in durations

Some authors have found evidence of a diurnal effect in durations. Let us now check whether such a pattern is also present in our data, by using the diurnal dummy model or the quadratic trend model introduced in the previous section:

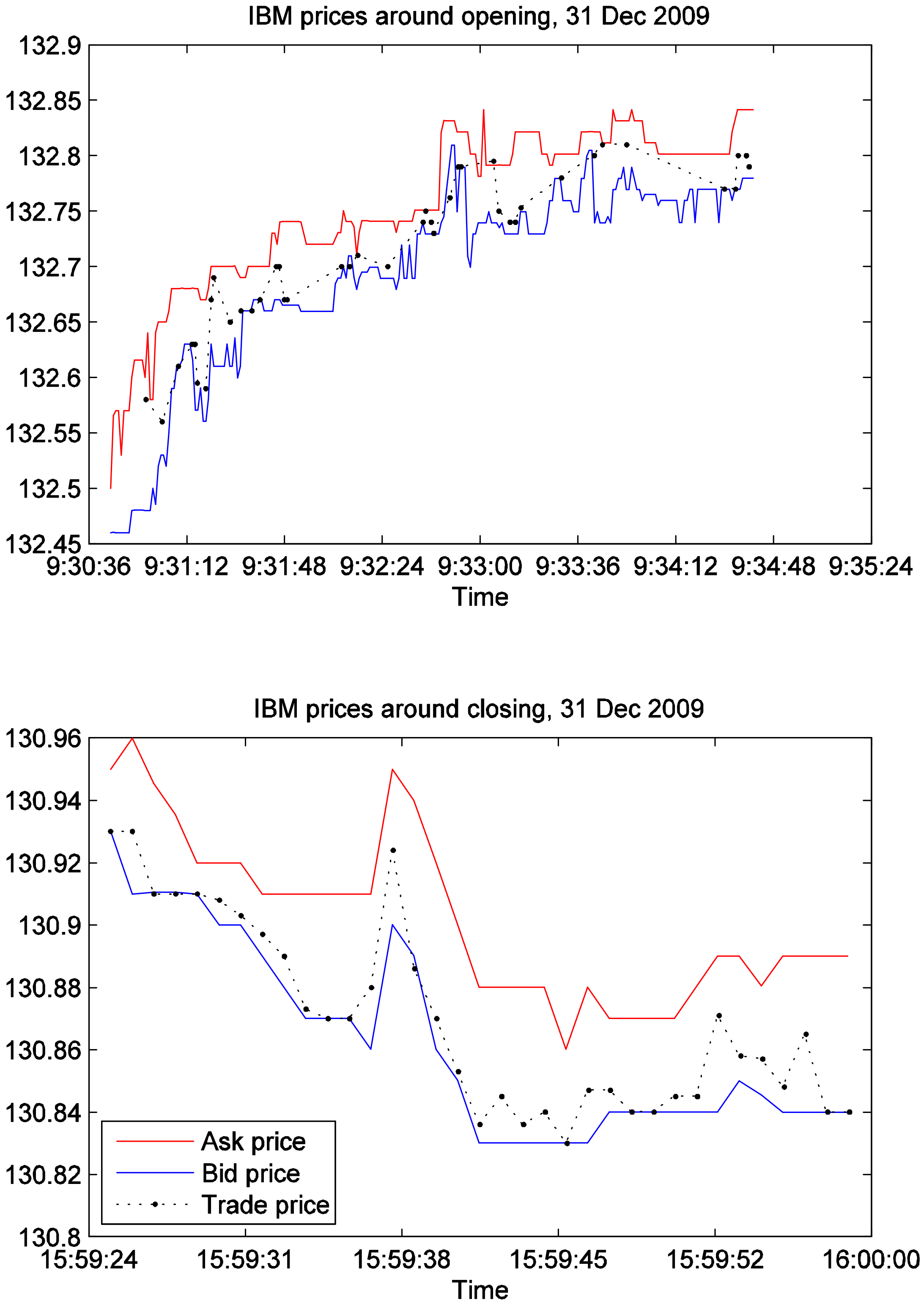


Figure 16.2: Trade and quote prices and bid-ask spreads for IBM, 31 Dec 2009. Top panel presents prices between 9:30am and 9:36am, lower panel between 3:59pm and 4pm.

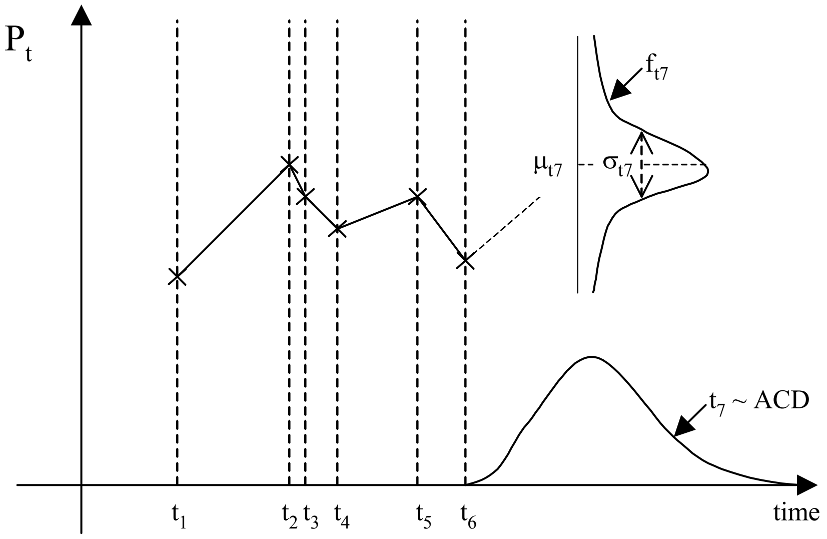


Figure 16.3: Forecasting both the return and the time of the next trade.

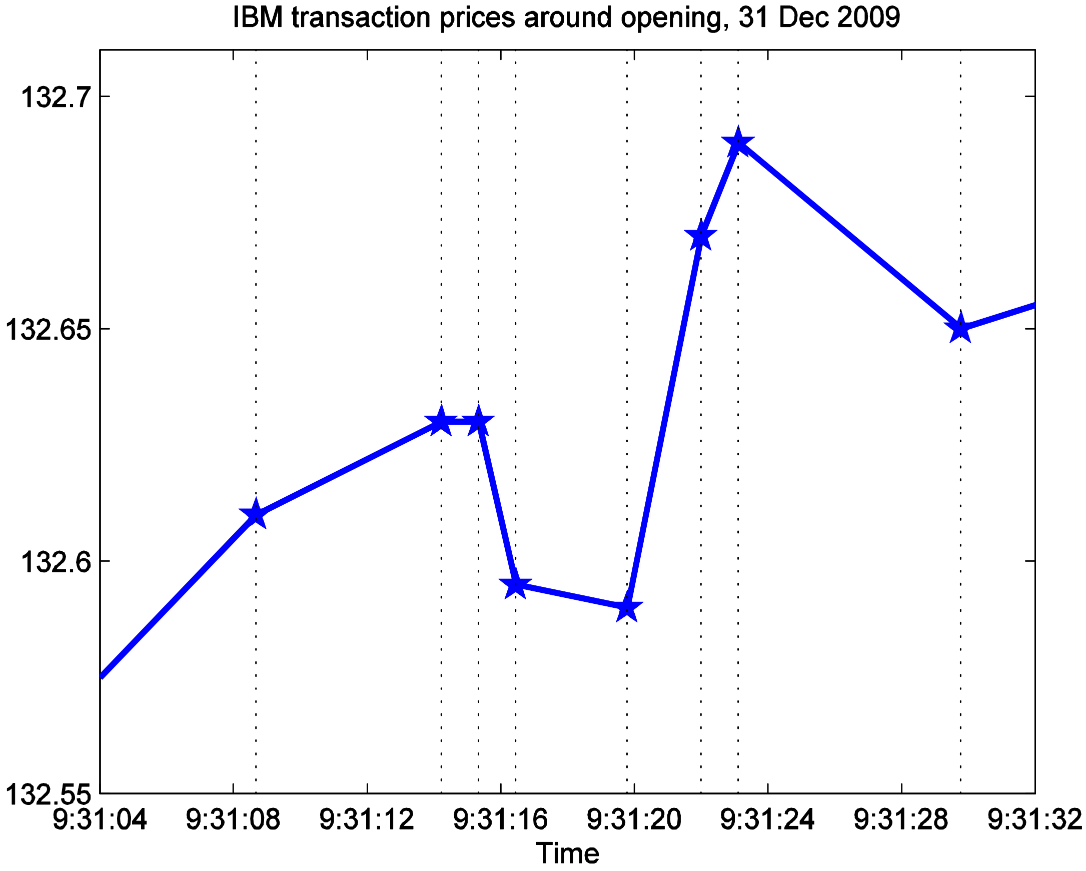


Figure 16.4: Transaction prices on IBM in the morning of 31 December 2009.

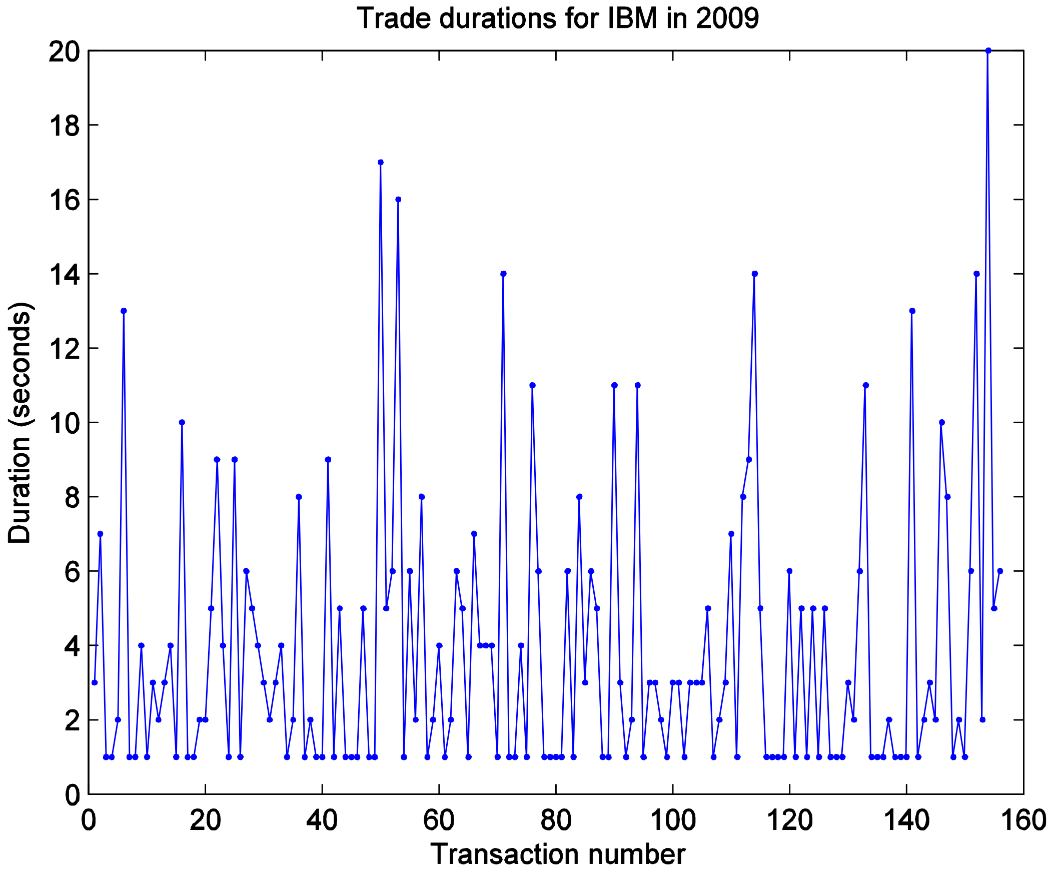


Figure 16.5: Trade durations for IBM in 2009.

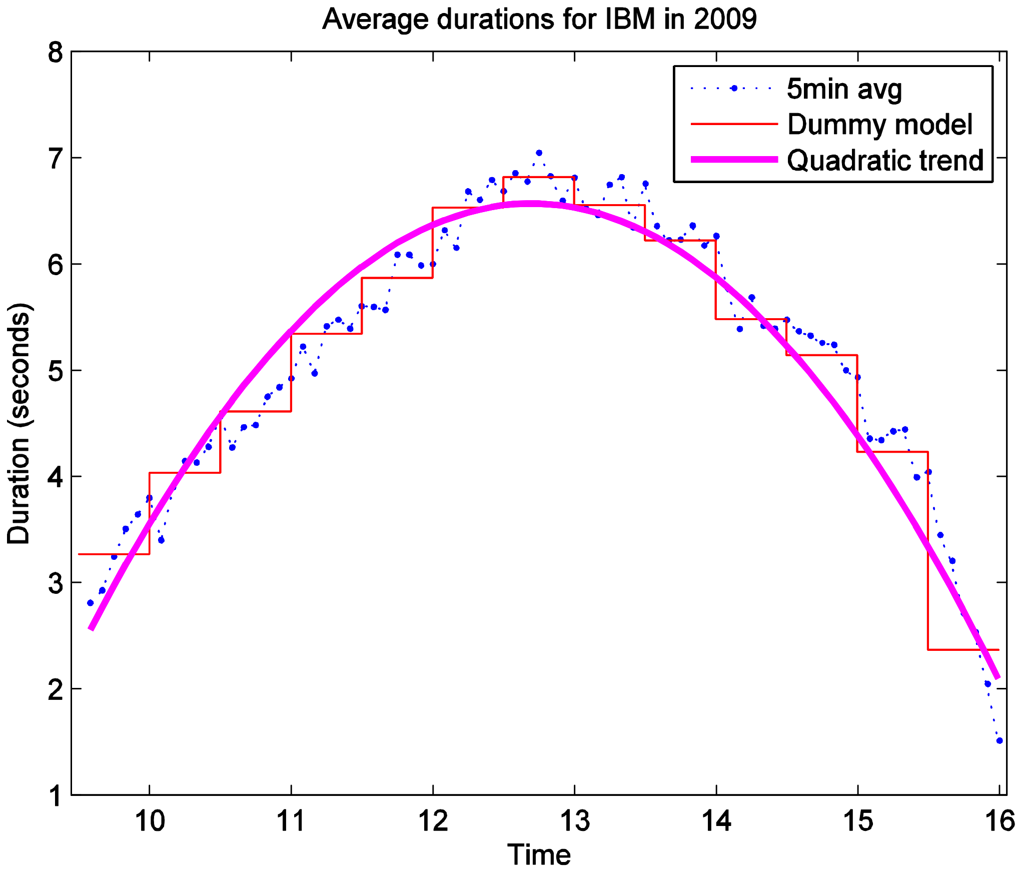


Figure 16.6: Intra-daily pattern in the trade durations for IBM in 2009.

$$
\begin{aligned}
x_j &= s_j x_j^* \\
\text{so} \quad \log x_j &= \log s_j + \log x_j^* \\
\text{where} \quad \log s_j &= \sum_{i=1}^{13} \beta_i D_{i,j} \\
\log s_j &= \gamma_0 + \gamma_1 HRS_j + \gamma_2 HRS_j^2
\end{aligned}
$$

where s_j is the diurnal component, and x_j^* is the de-seasonalised duration. The results are presented in Figure 16.6. Tests for the significance of diurnality using the dummy variable or the trend model yielded p-values of 0.00 in both cases, indicating a significant diurnal pattern.

Given the significance of the trade duration diurnality, we will de-seasonalise the duration data (using the estimates from the quadratic trend regression) before moving on to further analysis:

$$\begin{aligned}
x_j &= s_j x_j^* \\
\log x_j &= \log s_j + u_j, \quad \text{where } u_j = \log\left(x_j^*\right) \\
&= \widehat{\log s_j} + \hat{u}_j, \text{ from the diurnal dummy regression} \\
x_j &= \exp\left\{\widehat{\log s_j}\right\}\exp\left\{\hat{u}_j\right\} \\
E\left[x_j|s_j\right] &= \exp\left\{\widehat{\log s_j}\right\}E\left[\exp\left\{\hat{u}_j\right\}|s_j\right] \\
&= \exp\left\{\widehat{\log s_j}\right\}E\left[\exp\left\{\hat{u}_j\right\}\right], \text{ if } u_j \text{ is independent of } s_j \\
&\approx \exp\left\{\widehat{\log s_j}\right\}\frac{1}{n}\sum_{i=1}^{n}\exp\left\{\hat{u}_i\right\} \\
&\equiv \hat{s}_j \\
\text{and so} \quad x_j^* &= \frac{x_j}{\hat{s}_j}
\end{aligned}$$

16.3.2 Autoregressive Conditional Duration (ACD)

Let us now look at the serial correlation properties of the de-seasonalised duration data, see Figure 16.7. (To save space, I will refer to the data below as 'durations' even though they are the de-seasonalised durations.) Autocorrelation in raw durations is higher than in de-seasonalised durations due to the diurnal pattern in the durations, but the autocorrelations in de-seasonalised durations are still significant, suggesting that we may be able to build a model to forecast them.

One model, proposed by Engle and Russell (1998), for modelling durations is the 'autoregressive conditional duration' model. This model is similar in spirit to the GARCH model. In certain situations the similarity between the ACD and the GARCH models even allows certain results for GARCH models to be applied to ACD models. One possible model for durations is:

$$\begin{aligned}
x_j &= s_j x_j^* \\
x_j^* &= \psi_j \varepsilon_j \\
\text{where} \quad \varepsilon_j &\sim iid\,(1) \\
\log s_j &= \beta_0 + \beta_1 HRS_j + \beta_2 HRS_j^2 \\
x_j^* &= \frac{x_j}{s_j} \\
\psi_j &= \omega + \beta\psi_{j-1} + \alpha x_{j-1}^*,\ \omega > 0,\ \alpha, \beta \geq 0
\end{aligned}$$

That is, durations are decomposed into three parts: a deterministic diurnal component (s_j), an innovation term with conditional mean 1 (ε_j) and a time-varying conditional mean term (ψ_j), which is modelled as an ACD(1,1) process. If we look at de-seasonalised durations we get:

$$E_{j-1}\left[x_j^*\right] = E_{j-1}\left[\psi_j\varepsilon_j\right] = \psi_j E_{j-1}\left[\varepsilon_j\right] = \psi_j$$

since ε_j is *iid* with conditional mean 1.

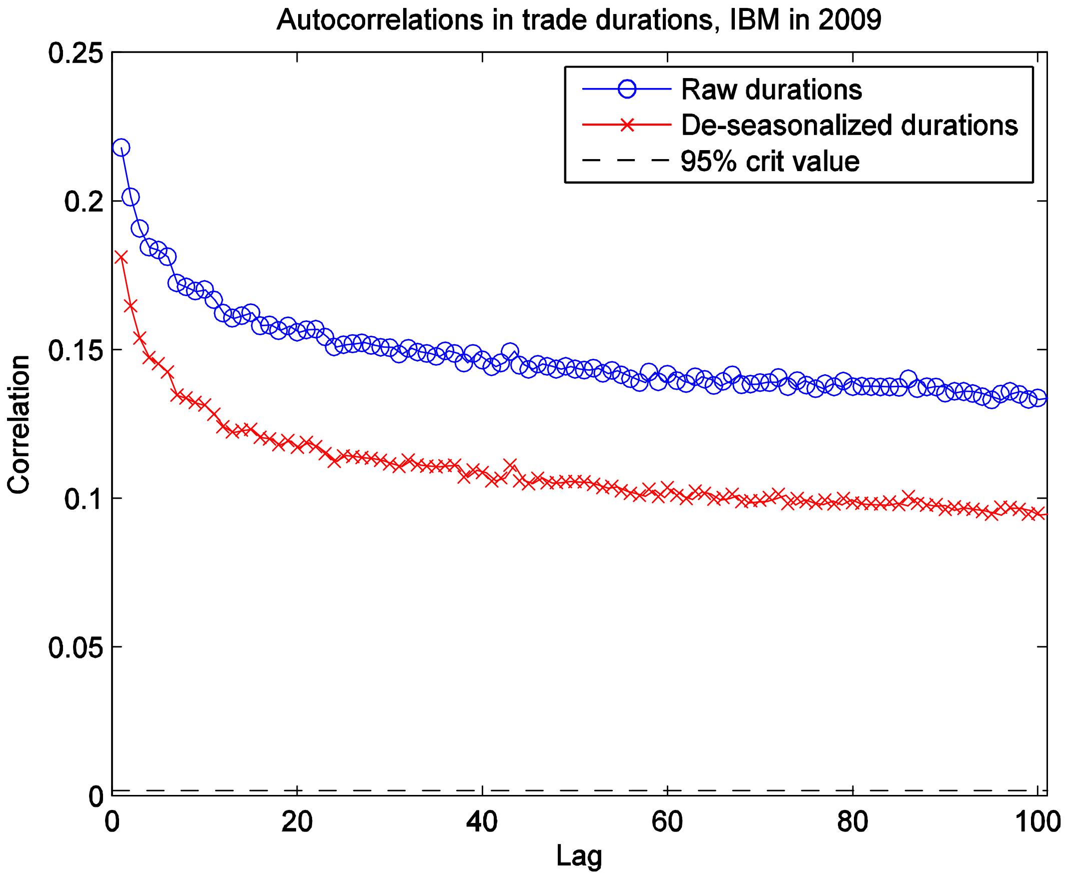


Figure 16.7: Autocorrelation function of durations and de-seasonalised durations for IBM in 2009.

Activity 16.1 Denote durations by x_j. Using an ACD(1,1) model (with *no* diurnal component), find an expression for the two-step ahead conditional mean of duration: $E_j\left[x_{j+2}\right]$.

16.3.3 Estimating the ACD model

Estimation of the parameters of the ACD model can be accomplished via maximum likelihood, once an assumption has been made about the distribution of the innovation terms, ε_j. Note that durations cannot be negative, and so the widely-used normal distribution is not appropriate for this application. The simplest density used in ACD modelling is the exponential density. Recall that a Exponential random variable with parameter γ satisfies:

$$\begin{aligned} Z &\sim \mathit{Exponential}\left(\gamma\right) \\ \text{so} \quad f\left(z|\gamma\right) &= \frac{1}{\gamma}\exp\left\{-\frac{z}{\gamma}\right\} \\ \text{and} \quad E\left[Z\right] &= \gamma \end{aligned}$$

We want to use this assumption for ε_j, which we know to have mean 1, and so we set $\gamma = 1$. Further, since $x_j^* = \psi_j\varepsilon_j$, the above assumption implies

$$x_j^* = \psi_j\varepsilon_j \sim \mathit{Exponential}\left(\psi_j\right)$$

and the conditional density of the de-seasonalised durations becomes:

$$\begin{aligned} f\left(x_j^*|\psi_j\right) &= f\left(x_j^*|\psi_j\left(\omega,\beta,\alpha\right)\right) = \frac{1}{\psi_j}\exp\left\{-\frac{x_j^*}{\psi_j}\right\} \\ &= \frac{1}{\omega+\beta\psi_{j-1}+\alpha x_{j-1}^*}\exp\left\{-\frac{x_j^*}{\omega+\beta\psi_{j-1}+\alpha x_{j-1}^*}\right\} \\ \text{so} \quad \log f\left(x_j^*|\omega,\beta,\alpha\right) &= -\log\left(\omega+\beta\psi_{j-1}+\alpha x_{j-1}^*\right) - \frac{x_j^*}{\omega+\beta\psi_{j-1}+\alpha x_{j-1}^*} \\ \log\mathcal{L}_{ACD} &= -\sum_{j=2}^{n}\log\left(\omega+\beta\psi_{j-1}+\alpha x_{j-1}^*\right) - \sum_{j=2}^{n}\frac{x_j^*}{\omega+\beta\psi_{j-1}+\alpha x_{j-1}^*} \end{aligned}$$

Activity 16.2 The log-likelihood of the ACD model looks quite similar to the estimation of a GARCH model using the normal likelihood function, which you will recall is:

$$\begin{aligned} \mathcal{L}_{GARCH} &= -\frac{n-1}{2}\log\left(2\pi\right) - \frac{1}{2}\sum_{t=2}^{n}\log\left(\omega+\beta\sigma_{t-1}^2+\alpha\left(r_{t-1}-\mu_{t-1}\right)^2\right) \\ &\quad -\frac{1}{2}\sum_{t=2}^{n}\frac{\left(r_t-\mu_t\right)^2}{\omega+\beta\sigma_{t-1}^2+\alpha\left(r_{t-1}-\mu_{t-1}\right)^2} \end{aligned}$$

Show that if we set $\mu_t = 0 \; \forall \; t$, and $r_j = \sqrt{x_j^*}$ then:

a. ψ_j performs the same function as σ_j^2 in a GARCH(1,1) model

b. The GARCH likelihood is simply a linear transformation of the ACD likelihood.

16.3.4 Moments and parameter restrictions for the ACD model

The similarity between the GARCH and the ACD models also extends to required parameter restrictions and moments. Conditions for covariance stationarity, identification and positivity of the duration series are the same for the ACD(1,1) as for the GARCH(1,1):

$$\begin{aligned} \text{Condition 1} &: \omega > 0,\ \alpha, \beta \geq 0, \text{ for positivity} \\ \text{Condition 2} &: \beta = 0 \text{ if } \alpha = 0, \text{ for identification} \\ \text{Condition 3} &: \alpha + \beta < 1, \text{ for covariance stationarity} \end{aligned}$$

If $\alpha + \beta < 1$ then

$$E[\psi_j] = E\left[E_{j-1}\left[x_j^*\right]\right] = E\left[x_j^*\right] = \frac{\omega}{1-\alpha-\beta}$$

and so the (unconditional) average duration implied by the ACD(1,1) is the same formula as that used to obtain the unconditional variance of the residuals from a GARCH(1,1) model.

16.3.5 Application to IBM durations

If we apply the ACD model to the de-seasonalised durations of the IBM data, we obtain the following parameter estimates:

Parameter	Estimate
ω	0.0061
β	0.9425
α	0.0517

The unconditional expected duration implied by the two models can be computed in the same way as the unconditional variance in a GARCH(1,1), if the requirement for weak stationarity ($\alpha + \beta < 1$) is met:

$$\begin{aligned} E[x_j] &= E\left[x_j^*\right] E[s_j] \\ &\approx \frac{\omega}{1-\alpha-\beta} E\left[\exp\left\{\widehat{\log s_j}\right\}\right] \frac{1}{n}\sum_{i=1}^{n} \exp\{\hat{u}_i\} \\ &= \frac{0.0061}{1-0.9425-0.0517} \times 2.8249 \times 1.6253 \\ &= 4.83 \end{aligned}$$

which is very close to the unconditional average duration of 4.7 seconds. A plot of the expected durations for the first three days of the sample period is presented in Figure 16.8.

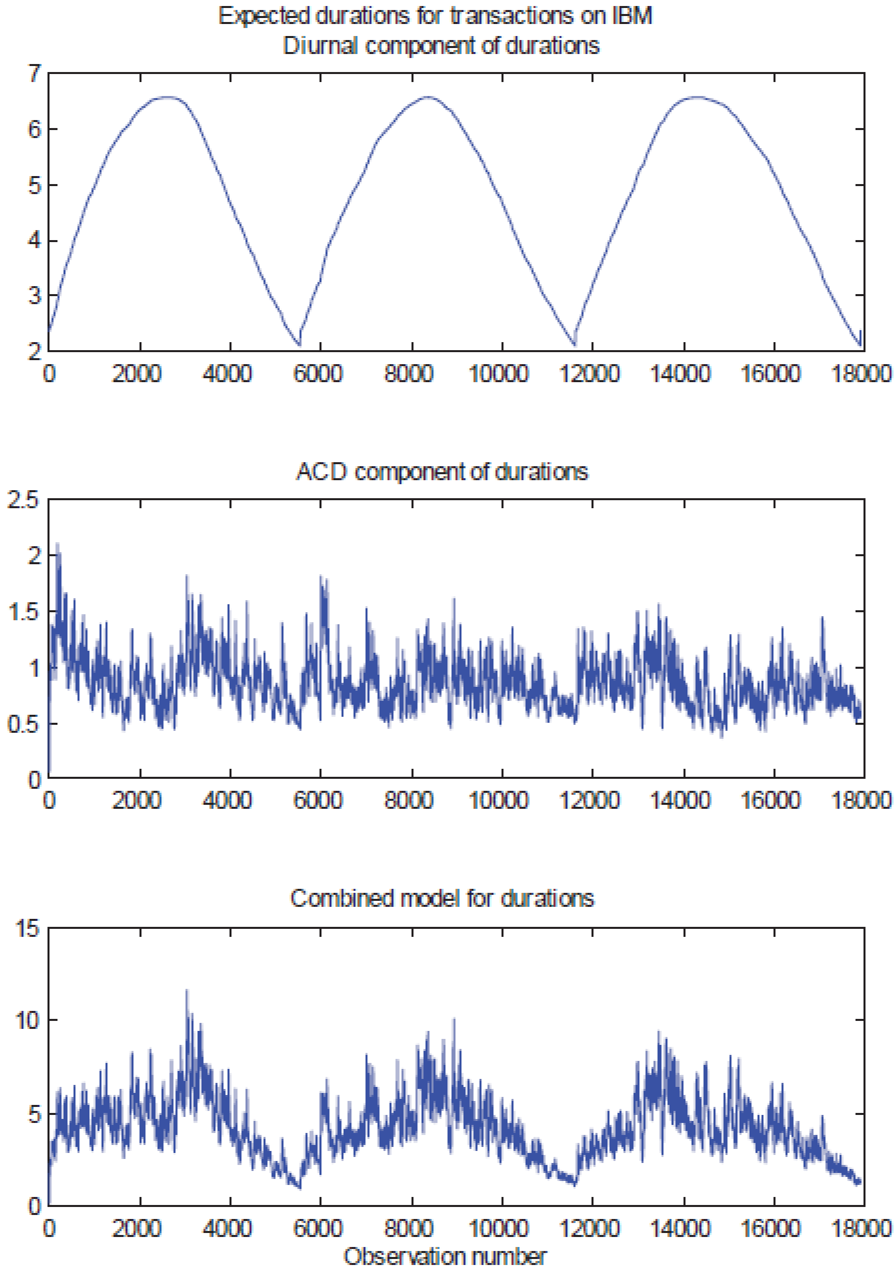


Figure 16.8: Components of the model for IBM trade durations from an ACD model with diurnality.

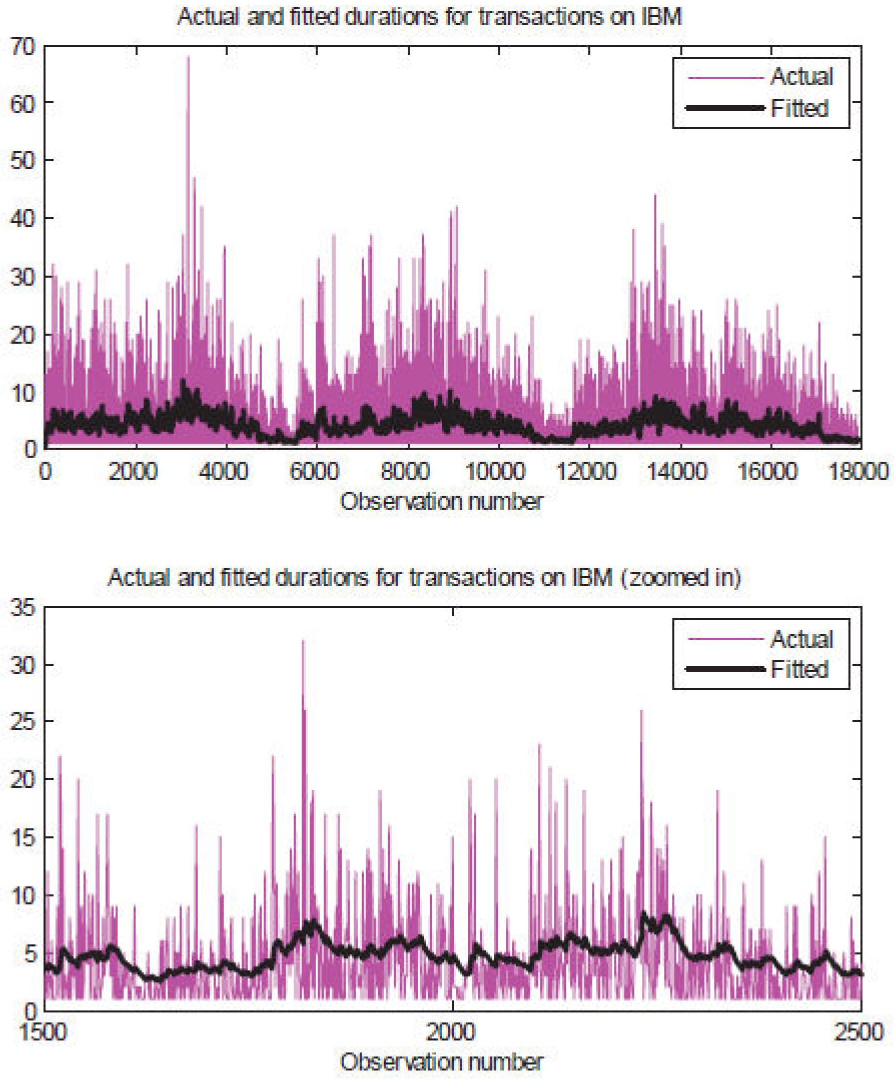


Figure 16.9: Actual and fitted trade durations on IBM, for the first three trading days of 2009.

16.3.6 Evaluating the ACD model

A plot of the actual and fitted durations for the first three days of the sample period is presented in Figure 16.9. There are around 18,000 trades over this period, and the lower panel of this figure zooms in on 1000 observations to provide a closer look at the two series. To formally evaluate the accuracy of these duration forecasts we can conduct a Mincer-Zarnowitz test:

$$\begin{aligned}
\text{Let } \hat{x}_j &\equiv \psi_j \exp\left\{\widehat{\log s_j}\right\} \frac{1}{n}\sum_{i=1}^{n} \exp\left\{\hat{u}_i\right\} \\
x_j &= \beta_0 + \beta_1 \hat{x}_j + e_j \\
H_0 &: \beta_0 = 0 \cap \beta_1 = 1 \\
\text{vs. } H_a &: \beta_0 \neq 0 \cup \beta_1 \neq 1
\end{aligned}$$

Running this regression yields:

$$x_j = \underset{(0.0361)}{-0.1787} + \underset{(0.0083)}{1.0532}\hat{x}_j + e_j\,, \quad R^2 = 0.1156$$

and a χ^2_2 statistic and p-value of 83.72 and 0.00. Thus while this model generates a reasonably high R^2, it is rejected as sub-optimal by the MZ test. With 1,254,720 observations, it is perhaps not surprising that a simple model such as the one above is rejected by the data: sample sizes this large mean that goodness-of-fit tests generally have very high power. The model might be improved by considering a more flexible diurnality model, an ACD(p, q) model with p or q set greater than one, or by considering a more flexible model for durations, such as the Weibull ACD model, the generalised Gamma ACD model or the log-ACD model.

16.4 Overview of chapter

This chapter discussed the issues that arise with high frequency prices arriving at *irregular* points in time. For some analyses, this problem can be avoided by only sampling the data at regular intervals, making it regularly spaced. However, in other applications we may wish to capture this irregular spacing, and we introduced the notion of a trade 'duration' and the ACD model as a way to do so.

16.5 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Define a trade 'duration,' and discuss how this variable changes, on average, through the trade day
- Describe the 'autoregressive conditional duration' (ACD) model for trade durations

16.6 Test your knowledge and understanding

1. (a) What are trade 'durations'? What features do they typically exhibit and how do we model these features?

   (b) What is the 'autoregressive conditional duration' model, and what is it designed to capture?

2. Describe one method for capturing diurnality in durations. How do we estimate the model and how do we test for the presence of a diurnal pattern?

3. You have studied a series of trade durations and noticed that, in addition to serial correlation in duration, it appears that the durations following a duration of longer than 30 seconds are much longer than those following a duration of less than 30 seconds. Propose an extension of the 'autoregressive conditional duration' model to capture this feature of the data. Explain your model.

4. Don't forget to check the VLE for additional practice problems for this chapter.

16.7 Solutions to activities

Activity 16.1

The model is

$$\begin{aligned} x_j &= \psi_j \varepsilon_j \\ \text{where} \quad \varepsilon_j &\sim iid\,(1) \\ \text{and} \quad \psi_j &= \omega + \beta\psi_{j-1} + \alpha x_{j-1} \end{aligned}$$

The two-step ahead conditional mean of duration, $E_j\left[x_{j+2}\right]$, is given by:

$$\begin{aligned} E_j\left[x_{j+2}\right] &= E_j\left[\psi_{j+2}\varepsilon_{j+2}\right] \\ &= E_j\left[\psi_{j+2}E_{j+1}\left[\varepsilon_{j+2}\right]\right] \quad \text{by the LIE} \\ &= E_j\left[\psi_{j+2}\right] \quad \text{since } \varepsilon_j \sim iid\,(1) \\ &= E_j\left[\omega + \beta\psi_{j+1} + \alpha x_{j+1}\right] \\ &= \omega + \beta\psi_{j+1} + \alpha E_j\left[x_{j+1}\right], \text{ since } \psi_{j+1} \text{ known at trade } j \\ &= \omega + \beta\psi_{j+1} + \alpha E_j\left[\psi_{j+1}\varepsilon_{j+1}\right] \\ &= \omega + \beta\psi_{j+1} + \alpha\psi_{j+1}, \text{ since } \psi_{j+1} \text{ known at trade } j \text{ and } \varepsilon_j \sim iid\,(1) \\ &= \omega + (\alpha + \beta)\,\psi_{j+1} \end{aligned}$$

and ψ_{j+1} is known at trade j.

Activity 16.2

If $\mu_t = 0 \ \forall \ t$, and $r_j = \sqrt{x_j^*}$ then

$$\begin{aligned}
\sigma_j^2 &\equiv E_{j-1}\left[r_j^2\right] \text{ in standard GARCH model} \\
&= E_{j-1}\left[x_j\right] \\
&= E_{j-1}\left[\psi_j \cdot \varepsilon_j\right] \\
&= \psi_j
\end{aligned}$$

Thus σ_j^2 and ψ_j have the same interpretation: they are the conditional mean of the squared dependent variable.

If $\mu_t = 0 \ \forall \ t$, and $r_j = \sqrt{x_j^*}$ then the likelihood from the GARCH model becomes

$$\begin{aligned}
\mathcal{L}_{GARCH} &= -\frac{n-1}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\log\left(\omega + \beta\sigma_{t-1}^2 + \alpha\left(r_{t-1} - \mu_{t-1}\right)^2\right) \\
&\quad -\frac{1}{2}\sum_{t=2}^{n}\frac{\left(r_t - \mu_t\right)^2}{\omega + \beta h_{t-1} + \alpha\left(r_{t-1} - \mu_{t-1}\right)^2} \\
&= -\frac{n-1}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\log\left(\omega + \beta\psi_{t-1} + \alpha x_{t-1}\right) \\
&\quad -\frac{1}{2}\sum_{t=2}^{n}\frac{x_t}{\omega + \beta\psi_{t-1} + \alpha x_{t-1}} \\
&= -\frac{n-1}{2}\log(2\pi) + \frac{1}{2}\mathcal{L}_{ACD}
\end{aligned}$$

and so the GARCH likelihood is simply a linear transformation of the ACD likelihood.

In addition to being an interesting link between the two models, it allowed Engle and Russell (1998) to obtain theoretical results on the ACD model from previously known results on the GARCH model. It also has the practical benefit that we may use GARCH software to estimate ACD models: we simply impose a zero mean and estimate a GARCH(1,1) model on the square root of the duration series. The resulting parameter estimates will be those that maximise the ACD log-likelihood, and the forecast 'variance' series from the GARCH software will instead be the time series of expected durations from the ACD model.

Chapter 17

Modelling high frequency financial data: Discreteness

17.1 Introduction

A third unusual feature of high frequency data is the *discreteness* of the observed price changes. Price changes of lower frequency data are often well approximated as continuous random variables, such as Normally-distributed variables, but at very high frequencies this approximation can break down, and there may be benefits to capturing the discreteness of the data. We will consider models for such variables in this chapter.

17.1.1 Aims of the chapter

The aims of this chapter are to:

- Highlight the discrete nature of price changes at very high frequencies
- Discuss methods for capturing persistence in discrete time series
- Describe and analyse the 'ADS' model for high frequency price changes

17.1.2 Learning outcomes

By the end of this chapter, and having completed the activities, you should be able to:

- Interpret 'transition matrices' for discrete time series
- Compute probabilities from the Geometric distribution
- Derive predictions from the ADS model

17.1.3 Essential reading

This chapter serves as the essential reading for this topic.

17.1.4 Further reading

- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354]. Chapter 5

17.1.5 References cited

- Rydberg, T. H. and N. Shephard, ‘Dynamics of trade-by-trade price movements: decomposition and models,’ *Journal of Financial Econometrics*, 2003, 1, pp.2-25.

17.2 Modelling discrete high frequency prices

On June 24 1997, the ‘tick size’ (that is, the minimum amount by which the price of a stock can change) of prices of stocks on the NYSE fell from one-eighth of a dollar to one-sixteenth, and on January 29 2001 the tick size moved to one cent, which remains its current size. When studying stock prices or returns at daily or lower frequencies the discrete nature of stock prices is generally not an issue, but when we study high frequency prices, their discreteness can become a more important issue.

For example, the daily close-to-close change in IBM’s stock price in 2009 took 210 unique values out of the 251 changes during that year, indicating that a approximating these changes by a continuous distribution is likely adequate. This is not too surprising: the average price of IBM in 2009 was $109.27, and the standard deviation of daily log-returns on IBM during 2009 was 1.72. So if we think that the tomorrow’s price will be within ± 2 standard deviations of today’s price, then on an ‘average’ day we expect the price to lie in the interval $[105.57, 113.10]$, which includes a total of 755 distinct prices.

17.3 Discreteness of high frequency IBM prices

When we study high frequency data the number of unique observed price changes can be quite small. For example, changes in high frequency transaction prices of IBM in 2009 were clustered on zero: 30% of changes in tick-level transaction prices were exactly zero. If we extend the range to $\{-1, 0, 1\}$ cents the probability rises to 62%. The range $\{-2, -1, 0, 1, 2\}$ cents covers 81% of changes, and if we include $\{-5, ..., 5\}$ we cover 97.4% of price changes. That is, with just 5 values we cover over 80% of price changes, and with only 11 values we cover over 97% of price changes. This is a level of discreteness that is not observed in lower frequency prices, and which necessitates the consideration of alternative econometric models. It should be noted that for less liquid stocks, or stocks with lower prices (where the 1 cent tick size becomes larger in proportion to the price), the level of discreteness can be even greater.

Note that here we study ΔP_t, the change in the price, rather than returns, $\Delta \log P_t$ or $\Delta P_t / P_{t-1}$. Using either of these definitions of returns obscures the discreteness of the data, which can be an important feature at very high frequencies and is something we want to study here explicitly. The price change ΔP_t is interpretable as a return when the sampling interval is very short, and so these are sometimes also called returns.

IBM price changes in 2009	
Price change (cents)	*Frequency*
Less than -5	1.25
-5	1.09
-4	2.28
-3	4.78
-2	9.42
-1	16.38
0	29.83
+1	16.09
+2	9.23
+3	4.79
+4	2.33
+5	1.17
More than +5	1.36
ALL	100.00

One way of looking at discrete time series data is to use a ‘transition matrix’. This matrix contains ‘transition probabilities’ which take the form:

$$\pi_{ij} \equiv \Pr\left[X_t = s_j | X_{t-1} = s_i\right] = \frac{\Pr\left[X_t = s_j \cap X_{t-1} = s_i\right]}{\Pr\left[X_{t-1} = s_i\right]}$$

where $s \in \{s_1, ..., s_J\}$ is the set of possible values for the variable. In words, π_{ij} measures the probability of observing $X_{t-1} = s_i$ last observation, and observing $X_t = s_i$ this observation. For data such as the IBM data above, we need to include an ‘outer’ state, to capture values of ΔP that are beyond the limits of the values we want to study. For example, to keep things tractable, in the table below I consider five states:

$$s_t = \begin{cases} -2^-, & \text{if } \Delta P_t \leq -2 \\ -1, & \text{if } \Delta P_t = -1 \\ 0, & \text{if } \Delta P_t = 0 \\ +1, & \text{if } \Delta P_t = 1 \\ +2^+, & \text{if } \Delta P_t \geq 2 \end{cases}$$

I estimate the transition probabilities using the following simple formula:

$$\hat{\pi}_{ij} \equiv \frac{\sum_{t=2}^{T} \mathbf{1}\left\{X_t = s_j \cap X_{t-1} = s_i\right\}}{\sum_{t=2}^{T} \mathbf{1}\left\{X_{t-1} = s_i\right\}}$$

The table below presents the resulting transition probability matrix, with the diagonal elements highlighted in bold. It is easiest to read this matrix across the rows. For example, if we observe a price change of less than or equal to -2 cents last observation, then there is a 20% of observing another change of less than or equal to -2 cents. There is a 25% chance that the next observation is a zero change, and there is a 26% chance that the next observation is a change of greater than or equal to 2 cents.

This matrix as a whole reveals that the 'no change' state is often the most likely, confirming how common it is to see zero price changes. The values in the (1,5) and (5,1) elements also reveal that there is *negative autocorrelation* in these prices: large down moves tend to be followed by large upward moves, and vice versa. Finally, the values in the corners and in the center reveal evidence of *volatility clustering:* zero price changes tend to be followed by zero changes, while changes of ±2 tend to be followed by changes of ±2.

Transition matrix for IBM price changes

	ΔP_t				
ΔP_{t-1}	*-2⁻*	*-1*	*0*	*+1*	*+2⁺*
-2⁻	**0.20**	0.15	0.25	0.14	0.26
-1	0.16	**0.17**	0.30	0.19	0.18
0	0.15	0.16	**0.37**	0.16	0.15
+1	0.18	0.20	0.29	**0.16**	0.17
+2⁺	0.26	0.14	0.24	0.15	**0.21**

Activity 17.1 Consider a two-state time series X_t with (one-period) transition matrix:

$$\Pi = \begin{bmatrix} \pi_{11} & 1-\pi_{11} \\ 1-\pi_{22} & \pi_{22} \end{bmatrix}$$

where as usual the (i,j) element represents $\pi_{ij} = \Pr\left[X_{t+1} = s_j | X_t = s_i\right]$, and $s_i \in \{1,2\}$. Find the two-period transition matrix, denoted $\Pi_{(2)}$, with (i,j) element equal to $\pi_{ij}^{(2)} = \Pr\left[X_{t+2} = s_j | X_t = s_i\right]$.

Activity 17.2 Given your answer from Activity 17.1, find the two-period transition matrix if the one-period matrix is

$$\Pi = \begin{bmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{bmatrix}$$

17.4 The 'ADS' model for discrete prices

Rydberg and Shephard (1998) propose a simple model for high frequency price changes based on a decomposition of the observed change into three components. First, assume that prices are measured in units of the 'tick' of the security. For example, for IBM this would correspond to measuring the price in cents. This is done so that the minimum non-zero price change is equal to 1, and larger price changes are equal to 2, 3, etc. Then decompose the price change into:

$$\begin{aligned}
\Delta P_t &= A_t D_t S_t \\
\text{where } A_t &= \begin{cases} 1, & \text{if } |\Delta P_t| > 0 \\ 0, & \text{if } \Delta P_t = 0 \end{cases} \\
D_t &= \begin{cases} -1, & \text{if } \Delta P_t < 0 \\ 0, & \text{if } \Delta P_t = 0 \\ +1, & \text{if } \Delta P_t > 0 \end{cases} \\
S_t &= |\Delta P_t|
\end{aligned}$$

A_t measures whether there was any *activity* at time t, that is, whether the price changed or remained the same. D_t measures the *direction* of the price change, and S_t measures the *size* of the price change in ticks. Decomposing the price change in this way allows us to: (i) capture the 'build up' of probability at $\Delta P_t = 0$ by giving $\Pr[A_t = 0]$ its own free parameter, (ii) allow for some sign predictability through the variable D_t,and (iii) capture the discreteness in the prices through the choice of distribution model for S_t.

Here we construct a simple ADS model using just summary statistics. It is also possible (see Tsay, Chapter 5) to construct a likelihood for this model, which can be used to consider more general specifications and to obtain standard errors. We will not cover that approach in this course.

Persistence in the activity variable can be seen by again using a transition probability matrix, presented below. There we see that if there was no activity last observation, the probability of no activity this observation is 37%, whereas if there was activity last observation then this probability falls to 27%.

	A_t	
A_{t-1}	*0*	*1*
0	**0.3728**	0.6272
1	0.2666	**0.7334**

Some predictability in the sign variable, D_t, can be captured using the following transition matrix, where the lagged variable D_{t-1} can take one of three values (-1, 0, +1), while the current variable D_t can only take the value ±1, as we condition on $A_t = 1$ (and so $\Pr[D_t = 0] = 0$).

		$D_t \vert A_t = 1$	
D_{t-1}	*-1*	*0*	*+1*
-1	0.4723	0	0.5277
0	0.5012	0	0.4988
+1	0.5315	0	0.4685

These numbers reveal that probability of seeing an upward move ($D_t = 1$), conditional on observing a move in either direction, is 53% if the previous observation was a *downward* move, and only 47% if the previous observation was an upward move. This is consistent with the negative autocorrelation discussed in the transition matrix for ΔP_t in the previous sub-section.

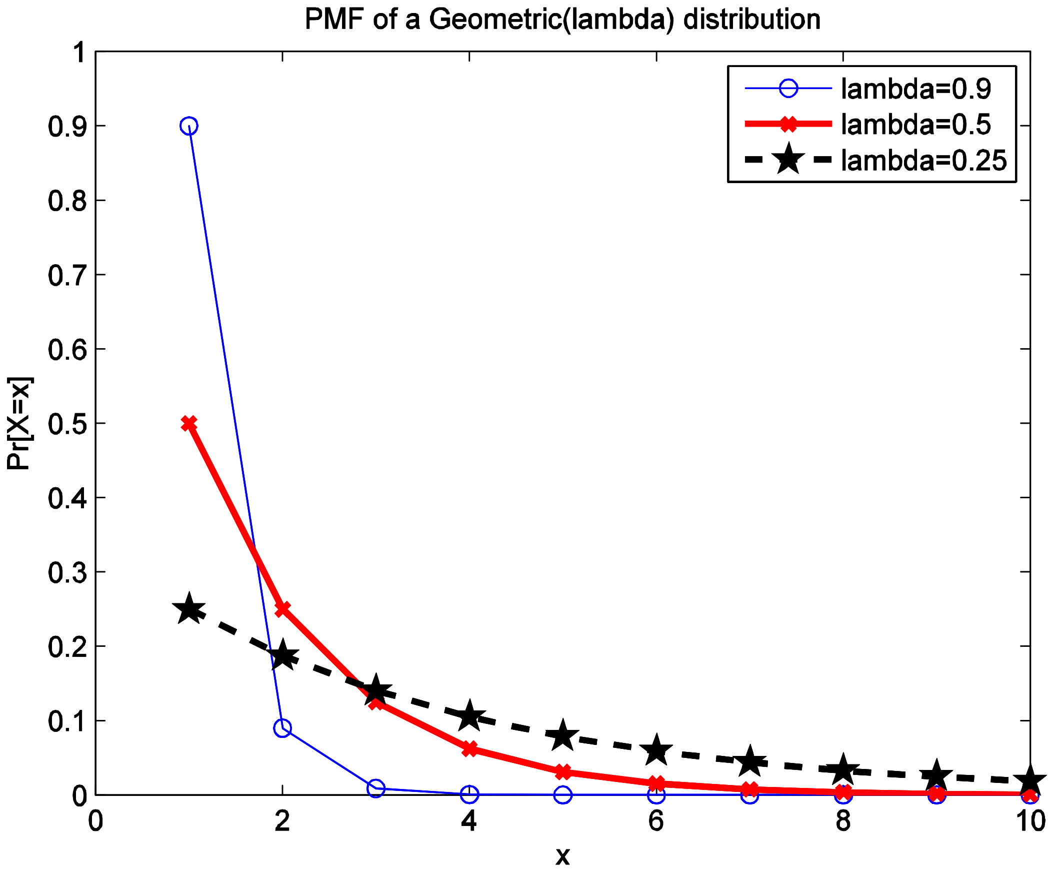


Figure 17.1: Probability mass functions (PMFs) for Geometric dsitributions with parameters 0.9, 0.5, and 0.25.

Finally, we can obtain a distribution for S_t, the size of the price moves. For this, we need a distribution that has discrete, but unbounded support. One possibility is the Geometric distribution, whose PMF is depicted in Figure 17.1.

$$\begin{aligned} X &\sim \mathit{Geometric}\,(\lambda)\,, \quad \text{for } \lambda \in [0,1] \\ \Pr[X=x] &= \lambda\,(1-\lambda)^{x-1}\,, \text{ for } x = 1, 2, ... \end{aligned}$$

This distribution has support on the positive integers (1,2,...) and distributes probability across these depending on the parameter λ. For values of λ near 1, the probability that $X = 1$ is high and declines rapidly to zero as x increases. For lower values of λ the probability of larger values of X is greater.

Activity 17.3 If $X \sim \mathit{Geometric}\,(0.6)$, find $\Pr[X \leq 2]$ and $\Pr[X > 4]$.

We use the Geometric distribution as a model for S_t, and one simple way to estimate its parameter is to use the fact that

$$E[X] = \frac{1}{\lambda} \Rightarrow \lambda = \frac{1}{E[X]}$$

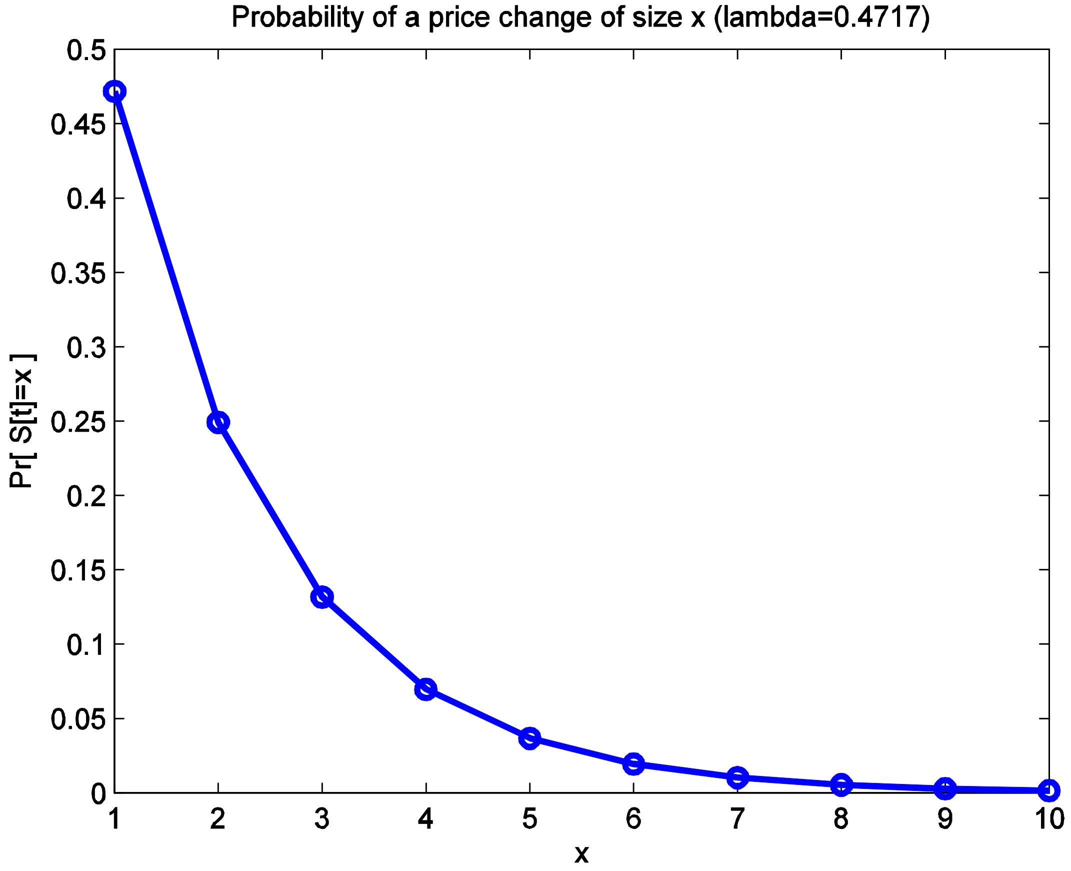


Figure 17.2: Fitted probability mass functions (PMFs) for IBM price chages ($S[t]$) in 2009.

Thus one possible estimator of λ is the inverse of the sample mean of $|S_t|$. In our sample the average size of a price move, conditional on the move being greater than zero, was 2.12 cents, implying $\hat{\lambda} = 1/2.12 = 0.4717$. The resulting PMF for the Geometric distribution with this parameter is presented in Figure 17.2.

Figure 17.3 shows the empirical and fitted distributions of price changes conditioning on three states of the market: $\{A_{t-1} = 0\}$, $\{A_{t-1} = 1 \cap D_{t-1} = -1\}$ and $\{A_{t-1} = 1 \cap D_{t-1} = +1\}$. For comparison, I also show the fit of the Normal density to these price changes. The top panel of this figure shows that the ADS model provides a very good fit in all three periods: following no activity, following a downward move, and following an upward move. Also, in all cases we see that the Normal distribution does a relatively poor job of approximating these distributions.

> **Activity 17.4** Given the fitted ADS model presented in this section, compute the following:
>
> 1. $\Pr\left[\Delta P_t = 0 | A_{t-1} = 0\right]$
> 2. $\Pr\left[|\Delta P_t| \leq 1 | A_{t-1} = 0\right]$
> 3. $E\left[\Delta P_t | A_{t-1} = 0\right]$

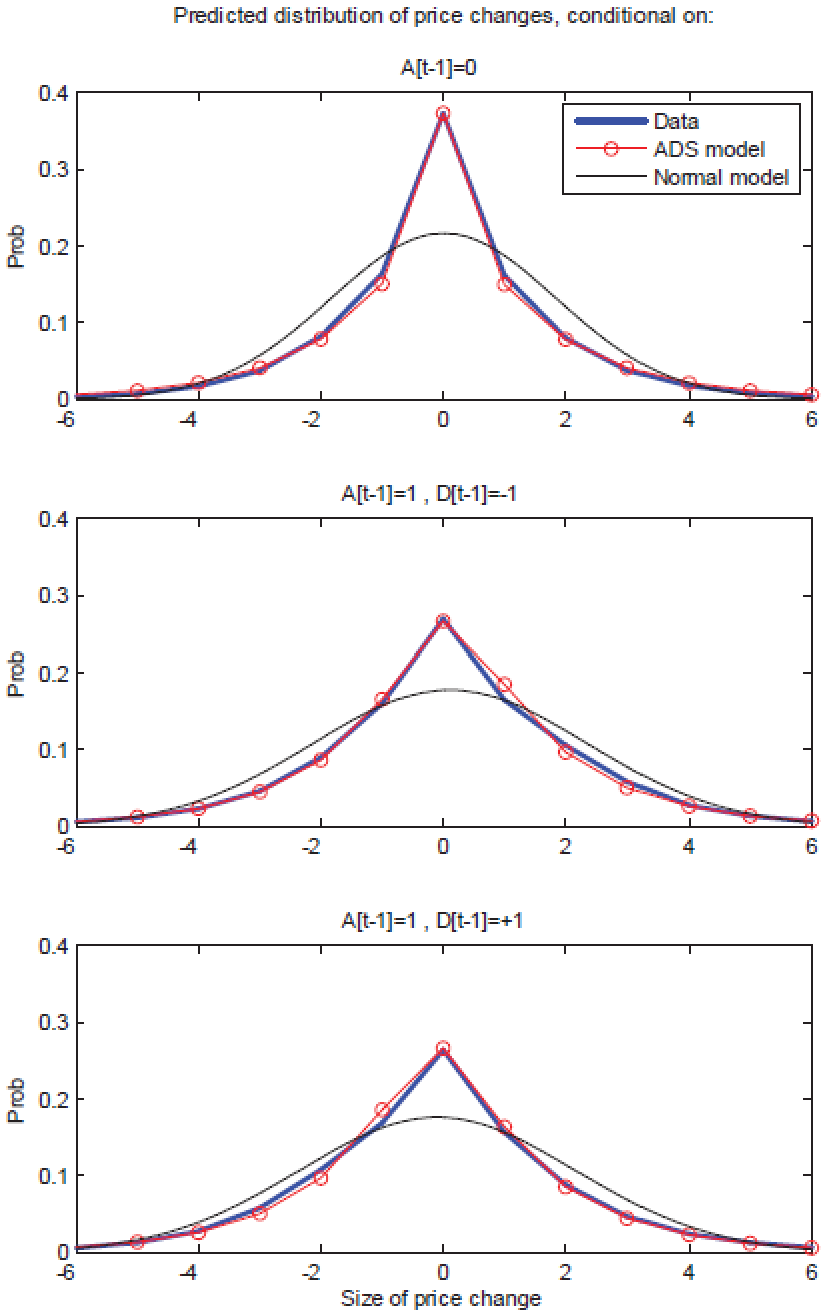


Figure 17.3: Empirical and fitted distributions of price changes for three different cases.

17.5 Overview of chapter

High frequency prices differ from low frequency prices in their degree of *discreteness*: low frequency price changes tend to take a wide variety of values, and can thus be reasonably approximated by continuous distributions (like the Normal). High frequency price changes tend to take only a few values (such as 0, ±1 penny, ±2 pennies, etc) and require different models for their analysis. This chapter review some methods for studying discrete-valued time series.

17.6 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Interpret 'transition matrices' for discrete time series
- Compute probabilities from the Geometric distribution
- Derive predictions from the ADS model

17.7 Test your knowledge and understanding

1. Consider the following transition matrix for a time series that can take values $(-2,-1,0,1,2)$. Does this time series exhibit evidence of volatility clustering?

Transition matrix for ΔP_t

	ΔP_{t+1}				
ΔP_t	*-2*	*-1*	*0*	*+1*	*+2*
-2	0.12	0.16	0.42	0.18	0.12
-1	0.07	0.18	0.49	0.17	0.09
0	0.05	0.17	0.56	0.17	0.05
+1	0.08	0.16	0.51	0.16	0.09
+2	0.14	0.14	0.41	0.17	0.14

2. Consider the following transition matrix for a time series that can take values $(1,2)$. Find the two-step transition matrix, containing the probabilities for going from ΔP_t to ΔP_t.

Transition matrix for ΔP_t

	ΔP_{t+1}	
ΔP_t	*+1*	*+2*
+1	0.8	0.2
+2	0.1	0.9

3. Don't forget to check the VLE for additional practice problems for this chapter.

17.8 Solutions to activities

Activity 17.1

The two-period transition matrix, like the one-period matrix, has four elements, but since the rows sum to one we only have to work out two of these.

Let's start with $\pi_{11}^{(2)}$, the probability of moving from $X_t = 1$ to $X_{t+2} = 1$. Now we think about all the possible paths to go from $X_t = 1$ to $X_{t+2} = 1$. Since this is a two-state time series, there are only two options: we could have $(X_t, X_{t+1}, X_{t+2}) = (1, 1, 1)$ or $(X_t, X_{t+1}, X_{t+2}) = (1, 2, 1)$. So

$$\begin{aligned}
\pi_{11}^{(2)} &= \Pr\left[X_{t+2} = 1|X_t = 1\right] \\
&= \Pr\left[X_{t+2} = 1, X_{t+1} = 1|X_t = 1\right] \\
&\quad + \Pr\left[X_{t+2} = 1, X_{t+1} = 2|X_t = 1\right] \\
&= \Pr\left[X_{t+2} = 1|X_t = 1, X_{t+1} = 1\right] \Pr\left[X_{t+1} = 1|X_t = 1\right] \\
&\quad + \Pr\left[X_{t+2} = 1|X_t = 1, X_{t+1} = 2\right] \Pr\left[X_{t+1} = 2|X_t = 1\right] \\
&= \Pr\left[X_{t+2} = 1|X_{t+1} = 1\right] \Pr\left[X_{t+1} = 1|X_t = 1\right] \\
&\quad + \Pr\left[X_{t+2} = 1|X_{t+1} = 2\right] \Pr\left[X_{t+1} = 2|X_t = 1\right] \\
&= \pi_{11}^2 + \pi_{21}\pi_{12}
\end{aligned}$$

where the third equality comes from breaking up the probability of two events into the conditional probability of the first given the second multiplied by the probability of the second, and fourth equality comes from the fact that X_{t+2} only depends on X_{t+1} not on both X_{t+1} and X_t, and the final equality uses the fact that all of these probabilities are one-step probabilities. Given $\pi_{11}^{(2)}$, we then have $\pi_{12}^{(2)}$:

$$\begin{aligned}
\pi_{12}^{(2)} &= 1 - \pi_{11}^{(2)} \\
&= 1 - \pi_{11}^2 - \pi_{21}\pi_{12}
\end{aligned}$$

Similar steps for $\pi_{22}^{(2)}$ yield:

$$\begin{aligned}
\pi_{22}^{(2)} &= \Pr\left[X_{t+2} = 2|X_t = 2\right] \\
&= \Pr\left[X_{t+2} = 2, X_{t+1} = 1|X_t = 2\right] \\
&\quad + \Pr\left[X_{t+2} = 2, X_{t+1} = 2|X_t = 2\right] \\
&= \Pr\left[X_{t+2} = 2|X_t = 2, X_{t+1} = 1\right] \Pr\left[X_{t+1} = 1|X_t = 2\right] \\
&\quad + \Pr\left[X_{t+2} = 2|X_t = 2, X_{t+1} = 2\right] \Pr\left[X_{t+1} = 2|X_t = 2\right] \\
&= \Pr\left[X_{t+2} = 2|X_{t+1} = 1\right] \Pr\left[X_{t+1} = 1|X_t = 2\right] \\
&\quad + \Pr\left[X_{t+2} = 2|X_{t+1} = 2\right] \Pr\left[X_{t+1} = 2|X_t = 2\right] \\
&= \pi_{12}\pi_{21} + \pi_{22}^2 \\
\text{and } \pi_{21}^{(2)} &= 1 - \pi_{22}^{(2)} = 1 - \pi_{12}\pi_{21} - \pi_{22}^2
\end{aligned}$$

and so we have shown

$$\Pi_{(2)} = \left[\begin{array}{cc} \pi_{11}^2 + \pi_{21}\pi_{12} & 1 - \pi_{11}^2 - \pi_{21}\pi_{12} \\ 1 - \pi_{12}\pi_{21} - \pi_{22}^2 & \pi_{12}\pi_{21} + \pi_{22}^2 \end{array}\right]$$

Aside: It turns out that this is equal to the matrix square of Π :

$$\begin{aligned}\Pi_{(2)} &= \Pi^2 \equiv \Pi \cdot \Pi \\ &= \left[\begin{array}{cc} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{array}\right]\left[\begin{array}{cc} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{array}\right] \\ &= \left[\begin{array}{cc} \pi_{11}^2+\pi_{12}\pi_{21} & \pi_{11}\pi_{12}+\pi_{12}\pi_{22} \\ \pi_{21}\pi_{11}+\pi_{22}\pi_{21} & \pi_{21}\pi_{12}+\pi_{22}^2 \end{array}\right]\end{aligned}$$

and in fact, the $p-$step transition matrix is obtained as the p^{th} matrix power of the one-step transition matrix:

$$\Pi_{(p)} = \Pi^p \equiv \underbrace{\Pi \cdot \Pi \cdot ... \cdot \Pi}_{\text{p times}}$$

Activity 17.2

If the one-period matrix is:

$$\Pi = \left[\begin{array}{cc} 0.9 & 0.1 \\ 0.05 & 0.95 \end{array}\right]$$

then using the above formula we have

$$\begin{aligned}\Pi_{(2)} &= \Pi \cdot \Pi \\ &= \left[\begin{array}{cc} 0.9 & 0.1 \\ 0.05 & 0.95 \end{array}\right]\left[\begin{array}{cc} 0.9 & 0.1 \\ 0.05 & 0.95 \end{array}\right] \\ &= \left[\begin{array}{cc} 0.8150 & 0.1850 \\ 0.0925 & 0.9075 \end{array}\right]\end{aligned}$$

Activity 17.3

If $X \sim \mathit{Geometric}\,(0.6)$,then $\Pr[X \leq 2] = \Pr[X=1] + \Pr[X=2]$, since the range of possible values for X are $(1,2,3,...)$. Thus we only need to calculate these two probabilities, which we can do using the formula for the PMF of a Geometric distribution:

$$\begin{aligned} \Pr[X=x] &= \lambda(1-\lambda)^{x-1}, \text{ for } x=1,2,... \\ \text{so} \quad \Pr[X=1] &= \lambda(1-\lambda)^0 = \lambda = 0.6 \\ \Pr[X=2] &= \lambda(1-\lambda)^1 = 0.24 \\ \Pr[X\leq 2] &= 0.84. \end{aligned}$$

Since X is unbounded, it is not convenient to obtain $\Pr[X>4]$ as $\Pr[X=4]+\Pr[X=5]+\Pr[X=6]$...Instead, we will use the fact that $\Pr[X>4] = 1-\Pr[X\leq 4] = 1-(\Pr[X=1]+\Pr[X=2]+\Pr[X=3]+\Pr[X=4])$. We obtained the first two of these probabilities above, and the other two are:

$$\begin{aligned} \Pr[X=3] &= \lambda(1-\lambda)^2 = 0.096 \\ \Pr[X=4] &= \lambda(1-\lambda)^3 = 0.0384 \\ \text{So } \Pr[X>4] &= 1-\Pr[X\leq 4] \\ &= 0.0256 \end{aligned}$$

Activity 17.4

(1):

$$\begin{aligned}
\Pr\left[\Delta P_t = 0|A_{t-1} = 0\right] &= \Pr\left[A_t D_t S_t = 0|A_{t-1} = 0\right] \\
&= \Pr\left[A_t = 0|A_{t-1} = 0\right], \text{ since } D_t = \pm 1 \text{ and } S_t = 1, 2, ... \\
&= 0.3728, \text{ from the table in the notes}
\end{aligned}$$

(2):

$$\begin{aligned}
\Pr\left[|\Delta P_t| \leq 1|A_{t-1} = 0\right] &= \Pr\left[\Delta P_t = 0|A_{t-1} = 0\right] \\
&\quad + \Pr\left[\Delta P_t = 1|A_{t-1} = 0\right] \\
&\quad + \Pr\left[\Delta P_t = -1|A_{t-1} = 0\right]
\end{aligned}$$

We have the first of these terms from part (1), now let's compute the second

$$\begin{aligned}
\Pr\left[\Delta P_t = 1|A_{t-1} = 0\right] &= \Pr\left[A_t D_t S_t = 1|A_{t-1} = 0\right] \\
&= \Pr\left[A_t = 0|A_{t-1} = 0\right] \Pr\left[A_t D_t S_t = 1|A_{t-1} = 0, A_t = 0\right] \\
&\quad + \Pr\left[A_t = 1|A_{t-1} = 0\right] \Pr\left[A_t D_t S_t = 1|A_{t-1} = 0, A_t = 1\right] \\
&= \Pr\left[A_t = 1|A_{t-1} = 0\right] \Pr\left[D_t S_t = 1|A_{t-1} = 0, A_t = 1\right] \\
&= \Pr\left[A_t = 1|A_{t-1} = 0\right] \times \\
&\quad \{\Pr\left[D_t = -1|A_{t-1} = 0, A_t = 1\right] \Pr\left[D_t S_t = 1|A_{t-1} = 0, A_t = 1, D_t = -1\right] \\
&\quad + \Pr\left[D_t = 1|A_{t-1} = 0, A_t = 1\right] \Pr\left[D_t S_t = 1|A_{t-1} = 0, A_t = 1, D_t = 1\right]\} \\
&= \Pr\left[A_t = 1|A_{t-1} = 0\right] \Pr\left[D_t = 1|A_{t-1} = 0, A_t = 1\right] \times \\
&\quad \Pr\left[S_t = 1|A_{t-1} = 0, A_t = 1, D_t = 1\right] \\
&= 0.6272 \times 0.4988 \times \left\{\lambda \left(1 - \lambda\right)^{1-1}\right\}, \text{ where } \lambda = 0.4717 \\
&= 0.1476
\end{aligned}$$

Similarly:

$$\begin{aligned}
\Pr\left[\Delta P_t = -1|A_{t-1} = 0\right] &= \Pr\left[A_t = 1|A_{t-1} = 0\right] \Pr\left[D_t = -1|A_{t-1} = 0, A_t = 1\right] \times \\
&\quad \Pr\left[S_t = 1|A_{t-1} = 0, A_t = 1, D_t = 1\right] \\
&= 0.6272 \times 0.5012 \times \left\{\lambda \left(1 - \lambda\right)^{1-1}\right\}, \text{ where } \lambda = 0.4717 \\
&= 0.1483
\end{aligned}$$

And pulling these together we find:

$$\begin{aligned}
\Pr\left[|\Delta P_t| \leq 1|A_{t-1} = 0\right] &= 0.3728 + 0.1476 + 0.1483 \\
&= 0.6687
\end{aligned}$$

(3):

$$\begin{aligned}
E\left[\Delta P_t|A_{t-1} = 0\right] &= E\left[A_t D_t S_t|A_{t-1} = 0\right] \\
&= \Pr\left[A_t = 0|A_{t-1} = 0\right] E\left[A_t D_t S_t|A_{t-1} = 0, A_t = 0\right] \\
&\quad + \Pr\left[A_t = 1|A_{t-1} = 0\right] E\left[A_t D_t S_t|A_{t-1} = 0, A_t = 1\right] \\
&= 0.3728 \times 0 + 0.6272 \times E\left[D_t S_t|A_{t-1} = 0, A_t = 1\right]
\end{aligned}$$

Now let's work out the last term

$$\begin{aligned}E\left[D_tS_t|A_{t-1}=0,A_t=1\right] &= E\left[D_t|A_{t-1}=0,A_t=1\right]E\left[S_t|A_{t-1}=0,A_t=1\right]\\ &= E\left[D_t|A_{t-1}=0,A_t=1\right]\times 2.12\end{aligned}$$

since $S_t \sim iid\ Geometric\left(\lambda\right)$, so S_t is independent of D_t, and $E\left[S_t\right]=1/\lambda=2.12$. So now we compute

$$\begin{aligned}E\left[D_t|A_{t-1}=0,A_t=1\right] &= -1\times\Pr\left[D_t=-1|A_{t-1}=0,A_t=1\right]\\ &\quad +0\times\Pr\left[D_t=0|A_{t-1}=0,A_t=1\right]\\ &\quad +1\times\Pr\left[D_t=+1|A_{t-1}=0,A_t=1\right]\\ &= -0.5012+0+0.4988\\ &= -0.0024\end{aligned}$$

Then we pull these terms together:

$$\begin{aligned}E\left[\Delta P_t|A_{t-1}=0\right] &= 0.3728\times 0+0.6272\times E\left[D_tS_t|A_{t-1}=0,A_t=1\right]\\ &= 0.6272\times(-0.0024)\times 2.12\\ &= -0.0032\end{aligned}$$

Chapter 18
Spurious regressions and persistent time series

18.1 Introduction

Early in this guide we discussed transforming asset prices to returns, and then conducting our analyses on the returns series. This was motivated by the claim that prices have statistical properties that make them more difficult to handle – in this chapter we will look at these properties. Firstly, we will discuss two popular models for *persistent* time series, and then we will look at the problem of *spurious regressions*. We will conclude with a famous test to detect whether a time series contains a ‘unit root’ (a form of persistence).

18.1.1 Aims of the chapter

The aims of this chapter are to:

- Introduce the time trend and random walk models for persistent time series
- Discuss the problem of ‘spurious regressions’ and how standard econometric inference breaks down when applied to very persistent data
- Introduce the Dickey-Fuller test for a unit root, and discuss how to implement and interpret this test.

18.1.2 Learning outcomes

By the end of this chapter, and having completed the essential reading and activities, you should be able to:

- Describe the differences between the time trend model and the random walk model for persistent time series
- Discuss the problem of ‘spurious regressions’ and know when that problem might arise
- Implement and interpret the Dickey-Fuller test for a unit root

18.1.3 Essential reading

- Diebold, F.X. *Elements of Forecasting.* (Thomson South-Western, Canada, 2006) fourth edition [ISBN 9780324323597]. Chapters 5 and 13.

18.1.4 Further reading

- Stock, J.H. and M.W. Watson *Introduction to Econometrics.* (Pearson Education, Boston, 2010) third edition. [ISBN 9781408264331]. Chapter 14, Section 6.
- Tsay, R.S., *Analysis of Financial Time Series.* (John Wiley & Sons, New Jersey, 2010) third edition. [ISBN 9780470414354]. Chapter 2, Section 7.
- Wooldridge, J.M. *Introductory Econometrics: A Modern Approach.* (South-Western, USA, 2012) fifth edition. [ISBN 9781111531041]. Chapter 18.

18.1.5 References cited

- Dickey, D. A. and W. A. Fuller, 'Distribution of the Estimators for Autoregressive Time Series with a Unit Root,' *Journal of the American Statistical Association*, 1979, 74(366), pp.427–431.
- Granger, C. W. J. and P. Newbold, P. 'Spurious regressions in econometrics,' *Journal of Econometrics*, 1974, 2(2), pp.111–120.

18.2 Random walks and time trends

Two main types of models have been used to capture non-stationarity in economic and financial data: models with a time trend, and random walk models:

$$\begin{aligned} Y_{t+1} &= \beta_0 + \beta_1 t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim WN\left(0, \sigma^2\right) \\ Y_{t+1} &= \phi_0 + Y_t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim WN\left(0, \sigma^2\right) \end{aligned}$$

An example of forecasts from these models is given in Figure 18.1. Notice that the forecasts from the random walk model (also known as a 'unit root' model) just extrapolate from where the series is at the point the forecast is made ($t = 50$ in the figure), in this case going up as ϕ_0 is positive, whereas the forecasts from the trend model assume that the series will revert back to its time trend. Notice also that the forecast intervals are different: the time trend model's intervals are constant; not a function of the forecast horizon. The intervals from the random walk model grow with the square-root of the forecast horizon (since the variances grow linearly with the horizon, which you will verify in Activity 18.1).

Random walk models are far more commonly used in finance, and they are usually used on *log* prices or exchange rates. This is so that the difference in the log variable is interpretable as the (continuously-compounded) return on the asset.

The random walk model above is also known as an AR(1) model with a 'unit root.' This is because the AR(1) coefficient is equal to one. (There is a large, advanced, econometrics literature on unit roots, see the further reading references for details.) A unit root process is *nonstationary*, and standard econometric methods often break down when applied to such data.

Consider how an AR(1) process adjusts to innovations through time. Let

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \ \varepsilon_t \sim WN\left(0\right)$$

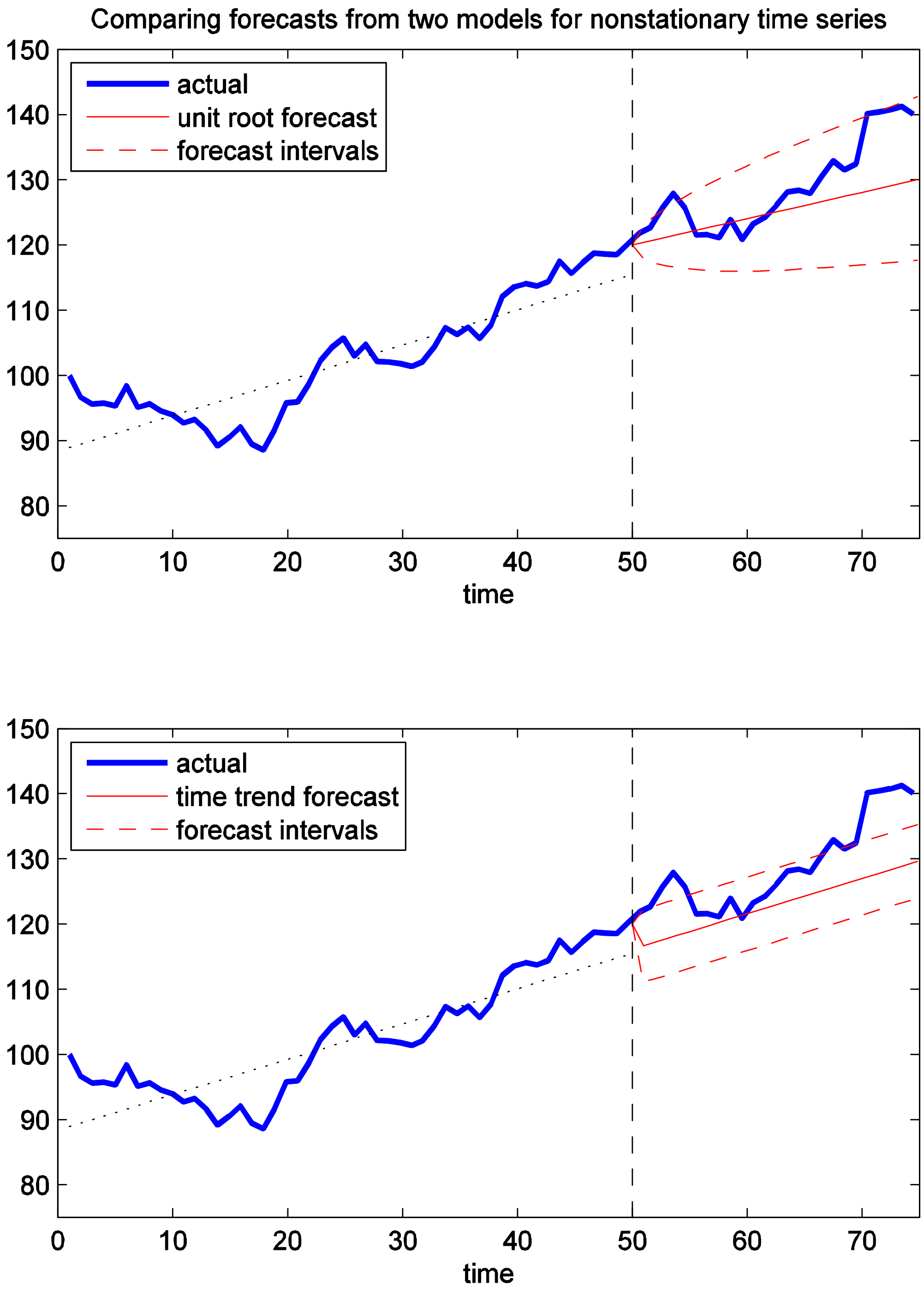


Figure 18.1: Forecasts from two models for nonstationary time series, with 80% forecast intervals.

How is the current value of Y affected by past innovations, ε_{t-j}?

$$\begin{aligned}
Y_t &= \phi Y_{t-1} + \varepsilon_t \\
&= \phi \left(\phi Y_{t-2} + \varepsilon_{t-1} \right) + \varepsilon_t \\
&= \phi^2 Y_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t \\
&= \phi^2 \left(\phi Y_{t-3} + \varepsilon_{t-2} \right) + \phi \varepsilon_{t-1} + \varepsilon_t \\
&= \phi^3 Y_{t-3} + \phi^2 \varepsilon_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t \\
&= ... \\
&= \phi^t Y_0 + \phi^{t-1} \varepsilon_1 + \phi^{t-2} \varepsilon_2 + ... + \phi \varepsilon_{t-1} + \varepsilon_t \\
&= \phi^t Y_0 + \sum_{j=0}^{t-1} \phi^j \varepsilon_{t-j}
\end{aligned}$$

If $|\phi| < 1$ then $\phi^j \to 0$ as $j \to \infty$, and so shocks die out; innovations to the series many periods into the past have only a small impact on the series today. This is the stationary case. Each innovation, or shock, causes the series to move away from its long-run (unconditional) mean, but the shocks fade eventually.

If $\phi = 1$ then $\phi^j = 1 \; \forall \; j$ and so every shock in the past is equally important for determining the level of the series today. The shocks are persistent.

If $\phi > 1$ then $\phi^j \to \infty$ as $j \to \infty$, and so old shocks have larger and larger impacts through time. This is an explosive case, and is not a realistic description of many economic or financial variables. For this reason most attention is paid to the cases that $|\phi| < 1$ and $\phi = 1$. In Figure 18.2 I plot the theoretical autocorrelation functions for AR(1) processes with autoregressive coefficients close to one. We can see that for large ϕ the autocorrelation function gets flatter.

> **Activity 18.1** Consider the following two models
>
> $$\begin{aligned}
> X_{t+1} &= \beta_0 + \beta_1 t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim iid\; N\left(0, \sigma^2\right) \\
> Y_{t+1} &= \phi_0 + Y_t + \eta_{t+1}, \quad \eta_{t+1} \sim iid\; N\left(0, \sigma^2\right)
> \end{aligned}$$
>
> 1. Find $E_t\left[X_{t+3}\right]$ and $E_t\left[Y_{t+3}\right]$
> 2. Find $V_t\left[X_{t+3}\right]$ and $V_t\left[Y_{t+3}\right]$

18.3 Spurious regressions

In this section we will look at one of the problems that can arise when standard econometric methods are applied to nonstationary series, such as price series or index levels, rather than the returns on these series. Consider the following example:

$$\begin{aligned}
X_{t+1} &= \mu_x + X_t + \varepsilon_{t+1}, \; t = 1, 2, ..., T \\
Y_{t+1} &= \mu_y + Y_t + \eta_{t+1}, \; t = 1, 2, ..., T \\
X_0 &= 0, \; Y_0 = 0 \\
\left(\varepsilon_{t+1}, \eta_{t+1}\right) &\sim iid\; N\left(0, 1\right)
\end{aligned}$$

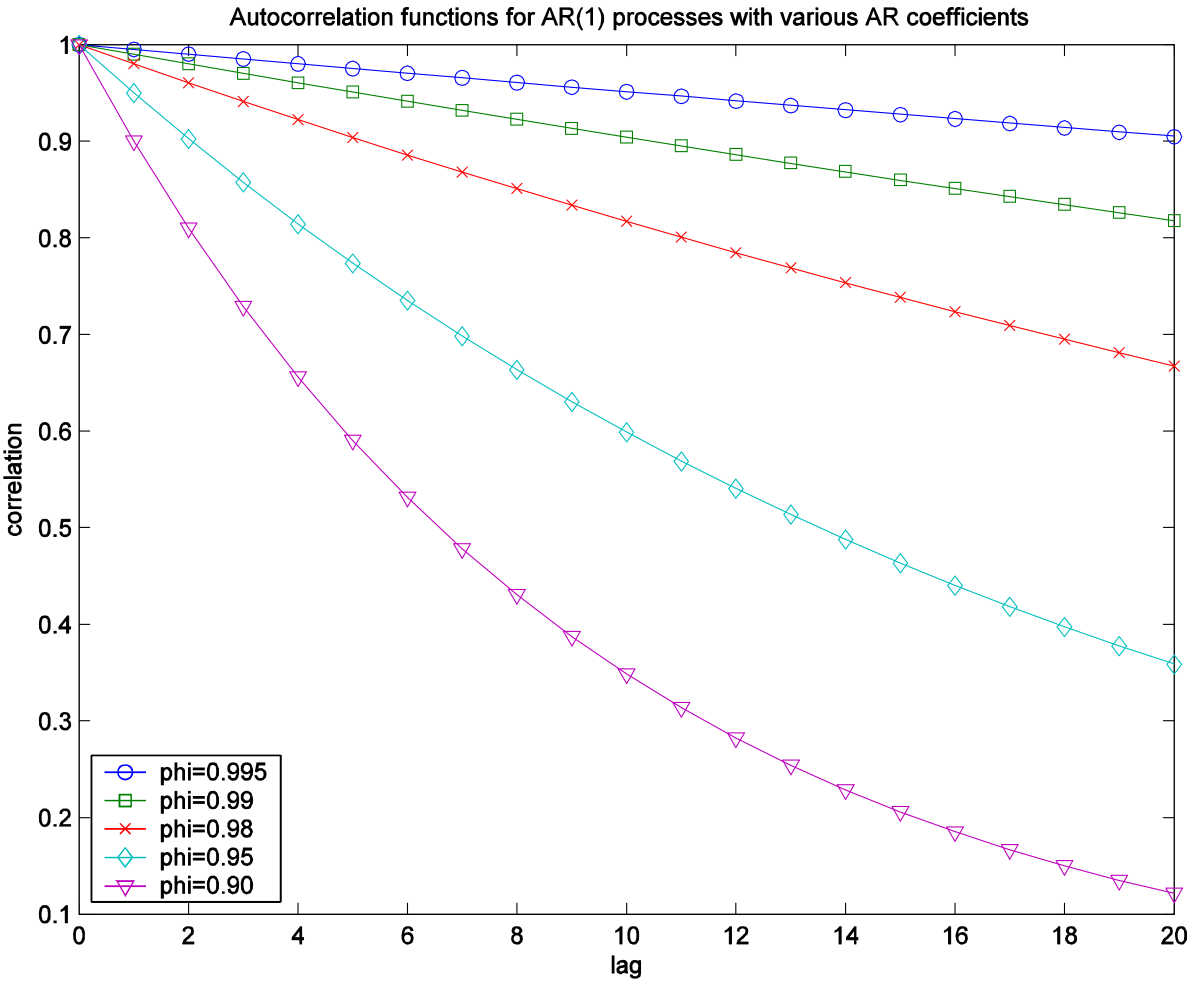


Figure 18.2: Autocorrelation functions for AR(1) processes with AR coefficients close to one.

X_t and Y_t are independent random walks, with standard normal innovations. In Figures 18.3 and 18.4 I plot 6 sample time paths of these two variables over 250 periods. The first figure has $\mu_x = \mu_y = 0$ while the second has $\mu_x = \mu_y = 0.1$. To the naked eye it would appear that in some cases these two series are related; they appear to move in the same direction over reasonably long periods. It might then be reasonable to conclude that one of the series could be useful for predicting the other. However this is *not* the case: these series are, by construction, independent of each other. Their apparent relationship is *spurious.*In a classic paper Granger and Newbold (1974) showed that the t-statistics and R^2 *s* from a regression of Y_t on X_t in the above framework are not what we would usually expect. The tables below are based on a replication of the simulation results of Granger and Newbold, for time series of length $T = 250$. They report statistics on the distribution of the t-statistics and R^2 *s* from the following regression:

$$Y_t = \beta_0 + \beta_1 X_t + e_t$$

The tables report both the simulated distribution results, and the distribution that we would expect using standard econometric theory: under H_0, $tstat \sim N(0,1)$, and $T \times R^2/(1-R^2) \sim \chi^2_1$. The top panel presents contains the results for the zero drift case (i.e., $\mu_x = \mu_y = 0$), and the lower panel presents results for a positive drift case ($\mu_x = \mu_y = 0.1$).

Distribution of t-statistics and R^2s from a regression using two independent random walks

t-statistic	$N(0,1)$	*sim.*	$\mathbf{R}^2$	χ^2_1	*sim.*
		Zero drift			
$\Pr[tstat < -2.58]$	0.005	0.40	$\Pr[R^2 \leq 0.10]$	1.00	0.38
$\Pr[tstat < -1.96]$	0.025	0.42	$\Pr[R^2 > 0.10]$	0.00	0.62
$\Pr[tstat < -1.65]$	0.05	0.44	$\Pr[R^2 > 0.20]$	0.00	0.47
$\Pr[tstat < 0]$	0.5	0.50	$\Pr[R^2 > 0.30]$	0.00	0.35
$\Pr[tstat > 1.65]$	0.05	0.44	$\Pr[R^2 > 0.50]$	0.00	0.17
$\Pr[tstat > 1.96]$	0.025	0.42	$\Pr[R^2 > 0.75]$	0.00	0.03
$\Pr[tstat > 2.58]$	0.005	0.40	$\Pr[R^2 > 0.90]$	0.00	0.0004
		Positive drift			
$\Pr[tstat < -2.58]$	0.005	0.10	$\Pr[R^2 \leq 0.10]$	1.00	0.16
$\Pr[tstat < -1.96]$	0.025	0.11	$\Pr[R^2 > 0.10]$	0.00	0.84
$\Pr[tstat < -1.65]$	0.05	0.12	$\Pr[R^2 > 0.20]$	0.00	0.75
$\Pr[tstat < 0]$	0.5	0.14	$\Pr[R^2 > 0.30]$	0.00	0.67
$\Pr[tstat > 1.65]$	0.05	0.83	$\Pr[R^2 > 0.50]$	0.00	0.50
$\Pr[tstat > 1.96]$	0.025	0.82	$\Pr[R^2 > 0.75]$	0.00	0.22
$\Pr[tstat > 2.58]$	0.005	0.81	$\Pr[R^2 > 0.90]$	0.00	0.03

When the drift is zero (top panel), we see that the t-statistic is significant at the 5% level for 84% of simulations, even though these series are independent. Further, for over

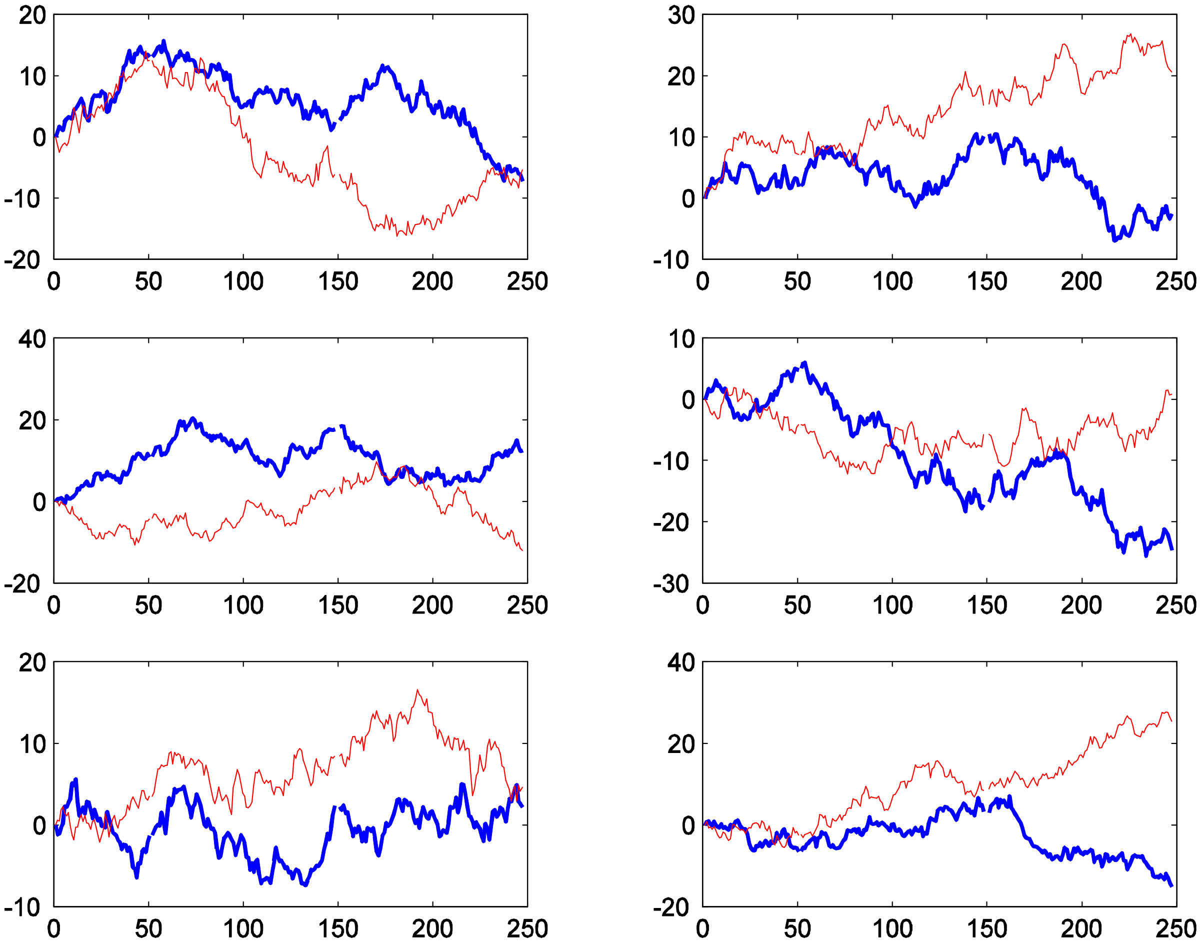


Figure 18.3: Six simulated sample paths of two independen random walks, over 250 periods.

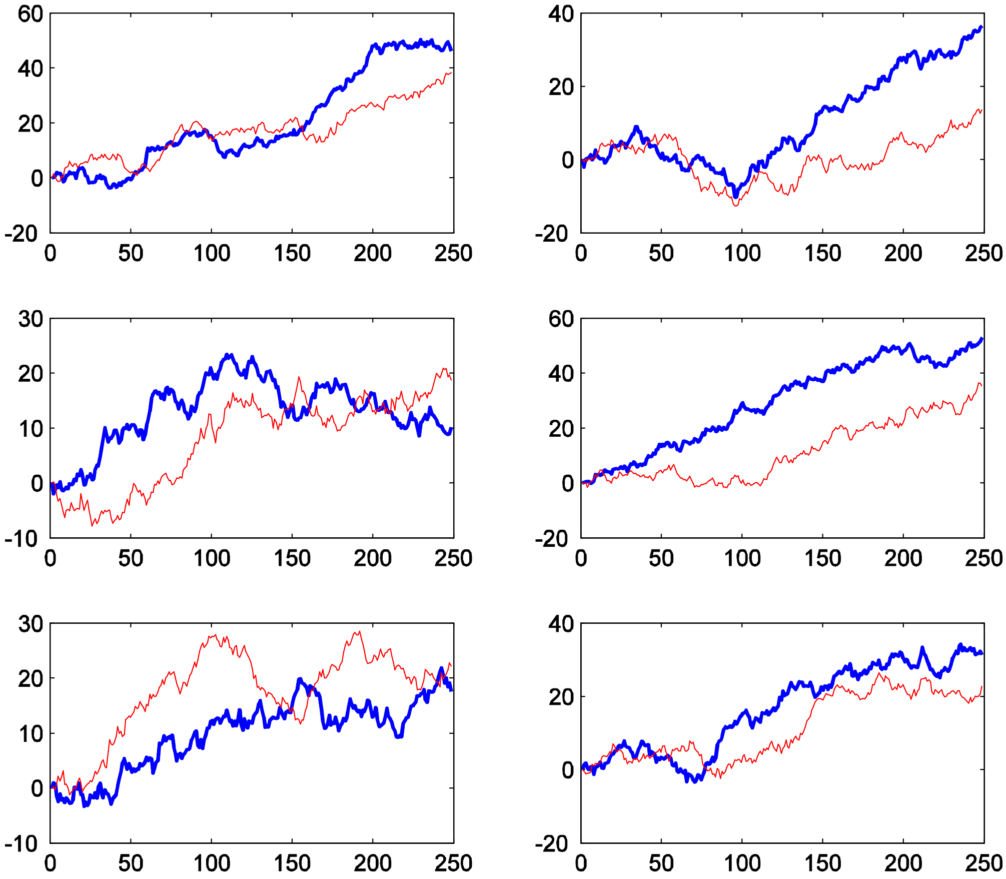


Figure 18.4: Six simulated sample paths of two independent random walks (both with drift of 0.1), over 250 periods.

60% of the simulations we see an R^2 of over 0.10, when we would expect none under the usual theory. Almost 20% of the R^2 coefficients are over 0.5. When we allow for a drift in the random walks, which is more relevant for financial time series, find even larger deviations from what we would expect using standard econometric theory: the t-statistic is significant at the 5% level for 93% of simulations. For 82% of simulations the t-statistic was greater than 1.96, indicating a high chance of finding an apparently positive relation between the series. This is driven by the fact that although the series are independent, they both have a positive drift and so they appear correlated. For over 80% of the simulations we see an R^2 of over 0.10, and 50% of the R^2 coefficients are over 0.5.

In Figure 18.5 I plot the simulated density of t-statistics and R^2 statistics for these two random walk scenarios, along with the asymptotic distributions of these statistics according to standard econometric theory.

These simulation studies show that when the two variables in a regression are random walks (or, more generally, highly persistent processes, such as AR(1) processes with an AR coefficient near 1) the usual statistics, such as t-statistics and R^2, no longer behave in the way that we usually expect. The distribution theory for these statistics needs to be adapted to deal with non-stationarity. So beware of looking at two price series and drawing conclusions about the relationship between them!

18.4 Testing for a unit root

If the series of interest Y_t has a unit root (i.e, it follows a random walk, possibly with drift), then the above simulation results indicate we need to use different econometric techniques. The simplest way of dealing with a unit root is to first-difference the series:

$$\begin{aligned} Y_{t+1} &= Y_t + \varepsilon_{t+1},\ \varepsilon_{t+1} \sim WN(0) \\ \text{then } \Delta Y_{t+1} &= Y_{t+1} - Y_t = \varepsilon_{t+1} \end{aligned}$$

and so the first-differenced series is white noise. More generally, if

$$\begin{aligned} Y_{t+1} &= Y_t + \nu_{t+1} \\ \nu_{t+1} &= \phi_0 + \phi_1 \nu_t + \theta \eta_t + \eta_{t+1},\ \eta_{t+1} \sim WN(0)\,,\ |\phi_1| < 1 \\ \text{then } \Delta Y_{t+1} &= \nu_{t+1} \sim ARMA(p,q) \end{aligned}$$

So a series with a unit root and a stationary component may be differenced to achieve stationarity. If a series is stationary after first-differencing, we say that it is ‘integrated of order 1,’ or $I(1)$. A stationary series requires no differencing, and so it is ‘integrated of order zero,’ or $I(0)$. Some series, for example consumer price indices, are thought to require second-differencing to achieve stationarity and so are called $I(2)$ series. We will focus on $I(1)$ and $I(0)$ series. Differencing a series more times than is actually required to achieve stationarity, first-differencing a series that is already $I(0)$, for example, will lead to innovations with an MA structure and will increase the variance of the series, but will *not* make the series non-stationary.

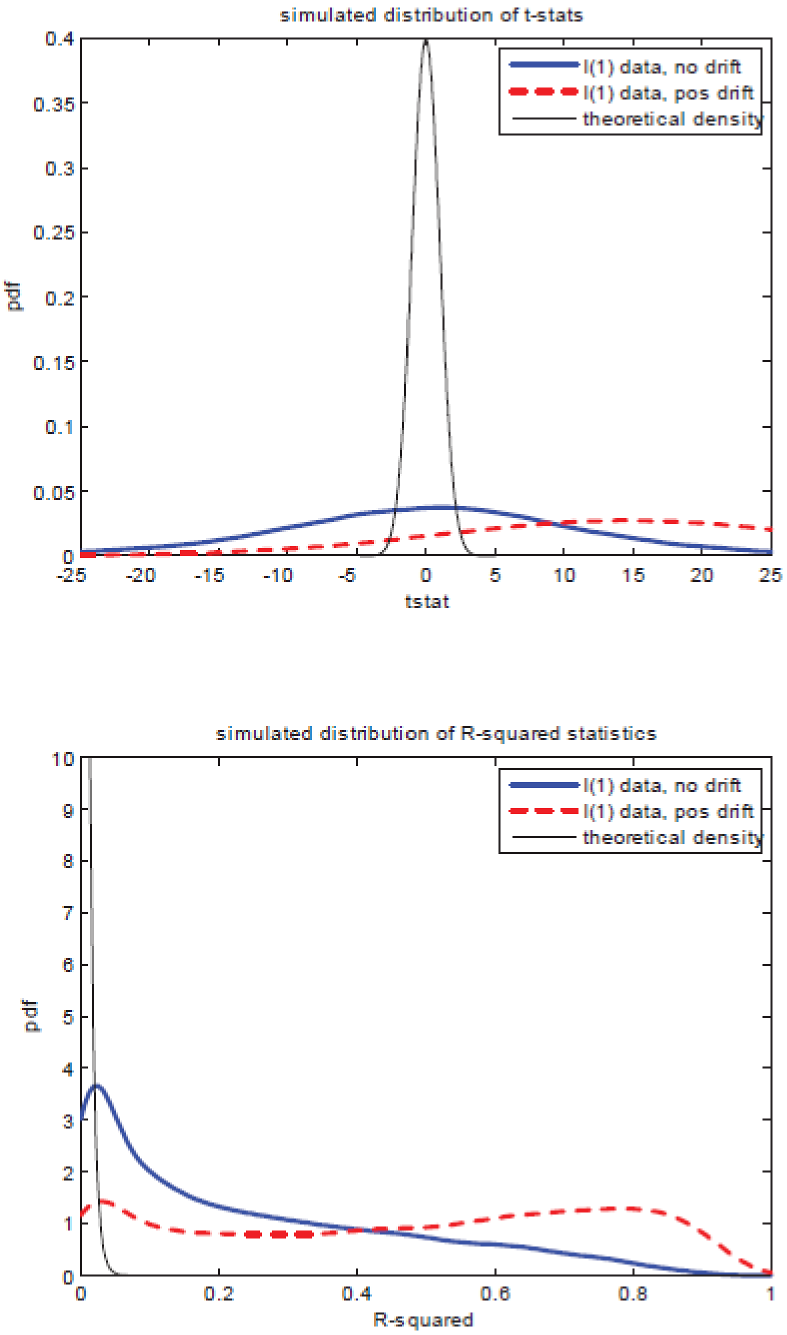


Figure 18.5: Simulated dsitribution of t-statistics and R^2 statistics for I(1) data with zero or positive drift, and the theoretical distributions for stationary data.

18.4.1 The Dickey-Fuller test

A number of tests are available for testing for the presence of a 'unit root' in a time series. Most of these test have a unit root as the null hypothesis, against a $I(0)$ alternative. So we test:

$$\begin{aligned} H_0 &: Y_t \sim I(1) \\ \text{vs.} \quad H_a &: Y_t \sim I(0) \end{aligned}$$

Notice that if the DGP is a simple AR(1), then

$$\begin{aligned} Y_t &= \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t, \;\; \varepsilon_t \sim WN(0) \\ \Delta Y_t &= \phi_0 + (\phi_1 - 1) Y_{t-1} + \varepsilon_t \\ &\equiv \phi_0 + \psi Y_{t-1} + \varepsilon_t \end{aligned}$$

and a test that $\phi = 1$ is equivalent to a test that $\psi = 0$:

$$\begin{aligned} H_0 &: \psi = 0 \\ \text{vs.} \quad H_a &: \psi < 0 \end{aligned}$$

This is a test that the series is at least $I(1)$ against the alternative that it is $I(0)$. One such test is the Dickey-Fuller test. The test statistic is the usual t-statistic on $\hat{\psi}$:

$$tstat = \frac{\hat{\psi}}{\sqrt{\hat{V}\left[\hat{\psi}\right]}}$$

but under the null hypothesis this does *not* have the Student's t or normal distribution, as the variable on the right-hand side of the regression is a random walk, see Figure 18.6. Dickey and Fuller used simulations to determine the critical values for this test statistic. When we include a constant term in the regression[1], the 95% critical value is -2.86 (compared with -1.64 critical value for a one-sided test based on the standard Normal distribution). If our test statistic is *less than* the critical value then we reject the null hypothesis of a unit root in favour of the hypothesis that the series is $I(0)$. If we fail to reject the null then we conclude that the series is $I(1)$.

The Dickey-Fuller test, however, only considers an AR(1) process. If the series has non-zero autocorrelations beyond lag one then the residuals from the regression will be autocorrelated, and thus ε_t will not be white noise. We may overcome this problem by using the Augmented Dickey-Fuller (ADF) test, which is conducted by estimating the following:

$$\Delta Y_t = \phi_0 + \psi Y_{t-1} + \alpha_1 \Delta Y_{t-1} + \alpha_2 \Delta Y_{t-2} + ... + \alpha_p \Delta Y_{t-p} + \varepsilon_t$$

and then performing the test on $\hat{\psi}$. The construction of the test statistic and the critical values do not change from the Dickey-Fuller test. By allowing more lags of ΔY_t to be used as regressors we (hopefully) remove all autocorrelation in the regression residuals. Choosing too few lags may lead to autocorrelated residuals, which can cause problems with the test. Choosing too many lags can lead to low power to reject the null hypothesis. It is a good idea to check the sensitivity of the test results to the choice of the number of lags.

[1] Unlike the usual t-test, the critical values for the Dickey-Fuller test change depending on whether a constant is included in the regression. With financial asset returns it is good practice to always include a constant term.

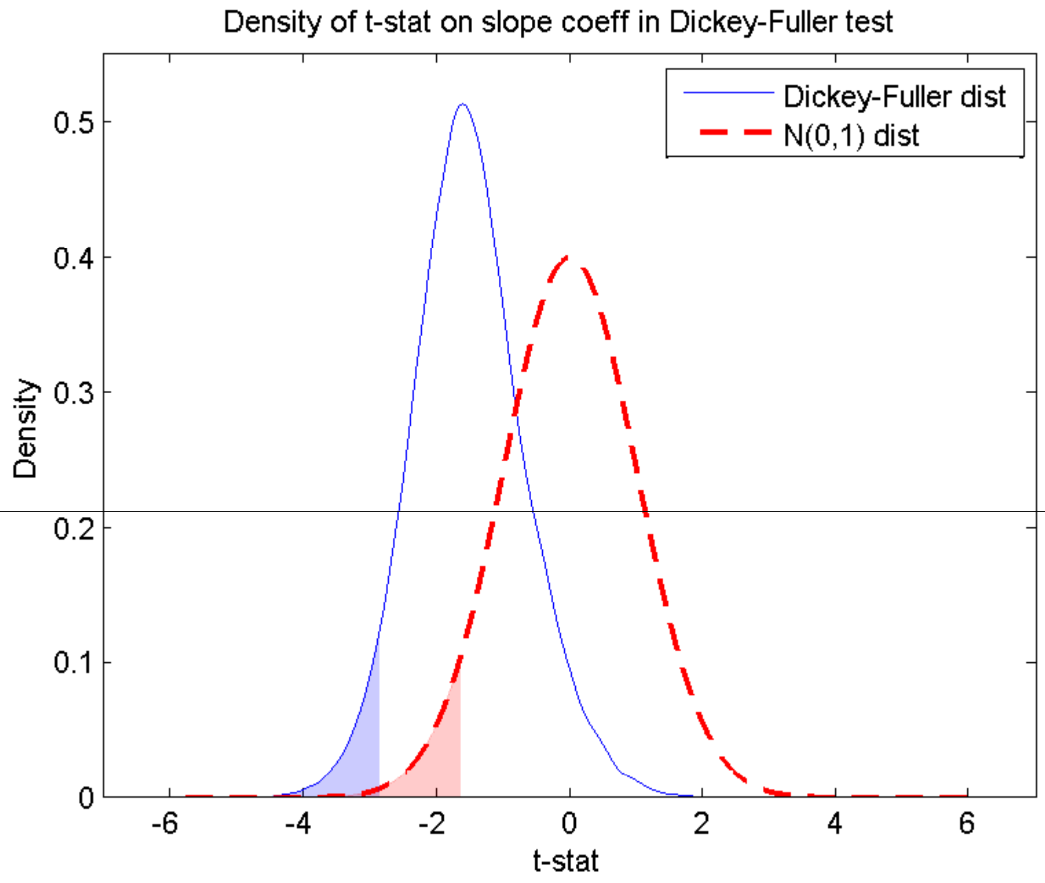


Figure 18.6: Density of t-statistic on psi in Dickey-Fuller test for a unit root, along with the standard Normal density.

18.4.2 Testing for higher-order integration

If we are unsure whether the series is $I\left(0\right)$, $I\left(1\right)$ or $I\left(2\right)$ (or more) it is necessary to conduct many tests: we first conduct the test on the level of the series, by running the regression above with the first-difference as the dependent variable. If we fail to reject the null hypothesis, then we know that the series is not $I\left(0\right)$, thus the series must be at least $I\left(1\right)$. We then conduct another test by running a regression with the *second*-difference as the dependent variable. If we again fail to reject the null then we know the series is at least $I\left(2\right)$, and we continue testing. If we *do* reject the null hypothesis then we conclude that the series is $I\left(1\right)$, and so the first-difference of the series will be $I\left(0\right)$.

Most financial asset prices have been found to be $I\left(1\right)$, and price indices are usually found to be $I\left(1\right)$ or $I\left(2\right)$. Integration of order higher than 2 is rare in economics and finance.

> **Activity 18.2** Consider the following table of results for two financial time series: the GBP/USD exchange rate and the FTSE 100 index. Both series are measured daily over the period 2004-2013. Interpret the results in this table. (Recall that the 95% critical value for the Dickey-Fuller test is -2.86.)

	GBP/USD	**FTSE 100**
$\text{Corr}[\log P_t, \log P_{t-1}]$	0.9979	0.9944
$\text{Corr}[\log P_t, \log P_{t-2}]$	0.9957	0.9891
$\text{Corr}[\log P_t, \log P_{t-3}]$	0.9935	0.9843
Augmented Dickey-Fuller statistic with 0 lags	-1.6077	-2.2359
Augmented Dickey-Fuller statistic with 1 lags	-1.6160	-2.1293
Augmented Dickey-Fuller statistic with 5 lags	-1.5374	-2.8641

18.5 Overview of chapter

This chapter considered models and methods to study persistent time series, that is, time series with autocorrelations that are close to one. We consider a *time trend* model and a *random walk* (or *unit root*) model for this purpose. We discussed the problem of spurious regressions, which arises when one or both variables included in a regression are persistent, and we looked at the Dickey-Fuller test for the presence of a unit root in a time series.

18.6 Reminder of learning outcomes

Having completed this chapter, and the essential reading and activities, you should be able to:

- Describe the differences between the time trend model and the random walk model for persistent time series
- Discuss the problem of 'spurious regressions' and know when that problem might arise
- Implement and interpret the Dickey-Fuller test for a unit root

18.7 Test your knowledge and understanding

1. Consider the following model for a stock price:

$$Y_t = \beta + Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN\left(0, \sigma^2\right)$$

   (a) Find $E_t[Y_{t+1}]$ and $E_t[Y_{t+2}]$. Conjecture a formula for $E_t[Y_{t+h}]$ for any $h > 0$.

   (b) Find $V_t[Y_{t+1}]$ and $V_t[Y_{t+2}]$. Conjecture a formula for $V_t[Y_{t+h}]$ for any $h > 0$.

   (c) Find $Cov_t[Y_{t+1}, Y_{t+2}]$.

2. The sample autocorrelation function presented in Figure 18.7 was obtained using the log of the DAX index of 30 German equities, measured daily over the period 2004 to 2013. The augmented Dickey Fuller test statistic for this series is -1.1704. Interpret these results.

3. The "Gibson paradox" refers to the apparently strong correlation between the (log) price level in a country and the level of interest rates. This paradox was noted as far back as Keynes (1925), who wrote that the "tendency of prices and interest rates to rise together ... and to fall together ... is one of the most completely established facts within the whole of quantitative economics". An example of this is given using U.S. data on the consumer price index (CPI) and the 1-year Treasury bill rate over the period 1953 to 2005. See Figure 18.8 for a plot of these two series, and see the table below for some summary statistics.

	log(CPI)	T-bill
mean	4.25	5.71
median	4.30	5.44
maximum	5.29	16.72
minimum	3.28	0.82
std dev	0.71	3.00
skewness	-0.03	0.99
kurtosis	1.38	4.22
autocorrelation lag 1	0.992	0.987
autocorrelation lag 2	0.986	0.968
correlation	0.213	

   (a) Why should we be cautious interpreting the results of regressions involving very persistent processes like log prices and the level of interest rates? Explain your answer.

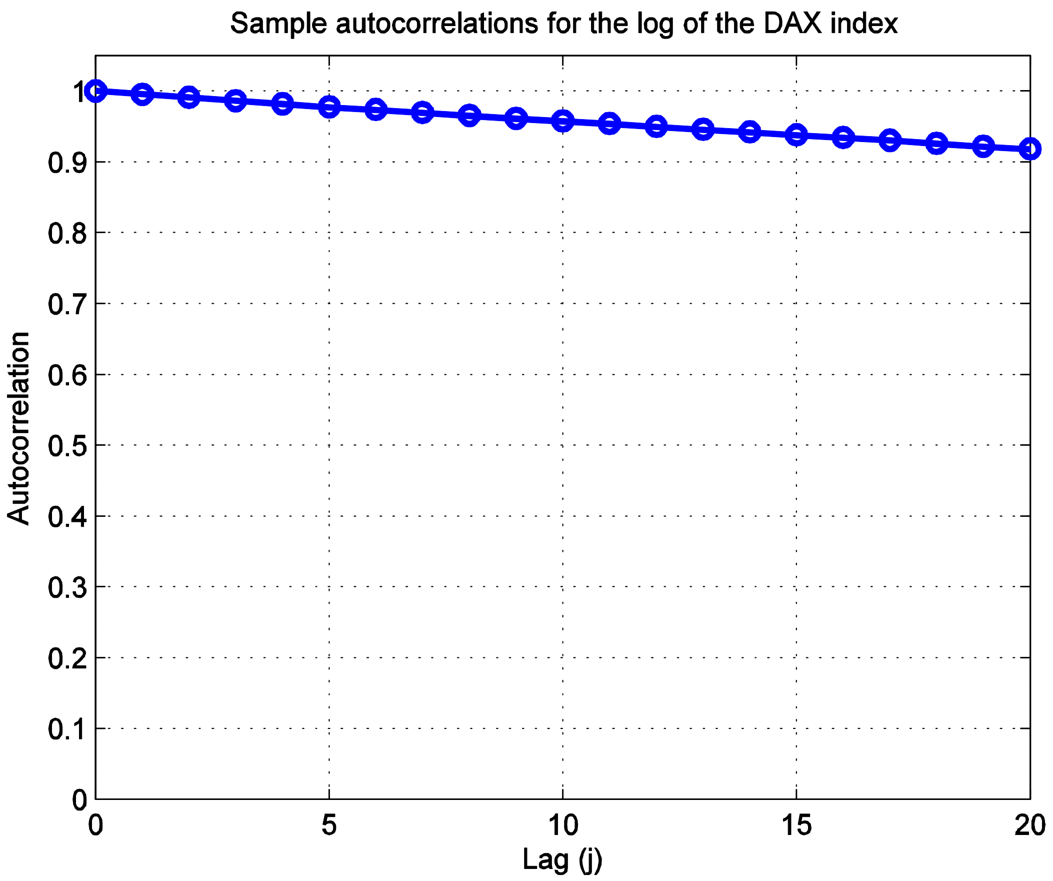


Figure 18.7: Sample autocorrelations for daily log-levels of the DAX equity index.

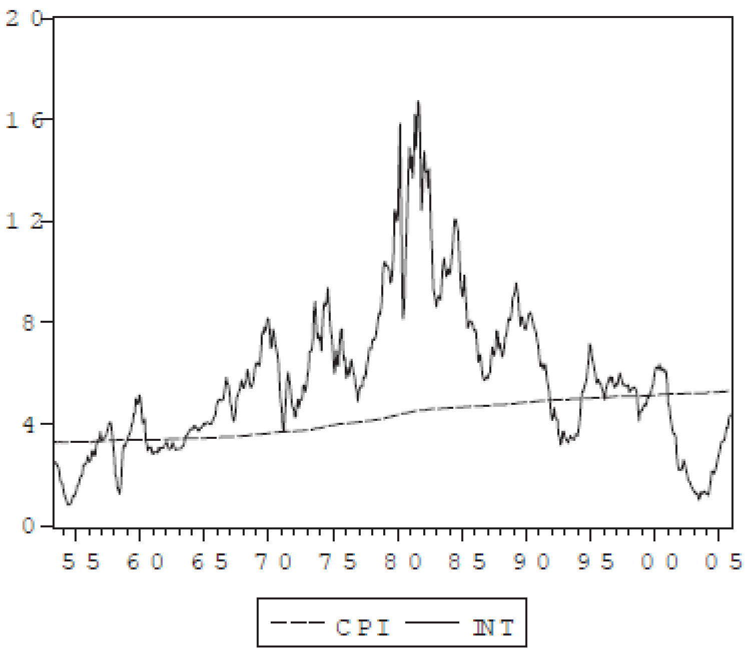


Figure 18.8: Log prices (CPI) and 1-year T-bill rates (INT) over the period 1953 to 2005.

Next consider an AR(1) process:

$$\begin{array}{rcl} Y_t & = & \phi Y_{t-1} + \varepsilon_t, \quad \text{for } t = 1, 2, ... \\ \text{where } \ \varepsilon_t & \sim & WN\left(0, \sigma^2\right) \\ \text{and } \ Y_s & = & 0 \ \ \text{for } s \leq 0 \end{array}$$

(b) Show that Y_t can be written as a (weighted) sum of ε_t, ε_{t-1}, ..., ε_1.

(c) Discuss how past shocks (ε_{t-j}) affect the current value of the series (Y_t) as a function of the size of ϕ.

4. Don't forget to check the VLE for additional practice problems for this chapter.

18.8 Solutions to activities

Activity 18.1

Given the two models:

$$\begin{array}{rcl} X_{t+1} & = & \beta_0 + \beta_1 t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim iid\ N\left(0, \sigma^2\right) \\ Y_{t+1} & = & \phi_0 + Y_t + \eta_{t+1}, \quad \eta_{t+1} \sim iid\ N\left(0, \sigma^2\right) \end{array}$$

First we find the conditional means:

$$\begin{array}{rcl} E_t\left[X_{t+3}\right] & = & E_t\left[\beta_0 + \beta_1\left(t+3\right) + \varepsilon_{t+3}\right] \\ & = & \beta_0 + \beta_1\left(t+3\right), \ \text{ since } E_t\left[\varepsilon_{t+3}\right] = 0 \\ E_t\left[Y_{t+3}\right] & = & E_t\left[\phi_0 + Y_{t+2} + \eta_{t+3}\right] \\ & = & E_t\left[\phi_0 + \phi_0 + Y_{t+1} + \eta_{t+2} + \eta_{t+3}\right] \\ & = & E_t\left[\phi_0 + \phi_0 + \phi_0 + Y_t + \eta_{t+1} + \eta_{t+2} + \eta_{t+3}\right] \\ & = & 3\phi_0 + Y_t, \ \text{ since } E_t\left[\eta_{t+j}\right] = 0 \ \forall j > 0 \end{array}$$

Then the conditional variances:

$$\begin{array}{rcl} V_t\left[X_{t+3}\right] & = & V_t\left[\beta_0 + \beta_1\left(t+3\right) + \varepsilon_{t+3}\right] \\ & = & V_t\left[\varepsilon_{t+3}\right], \ \text{ since only } \varepsilon_{t+3} \text{ is random} \\ & = & \sigma^2 \\ V_t\left[Y_{t+3}\right] & = & V_t\left[\phi_0 + \phi_0 + \phi_0 + Y_t + \eta_{t+1} + \eta_{t+2} + \eta_{t+3}\right] \\ & = & V_t\left[\eta_{t+1} + \eta_{t+2} + \eta_{t+3}\right], \ \text{ since } Y_t \text{ is known at time } t \\ & = & 3\sigma^2, \ \text{ since } \eta_{t+1} \sim iid\ N\left(0, \sigma^2\right) \end{array}$$

Activity 18.2

The upper panel of this table shows that both assets are very persistent: the first three autocorrelations are close to one for both assets. This is a first sign that these processes may have a 'unit root.' The lower panel formally tests for a unit root, using the Augmented Dickey Fuller test. Three test statistics are reported, corresponding to the inclusion of zero, one or five lags of the returns. The null and alternative hypotheses in a Dickey-Fuller test are

$$\begin{array}{rrcll} & H_0 & : & \psi = 0 & \Leftrightarrow H_0 : Y_t \sim I\,(1) \\ \text{vs} & H_a & : & \psi < 0 & \Leftrightarrow H_a : Y_t \sim I\,(0) \end{array}$$

For the British pound/US dollar exchange rate, we see that all three test statistics are greater than the critical value of -2.86. This means that we fail to reject the null in all three cases, and we conclude that the (log of the) GBP/USD exchange rate follows a random walk.

For the FTSE 100 stock index we see that the first two ADF test statistics are greater than -2.86, and so we again conclude that the (log of the) FTSE index follows a random walk. However the third test statistic, when five lagged returns are included, is just below -2.86, and so we would reject the null and conclude that the index is in fact stationary. Given that the other two test statistics fail to reject the null, and this rejection is very close to the border, a reasonable overall conclusion from these results is that the FTSE index follows a random walk.

Appendix A
Sample examination paper

Important note: This Sample examination paper reflects the examination and assessment arrangements for this course in the academic year 2014–2015. The format and structure of the examination may have changed since the publication of this subject guide. You can find the most recent examination papers on the VLE where all changes to the format of the examination are posted.

Time allowed: three hours.

Candidates should answer **THREE** of the following **FOUR** questions. All questions carry equal marks.

A calculator may be used when answering questions on this paper and it must comply in all respects with the specification given in paragraph 10.6 of the *General Regulations*.

1. Consider the following time series process:

$$Y_t = \alpha + \varepsilon_t + \theta\varepsilon_{t-1}, \quad \varepsilon_t \sim WN\left(0, \sigma^2\right)$$

   (a) Find $E\left[Y_t\right]$

   (b) Find $V\left[Y_t\right]$

   (c) Find $Cov\left[Y_t, Y_{t-1}\right]$

   (d) Find $Cov\left[Y_t, Y_{t-2}\right]$

   (e) Find $E_t\left[Y_{t+1}\right]$

2. (a) Describe the GARCH(1,1) model and how it is used in finance.

   (b) Briefly describe one extension of the GARCH model and what it is designed to capture.

   (c) Assume that

$$\begin{aligned} Y_{t+1} &= \phi_0 + \phi_1 Y_t + \varepsilon_{t+1} \\ \varepsilon_{t+1} &\sim iid\ N\left(0, \sigma^2\right) \end{aligned}$$

   Find the first-order autocorrelation of ε_{t+1}^2, and interpret.

   (d) If we instead assume an AR(1) for ε_{t+1}^2, what would we obtain for the conditional variance of Y_{t+1} from part (c)? Why would we employ such a model?

3. (a) What is Value-at-Risk? What are its pros and cons relative to variance as a measure of risk?

   (b) Describe the 'historical simulation' and RiskMetrics approaches to measuring Value-at-Risk.

   (c) Compare and contrast two different tests to evaluate Value-at-Risk forecasts.

4. Consider the problem of choosing between Model A and Model B as forecasting models, with the sample size $n = 150$. You have the following information:

$$\begin{array}{ll}
\frac{1}{n}\sum_{t=1}^{n} Y_t = \bar{Y} = 1.4, & \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \bar{Y}\right)^2 = 5 \\
\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t^a\right)^2 = 4.8, & \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t^b\right)^2 = 4.9 \\
\frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t^a\right| = 4.2, & \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t^b\right| = 3.9 \\
V\left[\left(Y_t - \hat{Y}_t^a\right)^2 - \left(Y_t - \hat{Y}_t^b\right)^2\right] = 3, & V\left[\left|Y_t - \hat{Y}_t^a\right| - \left|Y_t - \hat{Y}_t^b\right|\right] = 2.2
\end{array}$$

How can we test that model A is 'better than' model B? What does/do the test(s) tell us? (You can assume that there is no serial dependence in the data if that helps.)

END OF PAPER

Appendix B
Guidance on answering the Sample examination questions

Below I provide outlines of the answers to the sample examination paper. You are required to answer any four questions from the examination paper and each question carries equal marks. You are advised to divide your time accordingly.

Question 1

(a)

$$\begin{aligned}
E[Y_t] &= E[\alpha + \varepsilon_t + \theta\varepsilon_{t-1}] \\
&= \alpha + E[\varepsilon_t] + \theta E[\varepsilon_{t-1}] \\
&= \alpha, \text{ since } E[\varepsilon_t] = E[\varepsilon_{t-1}] = 0 \text{ as } \varepsilon_t \sim WN\left(0, \sigma^2\right)
\end{aligned}$$

(b)

$$\begin{aligned}
V[Y_t] &= V[\alpha + \varepsilon_t + \theta\varepsilon_{t-1}] \\
&= V[\varepsilon_t + \theta\varepsilon_{t-1}], \text{ since } \alpha \text{ is a constant} \\
&= V[\varepsilon_t] + \theta^2 V[\varepsilon_{t-1}] + 2\theta Cov[\varepsilon_t, \varepsilon_{t-1}] \\
&= \sigma^2\left(1 + \theta^2\right), \text{ since } Cov[\varepsilon_t, \varepsilon_{t-1}] = 0 \text{ and } V[\varepsilon_t] = V[\varepsilon_{t-1}] = \sigma^2 \text{ as } \varepsilon_t \sim WN\left(0, \sigma^2\right)
\end{aligned}$$

(c) Here you need to remember that $Cov[\varepsilon_{t-1}, \varepsilon_{t-1}] = V[\varepsilon_{t-1}] = \sigma^2$. This is the only term that is non-zero when you expand the covariance:

$$\begin{aligned}
Cov[Y_t, Y_{t-1}] &= Cov[\alpha + \varepsilon_t + \theta\varepsilon_{t-1}, \alpha + \varepsilon_{t-1} + \theta\varepsilon_{t-2}] \\
&= Cov[\varepsilon_t + \theta\varepsilon_{t-1}, \varepsilon_{t-1} + \theta\varepsilon_{t-2}], \text{ since } \alpha \text{ is a constant} \\
&= Cov[\varepsilon_t, \varepsilon_{t-1}] + \theta Cov[\varepsilon_t, \varepsilon_{t-2}] + \theta V[\varepsilon_{t-1}] + \theta^2 Cov[\varepsilon_{t-1}, \varepsilon_{t-2}] \\
&= \theta\sigma^2, \text{ since } Cov[\varepsilon_t, \varepsilon_{t-j}] = 0 \ \forall\ j \neq 0 \text{ and } V[\varepsilon_t] = \sigma^2 \text{ as } \varepsilon_t \sim WN\left(0, \sigma^2\right)
\end{aligned}$$

(d)

$$\begin{aligned}
Cov[Y_t, Y_{t-2}] &= Cov[\alpha + \varepsilon_t + \theta\varepsilon_{t-1}, \alpha + \varepsilon_{t-2} + \theta\varepsilon_{t-3}] \\
&= Cov[\varepsilon_t + \theta\varepsilon_{t-1}, \varepsilon_{t-2} + \theta\varepsilon_{t-3}], \text{ since } \alpha \text{ is a constant} \\
&= Cov[\varepsilon_t, \varepsilon_{t-2}] + \theta Cov[\varepsilon_t, \varepsilon_{t-3}] + \theta Cov[\varepsilon_{t-1}, \varepsilon_{t-2}] + \theta^2 Cov[\varepsilon_{t-1}, \varepsilon_{t-3}] \\
&= 0, \text{ since } Cov[\varepsilon_t, \varepsilon_{t-j}] = 0 \ \forall\ j \neq 0 \text{ as } \varepsilon_t \sim WN\left(0, \sigma^2\right)
\end{aligned}$$

(e) Here you need to remember that ε_t is *known* at time t. Unconditionally, ε_t has mean zero, but conditional on time t information, its expecation is just itself.

$$\begin{aligned} E_t\left[Y_{t+1}\right] &= E_t\left[\alpha + \varepsilon_{t+1} + \theta\varepsilon_t\right] \\ &= \alpha + E_t\left[\varepsilon_{t+1}\right] + \theta\varepsilon_t, \text{ since } \varepsilon_t \text{ is known at time } t \\ &= \alpha + \theta\varepsilon_t, \text{ since } E_t\left[\varepsilon_{t+1}\right] = 0 \end{aligned}$$

Question 2

(a) A statement of the GARCH(1,1) model, direct from the notes, and some discussion about the fact that it captures time-varying volatility (or volatility clustering) which is useful for predicting future volatility. It is based on an ARMA(1,1) model for the squared residual.

(b) Could discuss: Glosten-Jagannathan-Runkle (GJR) GARCH model, which allows for a 'leverage effect' in volatility; the EGARCH model of Nelson (1991) which allows for a leverage effect in a different way, and also ensures that the predicted volatility is positive; or the ARCH-in-mean model, which allows for the level of volatility to affect the expected return on the asset under analysis.

(c) Since $\varepsilon_{t+1} \sim iid$, as given in the question, we know that any function of ε_{t+1} is also iid. Thus $\varepsilon_{t+1}^2 \sim iid$ which implies $Corr\left[\varepsilon_{t+1}^2, \varepsilon_t^2\right] = 0$. This implies that assuming an iid innovation for the AR(1) model for Y_{t+1} rules out any volatility clustering in the innovation series ε_{t+1}.

(d) An AR(1) model for ε_{t+1}^2 leads to the ARCH(1) model for the conditional variance of Y_{t+1}. This is shown in detail in the notes. This model is useful if we find autocorrelation in the squared residuals from the AR(1) model for Y_{t+1}, and/or if we think there is volatility clustering in Y_{t+1}.

Question 3

(a) Here the ideal answer would give both the mathematical definition of VaR:

$$\Pr\left[r_{t+1} \leq VaR_{t+1}^{\alpha} | \mathcal{F}_t\right] = \alpha$$

and the verbal definition: 'The $\alpha\%$ VaR is the cut-off such that there is only an $\alpha\%$ probability that we will see a return as low or lower, as well as some discussion of common choices for α (1%, 5% and 10%), etc.

Pros: From the notes: 'Variance as a measure of risk has the drawback that it 'penalises' (by taking a higher value) large positive returns in the same way as large negative returns. As investors, however, we would generally only consider 'risk' to be the possibility of the value of our portfolio falling, not rising. VaR overcomes this by focussing on the lower tail of the distribution of returns only.'

Cons: the main drawback relative to variance is simply that it is less familiar to market participants. Any sensible discussion of alternative pros and cons would also make a good answer.

(b) Here the Examiners would be looking for a brief summary of these methods, with some discussion of their relative pros and cons. Historical simulation is simple and does not require a model, but it does require the assumption that returns are iid.

RiskMetrics allows for volatility clustering, which is good, but usually assumes Normality which is a drawback relative to historical simulation.

(c) Could compare two ‘unconditional’ tests (the simple t-test on the number of hits, equivalent to the ‘violation ratio’, or the unconditional test of Christoffersen) or two conditional tests (the regression-based test, or the conditional test of Christoffersen). Details on these are in the notes.

Question 4

We can use a Diebold-Mariano (DM) test to compare the models. We have sufficient information here to run 2 DM tests: one using MSE loss and the other using MAE loss.

$$\begin{aligned}
\text{Let } d_t &= L\left(Y_t, \hat{Y}_t^a\right) - L\left(Y_t, \hat{Y}_t^b\right) \\
\bar{d} &\equiv \frac{1}{n}\sum_{t=1}^{n} d_t \\
\text{DM statistic} &= \frac{\sqrt{n}\bar{d}}{V\left[d_t\right]} \\
\text{MSE } \ DM &= \frac{\sqrt{150}\,(4.8-4.9)}{\sqrt{3}} = -0.7071 \\
\text{MAE } \ DM &= \frac{\sqrt{150}\,(4.2-3.9)}{\sqrt{2.2}} = 2.4772
\end{aligned}$$

Under MSE loss forecast A is better than forecast B, but the difference is not significant (the DM statistic is less than 1.96 in absolute value). Thus under MSE loss we would conclude that there is no evidence against the null that the forecasts are equally accurate.

Under MAE the ranking is reversed, and the difference is significant. Thus we conclude that under MAE foreast B is significantly better than forecast A.

Thus depending on the preferences of the forecast user, forecast A is weakly preferred to forecast B, or forecast B is significantly preferred to forecast A.

Comment form

We welcome any comments you may have on the materials which are sent to you as part of your study pack. Such feedback from students helps us in our effort to improve the materials produced for the International Programmes.
If you have any comments about this guide, either general or specific (including corrections, non-availability of Essential readings, etc.), please take the time to complete and return this form.

Title of this subject guide:

Name

Address

Email

Student number

For which qualification are you studying?

Comments

Please continue on additional sheets if necessary.

Date:

Please send your completed form (or a photocopy of it) to:
Publishing Manager, Publications Office, University of London International Programmes,
Stewart House, 32 Russell Square, London WC1B 5DN, UK.