1. INTRODUCTION

Fintech (finance + technology) is playing a major role in the advancement and improvement of:

- **investment management industry** (such as assessment of investment opportunities, portfolio optimization, risk mitigation etc.).
- **investment advisory services** (e.g. Robo-advisors with or without intervention of human advisors are providing tailored, low-priced, actionable advice to investors).
- **financial record keeping, blockchain and distributed ledger technology (DLT**) through finding improved ways of recording, tracking or storing financial assets.

2. WHAT IS FINTECH

For the scope of this reading, term **'Fintech'** is referred to as technology-driven innovations in the field of financial services and products.

Note: In common usage, fintech may also refer to companies associated with new technologies or innovations.

Initially, the scope of fintech was limited to data processing and to the automation of routine tasks. Today, advanced computer systems are using artificial intelligence and machine learning to perform decision-making tasks including investment advice, financial planning, business lending/payments etc.

Some salient fintech developments related to the investment industry include:

- **Analysis of large data sets:** These days, professional investment decision making process uses extensive amounts of traditional data sources (e.g. economic indicators, financial statements) as well as non-traditional data sources (such as social media, sensor networks) to generate profits.
- **Analytical tools:** There is a growing need of techniques involving artificial intelligence (AI) to identify complex, non-linear relationships among such gigantic datasets.
- **Automated trading:** Automated trading advantages include lower transaction costs, market liquidity, secrecy, efficient trading etc.
- **Automated advice:** Robo-advisors or automated personal wealth management are low-cost alternates for retail investors.
- **Financial record keeping:** DLT (distributed ledger technology) provides advanced and secure means of record keeping and tracing ownership of financial assets on peer-to-peer (P2P) basis. P2P lowers involvement of financial intermediaries.

3. BIG DATA

Big data refers to huge amount of data generated by traditional and non-traditional data sources.

Details of traditional and non-traditional sources are given in the table below.

Traditional		Non-traditional (alternate)	
Sources	Institutions, Businesses, Government, Financial Markets	Sources	Social media, Sensor networks Company-used data, Electronic devices, Smart phones, Cameras, Microphones, Radio-frequency identification (RFID)
Forms of Data	Annual reports, Regulatory filings, Sales & earnings, Conference calls, Trade prices & volumes	Forms of Data	Posts, Tweets, Blogs, Email, Text messages, Web-traffic, Online news sites

Big data typically have the following features:

- Volume
- Velocity
- Variety

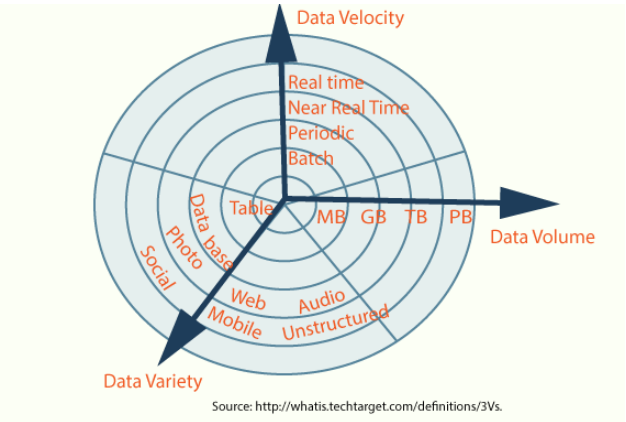
Volume: Quantities of data denoted in millions, or even billions, of data points. Exhibit below shows data grow from MB to GB to larger sizes such as TB and PB.

Velocity: Velocity determines how fast the data is communicated. Two criteria are Real-time or Near-time data, based on time delay.

Variety: Data is collected in a variety of forms including:

- **structured data** – data items are often arranged in tables where each field represent a similar type of information. (e.g. SQL tables, CSV files)
- **unstructured data** – cannot be organized in table and requires special applications or programs (e.g. social media, email, text messages, pictures, sensors, video/voice messages)
- **semi-structured data** – contains attributes of both structured and unstructured data (e.g. HTML codes)

Exhibit: Big Data Characteristics: Volume, Velocity & Variety



3.1 Sources of Big Data

In addition to traditional data sources, alternative data sources are providing further information (regarding consumer behaviors, companies' performances and other important investment-related activities) to be used in investment decision-making processes.

Main sources of alternative data are data generated by:

1. **Individuals:** Data in the form of text, video, photo, audio or other online activities (customer reviews, e-commerce). This type of data is often unstructured and is growing considerably.

2. **Business processes:** data (often structured) generated by corporations or other public entities e.g. sales information, corporate exhaust. Corporate exhaust includes bank records, point of sale, supply chain information.

Note:

- Traditional corporate metrics (annual, quarterly reports) are lagging indicators of business performance.
- Business process data are real-time or leading indicators of business performance.

3. **Sensors:** data (often unstructured) connected to devices via wireless networks. The volume of such data is growing exponentially compared to other two sources. **IoT** (internet of things) is the network of physical devices, home appliances, smart buildings that enable objects to share or interact information.

Alternative datasets are now used increasingly in the investment decision making models. Investment professionals will have to be vigilant about using information, which is not in the public domain regarding individuals without their explicit knowledge or consent.

3.2 Big Data Challenges

In investment analysis, using big data is challenging in terms of its quality (selection bias, missing data, outliers), volume (data sufficiency) and suitability. Most of the times, data is required to be sourced, cleansed and organized before use, however, performing these processes with alternative data is extremely challenging due to the qualitative nature of the data. Therefore, artificial intelligence and machine learning tools help addressing such issues.

4. ADVANCED ANALYTICAL TOOLS: ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Artificial intelligence (AI) technology in computer systems is used to perform tasks that involve cognitive and decision-making ability similar or superior to human brains.

Initially, AI programs were used in specific problem-solving framework following 'if-then' rules. Later, advanced processors enabled AI programs such as neural networks (which are based on how human brains process information) to be used in financial analysis, data mining, logistics etc.

Machine learning (ML) algorithms are computer programs that perform tasks and improve their performance overtime with experience. ML requires large amount of data (big data) to model accurate relationships.

ML algorithms use inputs (set of variables or datasets), learn from data by identifying relationships in the data to refine the process and model outputs (targets). If no targets are given, algorithms are used to describe the underlying structure of the data.

ML divides data into two sets:

- **Training data:** that helps ML to identify relationships between inputs and outputs through historical patterns.
- **Validation data**: that validates the performance of the model by testing the relationships developed (using the training data).

ML still depends on human judgment to develop suitable techniques for data analysis. ML works on sufficiently large amount of data which is clean, authentic and is free from biases.

The problem of overfitting (too complex model) occurs when algorithm models the training data too precisely. Over-trained model treats noise as true parameters. Such models fail to predict outcomes with out-of-sample data.

The problem of underfitting (too simple model) occurs when models treat true parameters as noise and fail to recognize relationships within the training data.

Sometimes results of ML algorithms are unclear and are not comprehensible i.e. when ML techniques are not explicitly programmed, they may appear to be opaque or 'black box'.

4.1 Types of Machine Learning

ML approaches are used to identify relationships between variables, detect patterns or structure data. Two main types of machine learning are:

1. **Supervised leaning:** uses labeled training data (set of inputs supplied to the program), and process that information to find the output. Supervised learning follows the logic of 'X leads to Y'. Supervised learning is used to forecast a stock's future returns or to predict stock market performance for next business day.

2. **Unsupervised learning:** does not make use of labelled training data and does not follow the logic of 'X leads to Y'. There are no outcomes to match to, however, the input data is analyzed, and the program discovers structures within the data itself e.g. splitting data into groups based on some similar attributes.

Deep Learning Nets (DLNs): Some approaches use both supervised and unsupervised ML techniques. For example, deep learning nets (DLNs) use neural networks often with many hidden layers to perform non-linear data processing such as image, pattern or speech recognition, forecasting etc.

There is a significant role of advanced ML techniques in the evolution of investment research. ML techniques make it possible to

- render greater data availability
- analyze big data
- improve software processing speeds
- reduce storage costs

As a result, ML techniques are providing insights into individual firms, national or global levels and are a great help in predicting trends or events. Image recognition algorithms are used in store parking lots, shipping/manufacturing activities, agriculture fields etc.

5. DATA SCIENCE: EXTRACTING INFORMATION FROM BIG DATA

Data science is interdisciplinary area that uses scientific methods (ML, statistics, algorithms, computer-techniques) to obtain information from big data or data in general.

The unstructured nature of the big data requires some specialized treatments (performed by data scientist) before using that data for analysis purpose.

5.1 Data Processing Methods

Various data processing methods are used by scientists to prepare and manage big data for further examination. Five data processing methods are given below:

Capture: Data capture refers to how data is collected and formatted for further analysis. **Low-latency** systems are systems that communicate high data volumes with small delay times such as applications based on real-time prices and events. **High-latency** systems suffers from long delays and do not require access to real-time data and calculations.

Curation: Data curation refers to managing and cleaning data to ensure data quality. This process involves detecting data errors and adjusting for missing data.

Storage: Data storage refers to archiving and storing data. Different types of data (structured, unstructured) require different storage formats

Search: Search refers to how to locate requested data. Advanced applications are required to search from big data.

Transfer: Data transfer refers to how to move data from its storage location to the underlying analytical tool. Data retrieved from stock exchange's price feed is an example of direct data feed.

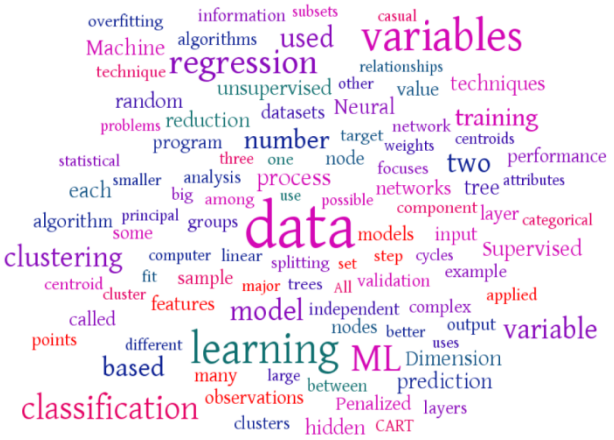
5.2 Data Visualization

Data visualization refers to how data will be formatted and displayed visually in graphical format.

Data visualization for

- **traditional structured** data can be done using tables, charts and trends.
- **non-traditional unstructured** data can be achieved using new visualization techniques such as:
  - interactive 3D graphics
  - multidimensional (more than three dimensional) data requires additional visualization techniques using colors, shapes, sizes etc.
  - tag cloud, where words are sized and displayed based on their frequency in the file.
  - Mind map, a variation of tag cloud, which shows how different concepts are related to each other.

Data visualization Tag Cloud Example



Source: https://worditout.com/word-cloud/create

6. SELECTED APPLICATIONS OF FINTECH TO INVESTMENT MANAGEMENT

6.1 Text Analytics and Natural Language Processing

Text analytics is a use of computer programs to retrieve and analyze information from large unstructured text or voice-based data sources (reports, earning calls, internet postings, email, surveys). Text analytics helps in investment decision making. Other analytical usage includes lexical analysis (first phrase of compiler) or analyzing key words or phrases based on word frequency in a document.

Natural language processing (NLP) is a field of research that focuses on development of computer programs to interpret human language. NLP field exists at the intersection of computer science, AI, and linguistics.

NLP functions include translation, speech recognition, sentiment analysis, topic analysis. Some NLP compliance related applications include reviewing electronic communications, inappropriate conduct, fraud detection, retaining confidential information etc.

With the help of ML algorithms, NLP can evaluate persons' speech – preferences, tones, likes, dislikes – to predict trends, short-term indicators, future performance of a company, stock, market or economic events in shorter timespans and with greater accuracy.

For example, NLP can help analyze subtleties in communications and transcripts from policy makers (e.g. U.S Fed, European central bank) through the choice of topics, words, voice tones.

Similarly, in investment decision making, NLP may be used to monitor financial analysts' *commentary* regarding EPS forecasts to detect shifts in sentiments (which can be easily missed in their written reports). NLP then assign sentiment ratings ranging from negative to positive, potentially ahead of a change in their recommendations.

Note: Analysts do not change their buy, hold and sell recommendations frequently; instead they may offer nuanced commentary reflecting their views on a company's near-term forecasts.

6.2 Robo-Advisory Services

Robo-advisory services provide online programs for investment solutions without direct interaction with financial advisors.

Robo-advisors just like other investment professionals are regulated by similar level of scrutiny and code of conduct. In U.S, Robo-advisors are regulated by the SEC. In U.K., they are regulated by Financial conduct authority. Robo advisors are also gaining popularity in Asia and other parts of the world.

How Robo-advisors work:
First, a client digitally enters his assets, liabilities, risk preferences, target investment returns in an investor questionnaire. Then the robo-adviser software composes recommendations based on algorithmic rules, the client's stated parameters and historical market data. Further research may be necessary overtime to evaluate the robo-advisor's performance.

Currently, robo-advisors are offering services in the area of automated asset allocation, trade execution, portfolio optimization, tax-loss harvesting, portfolio rebalancing.

Though robo-advisors cover both active and passive management styles, however, most robo-advisors follow a passive investment approach e.g. low cost, diversified index mutual funds or ETFs. Robo-advisors are low cost alternative for retail investors.

Two types of robo-advisory wealth management services are:

Fully Automated Digital Wealth Managers

- fully automated models that require no human assistance
- offer low cost investment portfolio solution e.g. ETFs
- services may include direct deposits, periodic rebalancing, dividend re-investment options

Advisor-Assisted Digital Wealth Managers

- automated services as well as human financial advisor who offers financial advice and periodic reviews through phone.
- such services provide holistic analysis of clients' assets and liabilities

Robo-advisors technology is offering a cost-effective financial guidance for less wealthy investors. Studies suggests that robo-advisors proposing a passive approach, tend to offer fairly conservative advice.

Limitations of Robo-advisors

- The role of robo-advisors dwindles in the time of crises when investors need some expert's guidance.
- Unlike human advisors, the rationale behind the advice of robo-advisors is not fully clear.
- The trust issues with robo-advisors may arise specially after they recommend some unsuitable investments.
- As the complexity and size of investor's portfolio increases, robo-advisor's ability to

deliver detailed and accurate services decreases. For example, portfolios of ultra-wealthy investors include a number of asset-types, and require customization and human assistance.

6.3 Risk Analysis

Stress testing and *risk assessment* measures require wide range of quantitative and qualitative data such as balance sheet, credit exposure, risk-weighted assets, risk parameters, firm and its trading partners' liquidity position. Qualitative information required for stress testing include capital planning procedures, expected changes in business plan, operational risk, business model sustainability etc.

To monitor risk is real time, data and associated risks should be identified and/or aggregated for reporting purpose as it moves within the firm. Big data and ML techniques may provide intuition into real time to help recognize changing market conditions and trends in advance.

Data originated from many alternative sources may be dubious, contain errors or outliers. ML techniques are used to asses data quality and help in selecting reliable and accurate data to be used in risk assessment models and applications.

Advanced AI techniques are helping portfolio managers in performing scenario analysis i.e. hypothetical stress scenario, historical stress event, what if analysis, portfolio backtesting using point-in-time data to evaluate portfolio liquidation costs or outcomes under adverse market conditions.

6.4 Algorithmic Trading

Algorithmic trading is a computerized trading of financial instruments based on some pre-specified rules and guidelines.

Benefits of algorithmic trading includes:

- Execution speed
- Anonymity
- Lower transaction costs

Algorithms continuously update and revise their trading strategy and trading venue to determine the best available price for the order.

Algorithmic trading is often used to slice large institutional orders into smaller orders, which are then executed through various exchanges.

High-frequency trading (HFT) is a kind of algorithmic trading that execute large number of orders in fractions of seconds. HFT makes use of large quantities of granular data (e.g. tick data) real-time prices, market conditions and place trade orders automatically when certain conditions are met. HFT earn profits from intraday market mispricing.

As real-time data is accessible, algorithmic trading plays a vital role in the presence of multiple trading venues, fragmented markets, alternative trading systems, dark-pools etc.

7. DISTRIBUTED LEDGER TECHNOLOGY

Distributed ledger technology (DLT) – advancements in financial record keeping systems – offers efficient methods to generate, exchange and track ownership of financial assets on a peer-to-peer basis.

Potential **advantages** of DLT networks include:

- accuracy
- transparency
- secure record keeping
- speedy ownership transfer
- peer-to-peer interactions

Limitations:

- DLT consumes excessive amount of energy.
- DLT technology is not fully secure, there are some risks regarding data protection and privacy.

Three basic elements of a DLT network are:

i. Digital ledger
ii. A consensus mechanism
iii. Participant network

A **distributed ledger** is a digital database where transactions are recorded, stored and distributed among entities in a manner that each entity has a similar copy of digital data.

Consensus is a mechanism which ensures that entities (nodes) on the network verify the transactions and agree on the common state of the ledger. Two essential steps of consensus are:

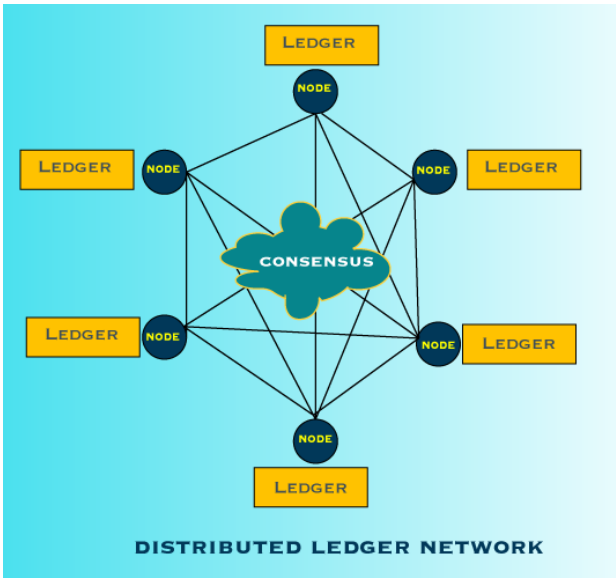
- Transaction validation
- Agreement on ledger update

These steps ensure transparency and data accessibility to its participants on near-real time basis.

Participant network is a peer-to-peer network of nodes (participants).

DLT process uses **cryptography** to verify network participant identity for secure exchange of information among entities and to prevent third parties from accessing the information.

Smart contracts – self-executed computer programs based on some pre-specified and pre-agreed terms and conditions - are one of the most promising potential applications of DLT. For example, automatic transfer of collateral when default occurs, automatic execution of contingent claims etc.



Blockchain:
Blackchain is a digital ledger where transactions are recorded serially in blocks that are then joined using cryptography. Each block embodies transaction data (or entries) and a secure link (hash) to the preceding block so that data cannot be changed retroactively without alteration of previous blocks. New transactions or changes to previous transactions require authorization of members via consensus using some cryptographic techniques.
It is extremely difficult and expensive to manipulate data as it requires very high level of control and huge consumption of energy.

7.1 Permission and Permissionless Networks

DLT networks can be permissionless or permissioned.

Permissionless networks are open to new users. Participants can see all transactions and can perform all network functions.

In permissionless networks:
- *'no central authority'* is required to verify the transaction.
- all transactions are recorded on single database and each node stores a copy of that database.
- records are *immutable* i.e. once data has been entered to the blockchain no one can change it.
- trust is not a requirement between transacting party.

Bitcoin is a renowned model of open, permissionless network.

Permissioned networks are closed networks where activities of participants are well-defined. Only pre-approved participants are permitted to make changes. There may be varying levels of access to ledger from adding data to viewing transaction to viewing selecting details etc.

7.2 Application of Distributed Ledger Technology to Investment Management

In the field of investment management, potential, DLT applications may include:

i. Cryptocurrencies
ii. Tokenization
iii. Post-trade clearing and settlement
iv. Compliance

7.2.1.) Cryptocurrencies

A cryptocurrency is a digital currency that works as a medium of exchange to facilitate near-real-time transactions between two parties without involvement of any intermediary. In contrast to traditional currencies, cryptocurrencies are not government backed or regulated, and are issued privately by individuals or companies. Cryptocurrencies use open DLT systems based on decentralized distributed ledger.

Many cryptocurrencies apply self-imposed limits on the total amount of currency issued which may help to sustain their store of value. However, because of a relatively new concept and ambiguous foundations, cryptocurrencies have faced strong fluctuations in purchasing power.

Nowadays, many start-up companies are interested in funding through cryptocurrencies by *initial coin offering (ICO)*. ICO is a way of raising capital by offering investors units of some cryptocurrency (digital tokens or coins) in exchange for fiat money or other form of digital currencies to be traded in cryptocurrency exchanges. Investors can use digital tokens to purchase future products/services offered by the issuer.

In contrast to IPOs (initial public offerings), ICOs are low-cost and time-efficient. ICOs typically do not offer voting

rights. ICOs are not protected by financial authorities, as a result, investors may experience losses in fraudulent projects. Many jurisdictions are planning to regulate ICOs.

7.2.2.) Tokenization
Tokenization helps in authenticating and verifying ownership rights to assets (such as real estate, luxury goods, commodities etc.) on digital ledger by creating a single digital record. Physical ownership verification of such assets is quite labor-intensive, expensive and requires involvement of multiple parties.

7.2.3.) Post-trade Clearing and Settlement
Another blockchain application in financial securities market is in the field of post-trade processes including clearing and settlement, which traditionally are quite complex, labor-intensive and require several dealings among counterparties and financial intermediaries.

DLT provides near-real time trade verification, reconciliation and settlement using single distributed record ownership among network peers, therefore reduces complexity, time, costs, trade fails and need for third-party facilitation and verification. Speedier process reduces time exposed to counterparty risk, which in turn eases collateral requirements and increases potential liquidity of assets and funds.

7.2.4.) Compliance
Today, amid stringent reporting requirements and transparency needs imposed by regulators, companies are required to perform many risk-related functions to comply with those regulations. DLT has the ability to provide advanced and automated compliance and regulatory reporting procedures which may provide greater transparency, operational efficiency and accurate record-keeping.

DLT-based compliance may provide well-thought-out structure to share information among firms, exchanges, custodians and regulators. Permissioned networks can safely share sensitive information to relevant parties with great ease. DLT makes it possible for authorities to uncover fraudulent activity at lower costs through regulations such as '*know-your-customer*' (KYC) and '*anti-money laundering*' (AML).

Practice: End of Chapter Practice Problems for Reading 6 & FinQuiz Item-sets and questions from FinQuiz Question-bank.

2. CORRELATION ANALYSIS

Scatter plot and correlation analysis are used to examine how two sets of data are related.

2.1 Scatter Plots

A scatter plot graphically shows the relationship between two varaibles. If the points on the scatter plot cluster together in a straight line, the two variables have a strong linear relation. Observations in the scatter plot are represented by a point, and the points are not connected.

2.2 & 2.3 Correlation Analysis & Calculating and Interpreting the Correlation Coefficient

The sample covariance is calculated as:

$$cov_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

where,

n = sample size
X_i = ith observation on variable X
$\bar{X}$ = mean of the variable X observations
Y_i = ith observation on variable Y
$\bar{Y}$ = mean of the variable Y observations

- The covariance of a random variable with itself is simply a variance of the random variable.
- **Covariance** can range from $-\alpha$ to $+\alpha$.
- The covariance number doesn't tell the investor if the relationship between two variables (e.g. returns of two assets X and) is strong or weak. It only tells the direction of this relationship. For example,
  - ***Positive number of covariance*** shows that rates of return of two assets are moving in the same direction: when the rate of return of asset X is negative, the returns of other asset tend to be negative as well and vice versa.
  - ***Negative number of covariance*** shows that rates of return of two assets are moving in the opposite directions: when return on asset X is positive, the returns of the other asset Y tend to be negative and vice versa.

NOTE:

- If there is positive covariance between two assets then the investor should evaluate whether or not he/she should include both of these assets in the same portfolio, because their returns move in the same direction and the risk in portfolio may not be diversified.
- If there is negative covariance between the pair of assets then the investor should include both of these assets to the portfolio, because their returns move in the opposite directions and the risk in portfolio could be diversified or decreased.
- If there is zero covariance between two assets, it means that there is no relationship between the rates of return of two assets and the assets can be included in the same portfolio.

Correlation coefficient measures the direction and strength of linear association between two variables. The correlation coefficient between two assets X and Y can be calculated using the following formula:

$$r_{XY} = \frac{covariance\ of\ X\ and\ Y}{(sample\ standard\ deviation\ of\ X)(sample\ standard\ deviation\ of\ Y)}$$

$$= \frac{cov_{XY}}{(s_X)(s_Y)}$$

or

$$r = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}$$

NOTE:

Unlike Covariance, Correlation has no unit of measurement; it is a simple number.

Example:

$Cov_{xy} = 47.78 \qquad S_x^2 = 40 \qquad S_y^2 = 250$

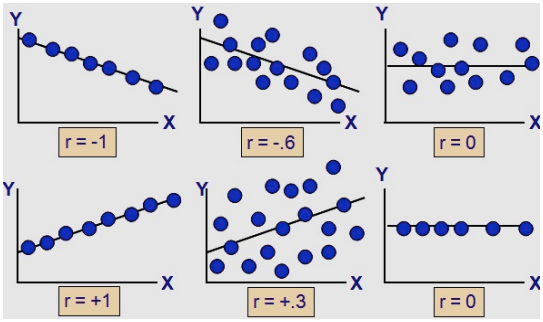
$$r = \frac{47.78}{\sqrt{(40)(250)}} = 0.478$$

- The correlation coefficient can range from -1 to +1.
- Two variables are perfectly positively correlated if correlation coefficient is +1.
- Correlation coefficient of -1 indicates a perfect inverse (negative) linear relationship between the returns of two assets.
- When correlation coefficient equals 0, there is no linear relationship between the returns of two assets.
- The closer the correlation coefficient is to 1, the stronger the relationship between the returns of two assets.

Note: *Correlation of +/- 1 does not imply that slope of the line is +/- 1.*

NOTE:

Combining two assets that have zero correlation with each other reduces the risk of the portfolio. A negative correlation coefficient results in greater risk reduction.



Difference b/w Covariance & Correlation: The covariance primarily provides information to the investor about whether the relationship between asset returns is positive, negative or zero, but correlation coefficient tells the ***degree*** of relationship between assets returns.

NOTE:

Correlation coefficients are valid only if the means, variances & covariances of X and Y are finite and constant. When these assumptions do not hold, then the correlation between two different variables depends largely on the sample selected.

2.4 Limitations of Correlation Analysis

1. **Linearity:** Correlation only measures linear relationships properly.
2. **Outliers:** Correlation may be an unreliable measure when outliers are present in one or both of the series.
3. **No proof of causation:** Based on correlation we cannot assume x causes y; there could be third variable causing change in both variables.
4. **Spurious Correlations:** Spurious correlation is a correlation in the data without any causal relationship. This may occur when:

i. two variables have only chance relationships.
ii. two variables that are uncorrelated but may be correlated if mixed by third variable .
iii. correlation between two variables resulting from a third variable.

NOTE:

Spurious correlation may suggest investment strategies that appear profitable but actually would not be so, if implemented.

2.6 Testing the Significance of the Correlation Coefficient

t-test is used to determine if sample correlation coefficient, r, is statistically significant.

Two-Tailed Test:

Null Hypothesis H_0 : the correlation in the population is 0 ($\rho = 0$);

Alternative Hypothesis H_1: the correlation in the population is different from 0 ($\rho \neq 0$);

NOTE:

The null hypothesis is the hypothesis to be tested. The alternative hypothesis is the hypothesis that is accepted if the null is rejected.

The formula for the t-test is (for normally distributed variables):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$$

where,

r is the sample coefficient of correlation calculated by

$r = \frac{cov(X,Y)}{S_x S_y}$

t = t-statistic (or calculated t)
n – 2 = degrees of freedom

Decision Rule:

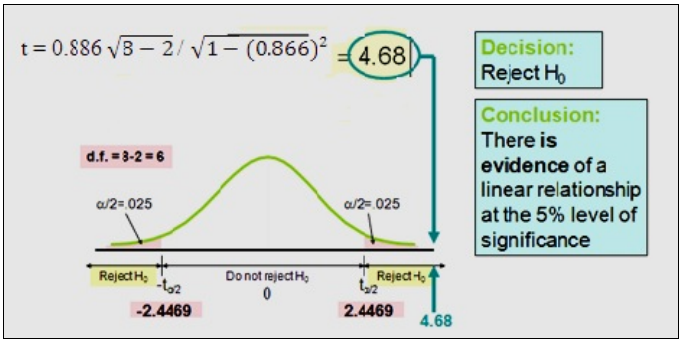
If test statistic is < – t-critical or > + t-critical with n-2 degrees of freedom, (if absolute value of t > t_c), Reject H_0; otherwise Do not Reject H_0.

Example:

Suppose r = 0.886 and n = 8, and t_C = 2.4469 (at 5% significance level i.e. α = 5%/2 and degrees of freedom = 8 – 2 = 6)

$t = 0.886\sqrt{\frac{8-2}{1-(0.886)^2}}$ = 4.68 → Since t-value > t_c, we reject

null hypothsis of no correlation.



Magnitute of r needed to reject the null hypothesis (H_0: $\rho = 0$) decreases as sample size n increases. Because as n increases the:

- number of degrees of freedom increases
- absolute value of t_c decreases.
- t-value increases

In other words, type II error decreases when sample size (n) increases, all else equal.

NOTE:

Type I error = reject the null hypothesis although it is true.
Type II error = do not reject the null hypothesis although it is wrong.

Practice: Example 7, 8, 9 & 10 Volume 1, Reading 7.

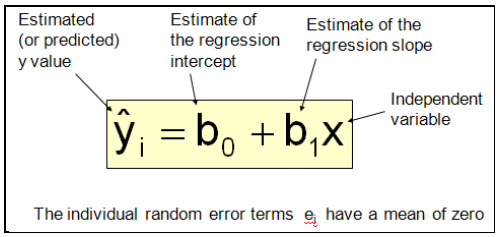
3. LINEAR REGRESSION

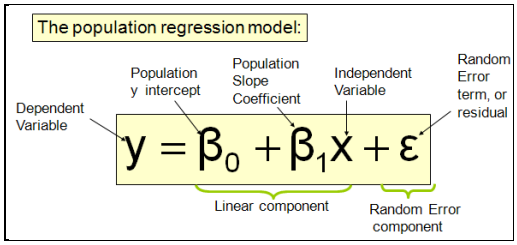
Regression analysis is used to:

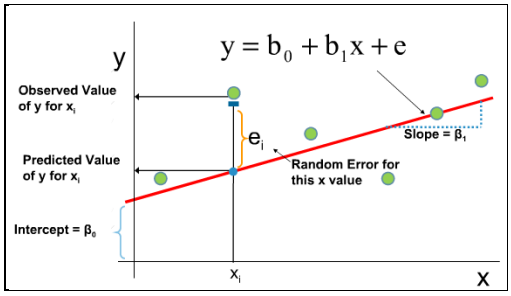
- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable.

Linear regression assumes a linear relationship between the dependent and the independent variables. Linear regression is also known as ***linear least squares*** since it selects values for the intercept b_0 and slope b_1 that minimize the sum of the squared vertical distances between the observations and the regression line.

Estimated Regression Model: The sample regression line provides an estimate of the population regression line. Note that population parameter values b_0 and b_1 are not observeable; only estimates of b_0 and b_1 are observeable.







Dependent variable: The variable to be explained (or predicted) by the independent variable. Also called endogenous or predicted variable.

Independent variable: The variable used to explain the dependent variable. Also called exogenous or predicting variable.

Intercept (b_0): The predicted value of the dependent variable when the independent variable is set to zero.

$$b_0 = \bar{y} - b_1\bar{x}$$

Slope Coefficient or regression coefficient (b_1): A change in the dependent variable for a unit change in the independent variable.

$$b_1 = \frac{cov(x,y)}{var(x)}$$

or

$$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

Error Term: It represents a portion of the dependent variable that cannot be explained by the independent varaiable.

Example:

n =100

$$\bar{x} = 36{,}009.45; \quad s_x^2 = \frac{\sum(x_i-\bar{x})^2}{n-1} = 43{,}528{,}688$$

$$\bar{y} = 5{,}411.41; \quad \text{cov}(X,Y) = \frac{\sum(x_i-x)(y_i-y)}{n-1} = -1{,}356{,}256$$

$$\hat{y} = b_0 + b_1 x = 6{,}535 - 0.0312x$$

$$b_1 = \frac{\text{cov}(X,Y)}{s_x^2} = \frac{-1{,}356{,}256}{43{,}528{,}688} = -0.0312$$

$$b_0 = \bar{y} - b_1\bar{x} = 5{,}411.41 - (-0.0312)(36{,}009.45) = 6{,}535$$

Types of data used in regression analysis:

1) **Time-series:** It uses many observations from ***different time periods*** for the ***same*** company, asset class or country etc.

2) **Cross-sectional:** It uses many observations for the ***same time period***of ***different*** companies, asset classes or countries etc.

3) **Panel data:** It is a mix of time-series and cross-sectional data.

3.2 Assumptions of the Linear Regression Model

1. The regression model is linear in its parameters b_0 and b_1 i.e. b_0 and b_1 are raised to power 1 only and neither b_0 nor b_1 is multiplied or divided by another regression parameter e.g. b_0 / b_1.

- When regression model is nonlinear in parameters, regression results are invalid.
- Even if the dependent variable is nonlinear but parameters are linear, linear regression can be used.

2. Independent variables and residuals are uncorrelated.
3. The expected value of the error term is 0.

- When assumptiuons 2 & 3 hold, linear regression produces the correct estimates of b_0 and b_1.

4. The variance of the error term is the same for all observations. (It is known as Homoskedasticity assumption).
5. Error values (ε) are statistically independent i.e. the error for one observation is not correlated with any other observation.
6. Error values are normally distributed for any given value of x.

3.3 The Standard Error of Estimate

Standard Error of Estimate (SEE) measures the degree of variability of the actual y-values relative to the estimated (predicted) y-values from a regression equation. Smaller the SEE, better the fit.

$$Standard\ Error\ of\ Estimate: S_E = \sqrt{\frac{SSE}{n-k-1}}$$

or

$$SEE = S_E = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}},$$

where,

SSE = Sum of squares error
n = Sample size
k = number of independent variables in the model

Example:

n = 100

SSE = 2,252,363

Thus,

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{2,252,363}{98}} = 151.60$$

Regression Residual is the difference between the actual values of dependent variable and the predicted value of the dependent variable made by regression equation.

3.4 The Coefficient of Determination

The coefficient of determination is the portion of the total variation in the dependent variable that is explained by the independent variable. The coefficient of determination is also called R-squared and is denoted as R^2.

$$Coefficient\ of\ determination\ (R^2)$$

$$= \frac{Total\ Variation\ (SST) - Unexplained\ Variation\ (SSE)}{Total\ Variation\ (SST)}$$

$$= \frac{Explained\ Variation (RSS)}{Total\ Variation\ (SST)}$$

where,

$0 \leq R^2 \leq 1$

In case of a single independent variable, the coefficient of determination is: $R^2 = r^2$

where,

R^2 = Coefficient of determination
r = Simple correlation coefficient

Example:

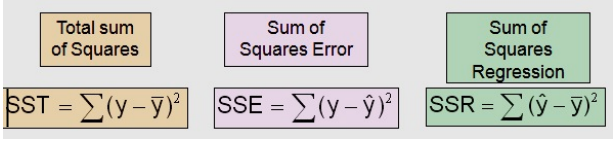
Suppose correlation coefficient between returns of two assets is + 0.80, then the coefficient of determination will be 0.64. The interpretation of this number is that approximately 64 percent of the variability in the returns of one asset (or dependent variable) can be explained by the returns of the other asset (or indepepnent variable). If the returns on two assets are perfectly correlated (r = +/- 1), the coefficient of determination will be equal to 100 %, and this means that if changes in returns of one asset are known, then we can exactly predict the returns of the other asset.

NOTE:

Multiple R is the correlation between the actual values and the predicted values of Y. The coefficient of determination is the square of multiple R.

Total variation is made up of two parts:

SST = SSE + SSR(or RSS)



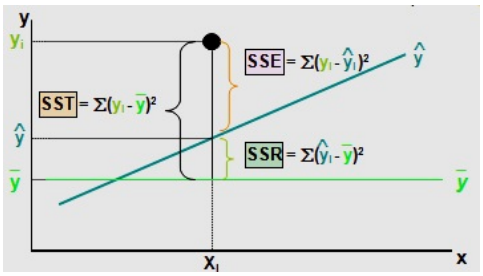
where,

$\bar{y}$ = Average value of the dependent variable
y = Observed values of the dependent variable
$\hat{y}$ = Estimated value of y for the given value of x

- **SST (total sum of squares):** Measures total variation

in the dependent variable i.e. the variation of the y_i values around their mean y.

- **SSE (error sum of squares):** Measures unexplained variation in the dependent variable.
- **SSR / RSS (regression sum of squares):** Measures variation in the dependent variable explained by the independent variable.



Practice: Example 13 Volume 1, Reading 7.

3.5 Hypothesis Testing

In order to determine whether there is a linear relationship between x and y or not, significance test (i.e. t-test) is used instead of just relying on b_1 value. t-statistic is used to test the significance of the individual coefficients (e.g. slope) in a regression.

Null and Alternative hypotheses

H_0: $b_1 = 0$ (no linear relationship)
H_1: $b_1 \neq 0$ (linear relationship does exist)

Test statistic = $$t = \frac{\hat{b}_1 - b_1}{s_{b_1}}$$

where,

$\hat{b}_1$ = *Sample regression slope coefficient*
b_1 = *Hypothesized slope*
S_{b_1} = *Standard error of the slope*
df= n–2

Decision Rule:

If test statistic is <– t-critical or > + t-critical with n-2 degrees of freedom, (if absolute value of t > t_c), Reject H_0; otherwise Do not Reject H_0.

Two-Sided Test	One-sided Test
H_0: $b_1 = 0$	H_0: $b_1 = 0$
H_A: $b_1 \neq 0$	H_A^+: $b_1 > 0$ or
	H_A^-: $b_1 < 0$

Confidence Interval Estimate of the Slope: Confidence interval is an interval of values that is expected to include the true parameter value b_1 with a given degree of freedom.

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

df = n - 2

Example:

$$n = 7 \quad \hat{b}_1 = -9.01, \quad \hat{s}_{\hat{b}_1} = 1.50, \quad b_1 = 0$$

Testing H_0: $b_1 = 0$ v/s H_A: $b_1 \neq 0$

$$T.S.: t_{obs} = \frac{-9.01 - 0}{1.50} = -6.01 \qquad R.R.: |t_{obs}| \geq t_{.025,5} = 2.571$$

95% Confidence Interval for b_1:

–9.01 ± 2.571(1.50) = –9.01 ± 3.86 = (–12.87 to –5.15)

- As this interval does not include 0, we can reject H_0. Therefore, we can say with 95% confidence that the regression slope is different from 0.

- Reject H_0 because t-value 6.01 > critical t_c 2.571.

NOTE:

Higher level of confidence or lower level of significance results in higher values of critical 't' i.e. t_c. This implies that:

- Confidence intervals will be larger.
- Probability of rejecting the H_0 decreases i.e. type –II error increases.
- The probability of Type-I error decreases.

Stronger regression results lead to smaller standard errors of an estimated parameter and result in tighter confidence interval. As a result probability of rejecting H_0 increases (or probability of Type-I error increases).

p-value: The p-value is the smallest level of significance at which the null hypothesis can be rejected.

Decision Rule: If p < significance level, H_0 can be rejected. If p > significance level, H_0 cannot be rejected.

For example, if the p-value is 0.005 (0.5%) & significance level is 5%, we can reject the hypothesis that true parameter equals 0.

Practice: Example 14, 15 & 16 Volume 1, Reading 7.

3.6 Analysis of Variance in a Regression with One Independent Variable

Analysis of Variance (ANOVA) is a statistical method used to divide the total variance in a study into meaningful pieces that correspond to different sources. In regression analysis, ANOVA is used to determine the

usefulness of one or more independent variables in explaining the variation in dependent variable.

ANOVA	df	SS	MSS	F
Regression	k	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\frac{SSR}{k}$	$\frac{SSR/k}{SSE/(n-k-1)}$
Error	n–k–1	$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$	$\frac{SSE}{n-k-1}$	
Total	n–1	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$		

Or

Source of Variability	DoF	Sum of Squares	Mean Sum of Squares
Regression (Explained)	1	RSS	MSR = RSS/1
Error (Unexplained)	n-2	SSE	MSE = SSE/n-2
Total	n-1	SST=RSS + SSE	

F-Statistic or F-Test evaluates how well a set of independent variables, as a group, explains the variation in the dependent variable. In multiple regression, the F-statistic is used to test whether at least one independent variable, in a set of independent variables, explains a significant portion of variation of the dependent variable. The F statistic is calculated as the ratio of the average regression sum of squares to the average sum of the squared errors,

$$\frac{MSR}{MSE} = \frac{(\frac{RSS}{k})}{(\frac{SSE}{n-k-1})}$$

df numerator = k = 1
df denominator = n – k – 1 = n – 2

Decision Rule: Reject H_0 if F>F-critical.

Note: F-test is always a one-tailed test.

In a regression with just one independent variable, the F statistic is simply the square of the t-statistic i.e. F= t^2. F-test is most useful for multiple independent variables while the t-test is used for one independent variable.

NOTE:

When independent variable in a regression model does not explain any variation in the dependent variable, then the predicted value of y is equal to mean of y. Thus, RSS = 0 and F-statistic is 0.

Practice: Example 17
Volume 1, Reading 7.

3.7 Prediction Intervals

$$\hat{Y} \pm t_c s_f$$

where,

$$s_f^2 = s^2\left[1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{(n-1)s_X^2}\right]$$

and

$$s_f = \sqrt{s_f^2}$$

s^2 = squared SEE
n = number of observations
X = value of independent variable
$\bar{X}$ = estimated mean of X
s^2_X= variance of independent variable
t_c = critical t-value for n –k –1 degrees of freedom.

Example:

Calculate a 95% prediction interval on the predicted value of Y. Assume the standard error of the forecast is 3.50%, and the forecasted value of X is 8%. And n = 36. Assume: Y = 3% + (0.50)(X)

The predicted value for Y is: Y =3% + (0.50)(8%)= 7%

The 5% two-tailed critical t-value with 34 degrees of freedom is 2.03. The prediction interval at the 95% confidence level is:

7% +/- (2.03 ×3.50%) = - 0.105% to 14.105%

This range can be interpreted as, "given a forecasted value for X of 8%, we can be 95% confident that the dependent variable Y will be between –0.105% and 14.105%".

Practice: Example 18
Volume 1, Reading 7.

Sources of uncertianty when using regression model & estimated parameters:

1. Uncertainty in Error term.
2. Uncertainty in the estimated parameters b_0 and b_1.

3.8 Limitations of Regression Analysis

- Regression relations can change over time. This problem is known as ***Parameter Instability.***
- If public knows about a relation, this results in no

relation in the future i.e. relation will break down.
- Regression is based on assumptions. When these assumptions are violated, hypothesis tests and predictions based on linear regression will be invalid.

Practice: End of Chapter Practice Problems for Reading 7 & FinQuiz Item-set ID# 15579, 15544 & 11437.

2. MULTIPLE LINEAR REGRESSION

Multiple linear regression is a method used to model the linear relationship between a dependent variable and more than one independent (explanatory or regressors) variables. A multiple linear regression model has the following general form:

Multiple Regression Model with k Independent Variables:

Y-intercept | Population slopes | Random Error

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon$$

where,

Y_i = i^{th} *observation of dependent variable Y*
X_{ki} = i^{th} *observation of* k^{th} *independent variable X*
β_0 = *intercept term*
β_k = *slope coefficient of* k^{th} *independent variable*
ε_i = *error term of ith observation*
n = number of observations
k = total number of independent variables

- ***A slope coefficient,*** β_j ***is known as partial regression coefficients or partial slope coefficients.*** It measures how much the dependent variable, Y, changes when the independent variable, X_j, changes by one unit, ***holding all other independent variables constant.***
- ***The intercept term*** (β_0) is the value of the dependent variable when the independent variables are all equal to zero.
- A regression equation has k slope coefficients and k + 1 regression coefficients.

Practice: Example 1 Volume 1, Reading 8.

Simple vs. Multiple Regression

Simple Regression	Multiple Regression
1. One dependent variable Y predicted from one independent variable X 2. One regression coefficient 3. **r^2**: proportion of variation in dependent variable Y predictable from X	1. One dependent variable Y predicted from a set of independent variables (X_1, X_2 …. X_k) 2. One regression coefficient for each independent variable 3. **R^2**: proportion of variation in dependent variable Y predictable by set of independent variables (X's)

2.1 Assumptions of the Multiple Linear Regression Model

The Multiple linear regression model is based on following six assumptions. When these assumptions hold, the regression estimators are ***unbiased, efficient*** and ***consistent***.

NOTE:

- Unbiased means that the expected value of the estimator is equal to the true value of the parameter.
- Efficient means that the estimator has a smaller variance than any other estimator.
- Consistent means that the biasness and variance of the estimator approach zero as the sample size increases.

Assumptions:

1. The relationship between the dependent variable, Y, and the independent variables, X_1, X_2, . . . ,X_k, is linear.
2. The independent variables (X_1, X_2, . . . ,X_k) are not random. Also, no exact linear relation exists between two or more of the independent variables.
3. The expected value of the error term, conditional on the independent variables, is 0: $E(\varepsilon | X_1, X_2, \ldots, X_k) = 0$.
4. The variance of the error term is constant for all observations i.e. errors are ***Homoskedastic***.
5. The error term is uncorrelated across observations (i.e. ***no serial correlation***).
6. The error term is normally distributed.

NOTE:

- Linear regression can't be estimated when an exact linear relationship exists between two or more independent variables. But when two or more independent variables are highly correlated, although there is no exact relationship, it leads to multicollinearity problem. (Discussed later in detail).
- Even if independent variable is random but uncorrelated with the error term, regression results are reliable.

Practice: Example 2 & 3 Volume 1, Reading 8.

2.2 Predicting the Dependent Variable in a Multiple Regression Model

The process of calculating the predicted value of dependent variable is the same as we did in Reading 11.

Prediction equation

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} + \ldots + \hat{b}_k X_{ki}$$

where,

$\hat{Y}_i$: Estimated or predicted value of Y
b_0: Estimated intercept
b_1, b_2,… & b_k: Estimated slope coefficients

Assumptions of the regression model must hold in order to have reliable prediction results.

Practice: Example 4
Volume 1, Reading 8.

Sources of uncertainity when using regression model & estimated parameters:

1. Uncertainity in error term.
2. Uncertainity in the estimated parameters of the model.

2.3 Testing Whether All Population Regression Coefficients Equal Zero

To test the significance of the regression as a whole, we test the null hypothesis that all the slope coefficients in a regression are simultaneously equal to 0.

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no linear relationship)
H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

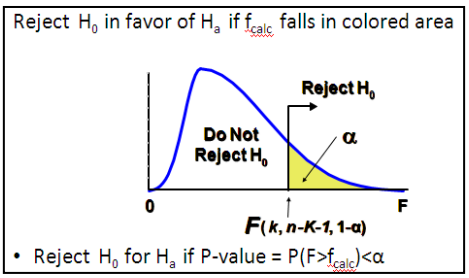
In multiple regression, the F-statistic is used to test whether at least one independent variable, in a set of independent variables, explains a significant portion of variation of the dependent variable. The F statistic is calculated as the ratio of the mean regression sum to squares of the mean squared error,

$$\frac{MSR}{MSE} = \frac{RSS/k}{SSE/(n-k-1)}$$

df numerator = k
df denominator = n – k – 1

Note: F-test is always a one-tailed test.

Decision Rule: Reject H_0 if F>F-critical.



NOTE:

When independent variable in a regression model does not explain any variation in the dependent variable, then the predicted value of y is equal to mean of y. Thus, RSS = 0 and F-statistic is 0.

- Larger R^2 produces larger values of F.
- Larger sample sizes also tend to produce larger values of F.
- *The lower the p-value, the stronger the evidence against that null hypothesis.*

Example:

k = 2
n = 1,819
df = 1,819 – 2 – 1 = 1,816
SSE = 2,236.2820
RSS = 2,681.6482
α = 5%
F-statistic = $\frac{MSR}{MSE}$ = (2,681.6482/2) / (2,236.2820/1,816) = 1,088.8325

F-critical with numerator df = 2 and denominator df = 1,816 is 3.00.

Since F-statistic > F-critical, Reject H_0 that coefficients of both independent variables equal 0.

2.4 Adjusted R^2

In multiple linear regression model, R^2 is less appropriate as a measure to test the "goodness of fit" of the model because R^2 always increases when the number of independent variables increases. It is important to keep in mind that a high R^2does not imply causation.

The ***adjusted R^2*** is used to deal with this artificial increase in accuracy. Adjusted R^2 does not automatically increase when another variable is added to a regression; it is adjusted for degrees of freedom. The *adjusted R^2* is given by

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2)$$

where,

n = sample size,
k = number of independent variables

- When $k \geq 1$, then R^2 is strictly > Adjusted R^2.
- Adjusted R^2 decreases if the new variable added does not have any significant explanatory power.

- Adjusted R^2 can be negative as well but R^2 is always positive.
- Adjusted R^2 is always ≤ R^2.

NOTE:

When Adjusted R^2 is used to compare regression models, both the dependent variable definition and sample size must be same for each model.

3. USING DUMMY VARIABLES IN REGRESSIONS

Dummy variable is a qualitative variable that takes on a value of 1 if a particular condition is true and 0 if that condition is false. It is used to account for qualitative variables such as male or female, month of the year effects, etc.

Suppose we want to test whether total returns of one small-stock index, the Russell 2000 Index, differ by months. We can use dummy variables to estimate the following regression,

$Returns_t = b_0 + b_1 jan_t + b_2 Feb_t + \ldots + b_{11} Nov_t + \varepsilon_t$

- If we want to distinguish among n categories, we need n -1 dummy variables e.g. in above regression model we will need 12 – 1 = 11 dummy variables. If we take 12 dummy variables, Assumption 2 is violated.
- b_0 represents average return for stocks in December.
- $b_1, b_2, b_3, .., b_{11}$ represent difference between returns in that month and returns for December i.e.
  - Average stock returns in Dec = b_0
  - Average stock returns in Jan = $b_0 + b_1$
  - Average stock returns in Feb = $b_0 + b_2$
  - Average stock returns in Nov = $b_0 + b_{11}$

As with all multiple regression results, the F-statistic for the set of coefficients and the R^2 are evaluated to determine if the months, individually or collectively, contribute to the explanation of monthly return. We can also test whether the average stock return in each of the months is equal to the stock return in Dec (the omitted month) by testing the individual slope coefficient using the following null hypotheses:

H_0: $b_1 = 0$ (i.e. stock return in Dec = stock return in Jan)
H_0: $b_2 = 0$ (i.e. stock return in Dec = stock return in Feb)
and so on….

Practice: Example 5
Volume 1, Reading 8.

4. VIOLATIONS OF REGRESSION ASSUMPTIONS

4.1 Heteroskedasticity

Heteroskedasticity occurs when the variance of the errors differs across observations i.e. variances are not constant.

Types of Heteroskedasticity:

1. ***Unconditional Heteroskedasticity***: It occurs when Heteroskedasticity of the error variance does not systematically increase or decrease with changes in the value of the independent variable. Although it violates Assumption 4, but it creates no serious problems with regression.

2. ***Conditional Heteroskedasticity:*** Conditional heteroskedasticity exists when Heteroskedasticity of the error variance increases as the value of independent variable increases. It is more problematic than unconditional hetroscadasticity.

4.1.1) Consequences of (Conditional) Heteroskedasticity:

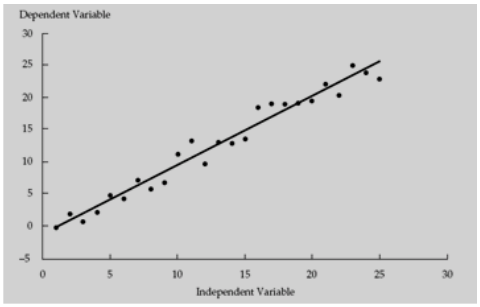
- It does not affect consistency but it can lead to wrong inferences.
- Coefficient estimates are not affected.
- It causes the F-test for the overall significance to be unreliable.
- It introduces biasness into estimators of the standard error of regression coefficients; thus, t-tests for the significance of individual regression coefficients are unreliable.

When Heteroskedasticity results in underestimated standard errors, t-statistics are inflated and probability of Type-I error increases. The opposite will be true if standard errors are overestimated.

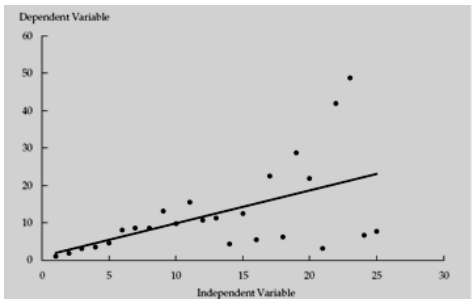
4.1.2) Testing for Heteroskedasticity:

1. ***Plotting residuals:*** A scatter plot of the residuals versus one or more of the independent variables can describe patterns among observations (as shown below).

Regressions with Homoskedasticity



Regressions with Heteroskedasticity



2. Using ***Breusch–Pagan test***: The Breusch–Pagan test involves regressing the squared residuals from the estimated regression equation on the independent variables in the regression.

H_0 = No conditional Heteroskedasticity exists
H_A = Conditional Heteroskedasticity exists

Test statistic = $n \times R^2_{residuals}$

where,

$R^2_{residuals}$ = *R^2 from a second regression of the squared residuals from the first regression on the independent variables*
n = number of observations

- Critical value is calculated from χ^2 distribution table with df = k.
- It is a one-tailed test since we are concerned only with large values of the test statistic.

Decision Rule: When test statistic > critical value, Reject H_0 and conclude that error terms in the regression model are conditionally Heteroskedastic.

- If no conditional heteroskedasticity exists, the independent variables will not explain much of the variation in the squared residuals.
- If conditional heteroskedasticity is present in the original regression, the independent variables will explain a significant portion of the variation in the squared residuals.

Practice: Example 8
Volume 1, Reading 8.

4.1.3) Correcting for Heteroskedasticity:

Two different methods to correct the effects of conditional heteroskedasticity are:

1. Computing ***robust standard errors*** (heteroskedasticity-consistent standard errors or white-corrected standard errors), corrects the standard errors of the linear regression model's estimated coefficients to deal with conditional heteroskedasticity.

2. ***Generalized least squares*** (GLS) method is used to modify the original equation in order to eliminate the heteroskedasticity.

4.2 Serial Correlation

When regression errors are correlated across observations, then errors are serially correlated (or auto correlated). Serial correlation most typically arises in time-series regressions.

Types of Serial Correlation:

1. ***Positive serial correlation*** is a serial correlation in which a positive (negative) error for one observation increases the probability of a positive (negative) error for another observation.

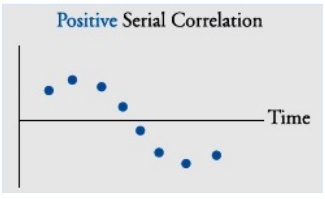
2. ***Negative serial correlation*** is a serial correlation in which a positive (negative) error for one observation increases the probability of a negative (positive) error for another observation.

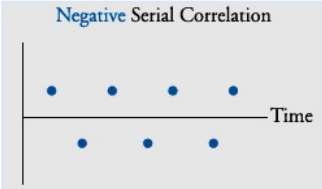
4.2.1) Consequences of Serial Correlation:

- The principal problem caused by serial correlation in a linear regression is an incorrect estimate of the regression coefficient standard errors.
- When one of the independent variables is a lagged value of the dependent variable, then serial correlation causes all the parameter estimates to be inconsistent and invalid. Otherwise, serial correlation does not affect the consistency of the estimated regression coefficients.
- Serial correlation leads to wrong inferences.
- In case of positive (negative) serial correlation: *Standard errors are underestimated (overestimated) → T-statistics (& F-statistics) are inflated (understated) →Type-I (Type-II) error increases.*

4.2.2) Testing for Serial Correlation:

1. ***Plotting residuals*** i.e. a scatter plot of residuals versus time (as shown below).





2. ***Using Durbin-Watson Test:*** The Durbin Watson statistic is used to test for serial correlation

$$DW = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\varepsilon}_t^{\,2}}$$

where,

$\hat{\varepsilon}_t$ is the regression residual for period t.
The DW statistic tests the null hypothesis of no autocorrelation against the alternative hypothesis of positive (or negative) autocorrelation. In case of large sample size, Durbin-Watson statistic (d) is approximately equal to

$$\mathbf{d} \approx 2\,(1 - r)$$

where,

r = sample correlation b/w regression residuals from one period and from the previous period.

The above equation implies that

- d = 2, if no autocorrelation (r = 0)
- d = 0, if autocorrelation is +1.0
- d = 4, if autocorrelation is –1.0

Decision Rule:

A. For positive autocorrelation, the decision rule is:

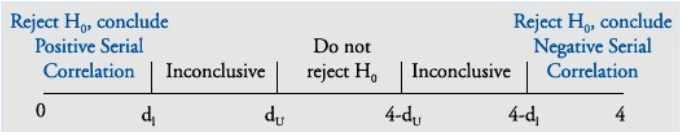
H_0; no positive auto correlation
H_a; positive auto correlation

- If $d < d_l \rightarrow$ Reject H_0
- If $d > d_u \rightarrow$ Do not reject H_0
- If $d_l \leq d \leq d_u \rightarrow$ Inconclusive

B. For negative autocorrelation, the decision rule is:

H_0; no negative auto correlation
H_a; negative auto correlation

- If $d > 4 - d_l \rightarrow$ Reject H_0
- If $d < 4 - d_u \rightarrow$ Do not reject H_0
- If $4 - d_u \leq d \leq 4 - d_l \rightarrow$ Inconclusive



4.2.3) Correcting for Serial Correlation:

The two different methods to correct effects of serial correlation are:

1. **Adjust** the coefficient standard errors for the linear regression parameter estimates to account for the serial correlation e.g. using **Hansen's method**. Hansen's method also simultaneously corrects for conditional heteroskedasticity. (Mostly this method is recommended).
2. **Modify** the regression equation itself to eliminate serial correlation.

4.3 Multicollinearity

Multicollinearity occurs when two or more independent variables (or combinations of independent variables) are highly (but not perfectly) correlated with each other.

4.3.1) Consequences of Multicollinearity:

- A high degree of multicollinearity can make it difficult to detect significant relationships.
- Multicollinearity does not affect the consistency of the estimates of the regression coefficients but estimates become extremely imprecise and unreliable.
- It does not affect F-statistic.
- The multicollinearity problem does not result in biased coefficient estimates; however, standard errors of regression coefficients can increase, causing insignificant t-tests and wide confidence intervals i.e. Type-II error increases.

4.3.2) Detecting Multicollinearity

- High pairwise correlations among independent variables do not necessarily indicate presence of multicollinearity while a low pairwise correlation among independent variables is not an evidence that multicollinearity does not exist. Correlation between independent variables is useful as an indicator of multicollinearity only in case of two independent variables.
- The classic symptom of multicollinearity is a high R^2 (and significant F-statistic) even though the t-statistics on the estimated slope coefficients are not significant.

4.3.3) Correcting for Multicollinearity

The problem of multicollinearity can be corrected by excluding one or more of the regression variables.

4.4 Summarizing the Issues

Problem	How to detect	Consequences	Possible Corrections
(Conditional) Heteroskedasticity i.e. Errors are correlated with earlier X	Plot residuals or use Breusch–Pagan test	Wrong inferences ; incorrect standard errors	Use robust standard errors or GLS
Serial correlation i.e. Errors are correlated with earlier errors	Durbin-Watson Test	Wrong inferences ; incorrect standard errors	Use robust standard errors (Hansen's method) or modifying equation
Multicollinearity i.e. independent variables are strongly correlated with each other	High R^2 and significant F-statistic but low t-statistic	Wrong inferences ;	Omit variable

5. MODEL SPECIFICATION AND ERRORS IN SPECIFICATION

Model specification refers to the set of variables included in the regression and the regression equation's functional form. Incorrect model specification can result in biased & inconsistent parameter estimates and violations of other assumptions.

5.1 Principles of Model Specification

1. The model should be based on logical economic reasoning.
2. The functional form chosen for the variables in the regression should be compatible with the nature of the variables.
3. The model should be *parsimonious* (i.e. economical both in terms of time & cost).
4. The model should be examined for any violation of regression assumptions before being accepted.
5. The model should be tested for its validity & usefulness out of sample before being accepted.

Types of misspecifications:

1. Misspecified Functional Form:

a) *Omitted variables bias*: One or more important variables are omitted from regression.

- When relevant variables are excluded, result can be biased & inconsistent parameter estimates (unless the omitted variable is uncorrelated with the included ones).
- When irrelevant variables are included, standard errors are overestimated.

b) One or more of the regression variables may need to be transformed (for example, by taking the natural logarithm of the variable) before estimating the regression.

c) The regression model pools data from different samples that should not be pooled.

2. *Independent variables are correlated with the error term.* This is a violation of Regression Assumption 3, that the error term has a mean of 0, and causes the estimated regression coefficients to be biased and inconsistent. Three common problems that cause this type of time-series misspecification are:

a) Including lagged dependent variables as independent variables in regressions (with serially correlated errors) e.g. $Y_t = b_0 + b_1X_t + b_2 Y_{t-1} + \varepsilon_t$

b) Including a function of dependent variables as an independent variable i.e. forecasting "past" instead of future e.g. $EPS_t = b_0 + b_1BV_t + \varepsilon_t$; we should rather use $Y_t = b_0 + b_1 BV_{t-1} + \varepsilon_t$

c) Independent variables that are measured with error i.e. due to use of wrong proxy variable. When this problem exists in a single independent variable regression, the estimated slope coefficient on that variable will be biased toward 0.

3. *Other types of Time-series Misspecification* e.g. nonstationarity problem, which results in non-constant mean and variance over time. (Discussed in detail in Reading 13)

6. MODEL WITH QUALITATIVE DEPENDENT VARIABLES

Qualitative dependent variables are dummy variables used as dependent variables instead of independent variables.

- The ***probit model*** is based on the normal distribution and estimates the probability that Y = 1 (a condition is fulfilled) given the value of the independent variable X.
- The ***logit model*** is identical to probit model, except that it is based on the logistic distribution rather than the normal distribution.
- ***Discriminant analysis*** is based on a linear function, similar to a regression equation, which is used to create an overall score. Based on the score, an observation can be classified into categories such as bankrupt or not bankrupt.

Economic meaning of the results of multiple regression analysis and criticism of a regression model and its results:

1. The validity of a regression model is based on its assumptions. When these assumptions do not hold, regression estimates and results are inaccurate and invalid.
2. Regression does not prove causality between variables; it only discovers correlations between variables.
3. Regression Analysis focuses on its use for statistical inference only. A relationship may be statistically significant but has no economic significance e.g. a regression model may identify a statistically significant abnormal return after the dividend announcement, but these returns may prove unprofitable when transactions costs are taken into account.

Practice: End of Chapter Practice Problems for Reading 8 &FinQuiz Item-set ID# 11514, 15830 & 16534.

7. Machine Learning

In real time, huge amount of data (commonly called Big Data) is being created by institutions businesses, governments, financial markets, individuals, and sensors (e.g. satellite imaging). Investors generally use big data information to find better investment opportunities.

Big data covers data from traditional and non-traditional sources. Analysis of big data is challenging because:

- non-traditional data sources are often unstructured
- theoretical methods do not perform well to establish relationships among the data at such a massive scale.

Machine learning (advanced computer techniques), computer algorithms and adaptable models are used to study relationships among the data. The information obtained from big data using such techniques is called data analysis (a.k.a 'data analytics').

7.1 Major Focuses of Data Analytics

Six focuses of data analytics include:

1. **Measuring correlations**
   Determining synchronous relationship between variables i.e. how variables tend to covary.
2. **Making predictions**
   Identifying variables that can help predict the value of variable of interest.
3. **Making casual inferences**
   Casual inference focuses on determining whether an independent variable cause changes to the dependent variable. Casual inference is a stronger relationship between variables than that of correlation and prediction. However, in real-world situation, estimating casual effect in the presence of confounding variables (variables that influence both dependent and independent variables) is challenging.
4. **Classifying data**
   Classification focuses on classifying variables into various categories. Variables can be continuous variables (such as time, weight) or categorical variables (countable distinct groups). In case of categorical variables, the econometric model is called a classifier. Many classification models are binary classifiers (two possible values 0 or 1), others are multicategory classifiers (such as ordinal or nominal). Ordinal variables follow some natural order (small, medium, large or low to high ratings etc.). Nominal variables do not follow any natural order (e.g. equity, fixed income, alternate).

5. **Sorting data into clusters**
   Clustering focuses on sorting observations into various groups based on similar attributes or set of criteria that may or may not be prespecified.

6. **Reducing the dimension of data**
   Dimension reduction is a process of reducing number of independent variables while retaining variation across observations.

- Dimension reduction when applied to data with large number of attributes makes easier to visualize the data on computer screens.
- For out of sample forecasting, simple models perform better than complex models.
- Dimension reduction improves performance by focusing on major factors that cause asset price movements in quantitative investment and risk management.

All these problems (prediction, clustering, dimension reduction, classification etc.) are often solved by machine learning methods.

7.2 What is Machine Learning?

Machine learning (ML) is a subset of artificial intelligence (AI). **Machine Learning (ML)** uses statistical techniques that give computer systems the ability to act by learning from data without being explicitly programmed.

The ML program uses inputs from historical database, trends and relationships to discover hidden insights and pattern in data.

7.3 Types of Machine Learning

Two broad categories of ML techniques are:

1. Supervised learning
Supervised learning uses labeled training data (set of inputs supplied to the program), and process that information to find the output. Supervised learning follows the logic of 'X leads to Y'.

For example, consider an ML program that predicts whether credit card transactions are fraudulent or not. This is a binary classifier where the transaction is either fraudulent (value = 1) or non-fraudulent (value = 0). The ML program collects input from the growing database of credit card transactions labeled 'fraudulent' or 'non-fraudulent' and learns the relationship from experience. The performance is measured by the percentage of transactions accurately predicted.

2. Unsupervised learning
Unsupervised learning does not make use of labelled training data and does not follow the logic of 'X leads to Y'. There are no outcomes to match to, however, the input data is analyzed, and the program discovers structures within the data itself.

One application of unsupervised learning is "clustering' where program identifies similarity among data points and automatically splits data into groups based on their similar attributes.

Note:
Some additional ML categories are '**deep learning**' (ML program using neural network with many hidden layers) and '**reinforcement learning**' (ML program that learns from interacting with itself).

Machine Learning Vocabulary
General ML terminologies are different from the terms used in statistical modeling. For example,

- Y variable (dependent variable in regression analysis) is called **target variable** (or tag variable) in ML.
- X variable (independent variable in regression analysis) is known as **feature** in ML.
- In ML terminology, organizing features for ML processing is called **feature engineering**.

Practice: Example 16, Reading 8.

7.4 Machine Learning Algorithms

The following sections provide description of some important models and procedures categorized under supervised and unsupervised learning.

	MODELS AND PROCEDURES	
SUPERVISED LEARNING	• Penalized Regression • CART • Random Forests	Neural Networks
UNSUPERVISED LEARNING	• Clustering Algorithms • Dimension Reduction	

Neural networks are commonly included under supervised learning but are also important in reinforcement learning, which is a part of unsupervised learning.

7.4.1) Supervised Learning

Supervised leaning is divided into two classes based on the nature of the Y variable. These classes are regression and classification. Both classes use different ML techniques.

i. Regression – when the Y variable is continuous – supervised ML techniques

include linear and non-linear models often used for prediction problems.

ii. Classification – when the Y variable is categorical or ordinal – classification techniques include CART (classification and regression trees), random forests and neural networks.

In the following description, for both regression and classification techniques assume
Y = Target variable
$X_1, \ldots X_n$ = n real-valued variables

7.4.1.1) Penalized Regression

Penalized regression is a computationally-efficient method used to solve prediction problems. Penalized regression (imposing penalty on the size of regression coefficients) improves prediction in large datasets by shrinking the number of independent variables and handling model complexity.

Penalized regression and other forms of linear regression, like multiple regression, are classified as special case of the generalized linear model (GLM).

GLM is linear regression in which specification can be changed based on two choices:

1. Maximum number of independent variables the researcher wants to use;
2. How good model fit is needed?

In large datasets, algorithm starts modelling unnecessary complex relationships among many variables, and estimates an output that does not perform well on new data. This problem is called **overfitting**. Penalized regression solves overfitting through **regularization** (by penalizing the statistical variability and magnitude of high dimensional data features). In prediction, parsimonious models (having less parameters) are less subject to overfitting.

Penalized regression is similar to ordinary linear regression with an added penalty which increases as the number of variables increase. The purpose is to regularize the model such that only variables that explain Y should remain in the model. Penalized regressions are usually subject to a trade-off between contribution to model fit versus penalty.

7.4.1.2) Classification and Regression Trees

CART is a common supervised ML method that can be used for predicting classification or regression related modeling issues.

CART model is
- computationally efficient
- adaptable to complex datasets
- usually applied where the target is binary
- useful for interpreting how observations are classified

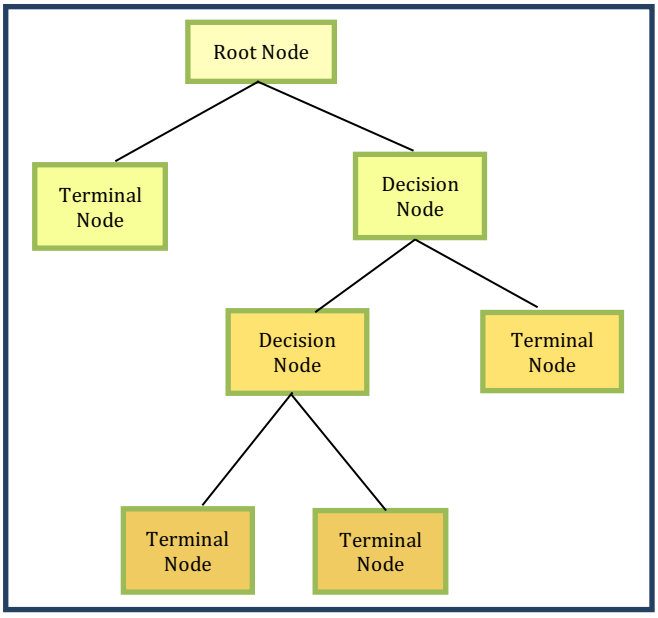
CART model is represented by a binary tree (two-way branches). CART works on a pre-classified training data. Each node signifies a single input variable (X). A classification tree is formed by splitting each node into two distinct subsets, and the process of splitting the derived subsets is repeated in a recursive manner. Process ends when further splitting is not possible (observations cannot be divided into two district groups). The last node is called terminal node that holds a category based on attributes shared by observations at that node.

The chosen cut-off value (splitting value into two groups) is the one that decreases classification error, therefore, observations in each subsequent division have lower error within that group. Some parts of the tree may turn out to be denser (a greater number of splits) while others simpler (a smaller number of splits).

Classification tree vs Regression Tree

i. Classification tree is used when the value of the target variable is categorical.
ii. Regression tree is used when the value of the target variable is continues or numeric.

Classification Tree Example



7.4.1.3.) Random Forests

Random forest, an ML technique that ensembles multiple decision trees together based on random selection of features that contribute more to the process in order to produce accurate and stable prediction. Splitting a node in random forests is based on the ***best*** features from a random subsets of n features. Therefore, each tree marginally varies from other tress.

The power of this model is based on the idea of 'wisdom of crowd' and ensemble learning (using numerous algorithms to improve prediction). All the classifier trees go for classification by majority vote for any new observation.

The involvement of random subsets in the pool of classification trees prevents overfitting problem and also reduces the ratio of noise to signal.

CART and random forest techniques are useful to resolve classification problems in investment and risk management (such as predicting IPO performance, classifying info concerning positive and negative sentiments etc.).

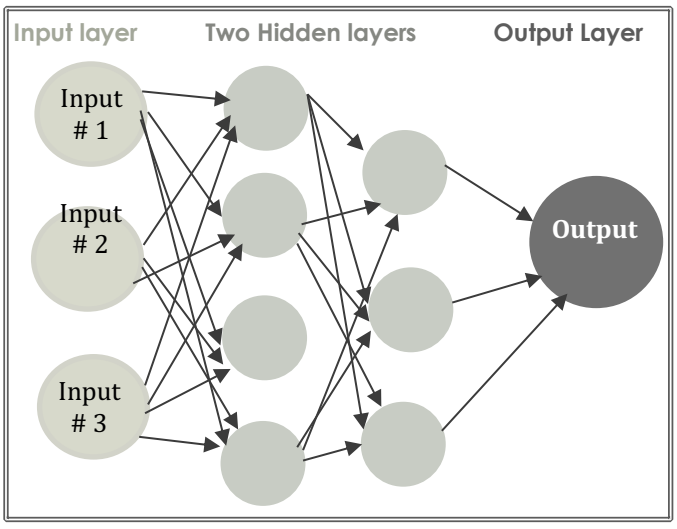
7.4.1.4.) Neural Networks

Neural networks are also known as artificial neural networks, or ANNs. Neural networks are appropriate for nonlinear statistical data and for data with complex connections among variables. Neural networks contain nodes that are linked to the arrows.

ANNs have three types of interconnected layers:
i. an input layer
ii. hidden layers
iii. an output layer

Input layer consists of nodes, and the number of nodes in the input layer represents the number of features used for prediction. For example, the neural network shown below has an input layer with three nodes representing three features used for prediction, two hidden layers with four and three hidden nodes respectively, and an output layer. For a sample network given below, the four numbers – 3,4,3, and 1 – are hyperparameters (variables set by humans that determine the network structure).

Sample: A Neural Network with Two Hidden Layers



Links (arrows) are used to transmit values from one node to the other. Nodes of the hidden layer(s) are called neurons because they process information. Nodes assign weights to each connection depending on the strength and the value of information received, and the weights typically varies as the process advances. A formula (activation function) is applied to inputs, which is generally nonlinear. This allows modeling of complex non-linear functions. Learning (improvement) happens through better weights being applied by neurons. Better weights are identified by improvement in some performance measure (e.g. lower errors). Hidden layer feeds the output-layer.

Deep learning nets (DLNs) are neural networks with many hidden layers (often > 20). Advanced DLNs are used for speech recognition and image or pattern detection.

7.4.2) Unsupervised Learning

In unsupervised ML, we only have input variables and there is no target (corresponding output variables) to which we match the feature set. Unsupervised ML algorithms are typically used for dimension reduction and data clustering.

7.4.2.1) Clustering Algorithms

Clustering algorithms discover the inherent groupings in the data without any predefined class labels. Clustering is different from classification. Classification uses predefined class labels assigned by the researcher.

Two common clustering approaches are:

i) Bottom-up clustering: Each observation starts in its own cluster, and then assemble with other clusters progressively based on some criteria in a non-overlapping manner.

ii) Top-down clustering: All observations begin as one cluster, and then split into smaller and smaller clusters gradually.

The selection of clustering approach depends on the nature of the data or the purpose of the analysis. These approaches are evaluated by various metrics.

K-means Algorithm: An example of a Clustering Algorithm

K-means is a type of bottom-up clustering algorithm where data is partitioned into k-clusters based on the concept of two geometric ideas 'Centroid' (average position of points in the cluster) and 'Euclidian' (straight line distance between two points). The number of required clusters (k-clusters) must have been mentioned beforehand.

Suppose an analyst wants to divide a group of 100 firms into 5 clusters based on two numerical metrics of corporate governance quality. Algorithm will work iteratively to assign suitable group (centroid) for each data point based on similarity of the provided features (in this case two-dimensional corporate governance qualities). There will be five centroid positions (initially located randomly).

Step 1. First step involves assigning each data point its nearest Centroid based on the squared Euclidian distance.

Step 2. The centroids are then recomputed based on the mean location of all assigned data points in each cluster.

The algorithm repeats step 1 and 2 many times until no movement of centroids is possible and sum of squared distances between points is minimum. The five clusters for 100 firms are considered to be optimal when average of squared-line distance between data points from centroid is at minimum.

However, final results may depend on the initial position selected for the centroids. This problem can be addressed by running algorithm many times for different initial positions of the centroids, and then selecting the best fit clustering.

Clustering is a valuable ML technique used for many portfolio management and diversification functions.

7.4.2.2) Dimension Reduction

Dimension reduction is another unsupervised ML technique that reduces the number of random variables for complex datasets while keeping as much of the variation in the dataset as possible.

Principal component analysis (PCA) is an established method for dimension reduction. PCA reduces highly correlated data variables into fewer, necessary, uncorrelated composite variables. Composite variables are variables that assemble two or more highly correlated variables.

The first principal component accounts for major variations in the data, after which each succeeding principal component obtains the remaining volatility, subject to constraint that it is uncorrelated with the preceding principal component. Each subsequent component has lower information to noise ratio. PCA technique has been applied to process stock market returns and yield curve dynamics.

Dimension reduction techniques are applicable to numerical, textual or visual data.

Practice: Example 17, Reading 8.

7.5 Supervised Machine Learning: Training

The process to train ML models includes the following simple steps.

1. Define the ML algorithm.
2. Specify the hyperparameters used in the ML technique. This may involve several training cycles.
3. Divide datasets in two major groups:
   - **Training sample** (the actual dataset used to train the model. Model actually learns from this dataset).
   - **Validation sample** (validates the performance of model and evaluates the model fit for out-of-sample data.)
4. Evaluate model-fit through validation sample and tune the model's hyperparameters.
5. Repeat the training cycles for some given number of times or until the required performance level is achieved.

The output of the training process is the 'ML model'.

The model may overfit or underfit depending on the number of training cycles e.g. model overfitting (excessive training cycles) results in bad out-of-sample predictive performance.

In step 3, the process randomly and repeatedly partition data into training and validation samples. As a result, a data may be labeled as training sample in one split and validation sample in another split. ***'Cross validation'*** is a process that controls biases in training data and improves model's prediction.

Note: Smaller datasets entail more cross validation whereas bigger datasets require less cross-validation.

Practice: End of Chapter Practice Problems for Reading 8.

1. INTRODUCTION TO TIME SERIES ANALYSIS

A time series is any series of data that varies over time e.g. the quarterly sales for a company during the past five years or daily returns of a security.

Time-series models are used to:

1. explain the past
2. predict the future of a time-series

2. CHALLENGES OF WORKING WITH TIMES SERIES

When assumptions of the regression model are not met, we need to transform the time series or modify the specifications of the regression model.

Problems in time series:

1. When the dependent and independent variables are distinct, presence of serial correlation of the errors does not affect the consistency of estimates of intercept or slope coefficients.

But in an autoregressive time-series regression, presence of serial correlation in the error term makes estimates of the intercept (b_0) and slope coefficient (b_1) to be inconsistent.

2. When mean and/or variance of the time series model change over time and is not constant, then using an autoregressive model will provide invalid regression results.

Because of these problems in time series, time series model is needed to be transformed for the purpose of forecasting.

3. TREND MODELS

3.1 Linear Trend Models

In a linear trend model, the dependent variable changes at a ***constant rate*** with time.

$$y_t = b_0 + b_1 t + \varepsilon_t$$

where,

y_t = *value of time series at time t (value of dependent variable)*
b_0 = *y-intercept term*
b_1 = *slope coefficient or trend coefficient*
t = *time, the independent or explanatory variable*
ε_t = *random error term*

The predicted or fitted value of y_t in period 1 is:

$$\hat{y}_1 = \hat{b}_0 + \hat{b}_1(1)$$

The predicted or fitted value of y_t in period 5 is:

$$\hat{y}_5 = \hat{b}_0 + \hat{b}_1(5)$$

The predicted or fitted value of y_t in period T + 1 is:

$$\hat{y}_{T+1} = \hat{b}_0 + \hat{b}_1(T+1)$$

NOTE:

Each consecutive observation in the time series increases by $\hat{b}_1$ in a linear trend model.

Practice: Example 1
Volume 1, Reading 11.

3.2 Log-Linear Trend Models

When time series has exponential growth rate, it is more appropriate to use log-linear trend model instead of linear trend model. Exponential growth rate refers to a ***constant growth*** at a particular rate.

$$y_t = e^{b_0 + b_1 t}$$

where,

t = 1, 2, 3,…,T

Taking natural log on both sides we have:

$\ln y_t = b_0 + b_1 t + \varepsilon_t$

where,

t = 1, 2, 3, …,T

Linear trend model	Log-linear trend model
Predicted trend value of y_t is $\hat{b}_0 + \hat{b}_1 t$,	Predicted trend value of y_t is $e^{\hat{b}_0+\hat{b}_1 t}$ because $e^{\ln y_t} = y_t$.
The model predicts that y_t grows by a ***constant amount*** from one period to the next.	The model predicts a ***constant growth rate*** in y_t of $e^{b_1} - 1$.
A linear trend model is appropriate to use when the residuals from a model are equally distributed above and below the regression line e.g. inflation rate.	A log-linear model is appropriate to use when the residuals of the model exhibit a persistent trend i.e. either positive or negative for a period of time e.g. financial data i.e. stock prices, sales, and stock indices.

Practice: Example 2 & 3
Volume 1, Reading 9.

Limitation of Trend Models: Trend model is based on only one independent variable i.e. time; therefore, it does not adequately incorporate the underlying dynamics of the model.

3.3 Trend Models and Testing for Correlated Errors

In case of presence of serial correlation, both the linear trend model and the log-linear trend model are not appropriate to use. In case of serial correlation, autoregressive time series models represent better forecasting models.

4. AUTOREGRESSIVE (AR) TIME-SERIES MODELS

An autoregressive (AR) model is a time series regression in which the independent variable is a lagged (past) value of the dependent variable i.e.

$$x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$$

First order autoregressive AR (1) for the variable x_t is:

$$x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$$

A pth-order autoregressive AR (p) for the variable x_t is:

$$x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-2} + \ldots.. + b_p x_{t-p} + \varepsilon_t$$

4.1 Covariance-Stationary Series

In order to obtain a valid statistical inference from a time-series analysis, the time series must be covariance stationary.

Time series is covariance stationary when:

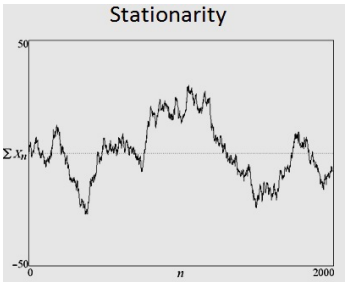
1. The expected value of the time series is constant and finite in all periods.
2. The variance of the time series is constant and finite in all periods.
3. The covariance of the time series with past or future values of itself is constant and finite in all periods.

NOTE:

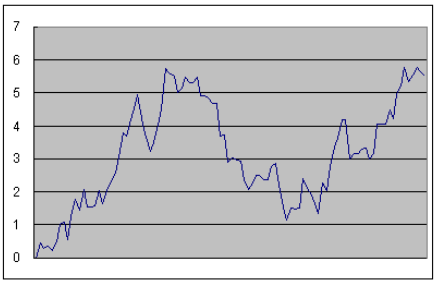
Weakly stationary also refers to covariance stationary.

Stationary Data: When a time series variable does not exhibit any significant upward or downward trend over time.

Nonstationary Data: When a time series variable exhibits a significant upward or downward trend over time.



Non-stationarity (upward trend)



Consequence of Covariance Non-Stationarity: When time series is not covariance stationary, the regression estimation results are invalid because:

- The "*t*-ratios" will not follow a *t*-distribution.
- The estimate of b_1 will be biased and any hypothesis tests will be invalid.

NOTE:

Stationarity in the past does not guarantee stationarity in the future because state of the world may change over time.

4.2 Detecting Serially Correlated Errors in an Autoregressive Model

An Autoregressive model can be estimated using ordinary least squares model (OLS) when the time series is covariance stationary and the errors are uncorrelated.

Detecting Serial Correlation in AR models: In AR models, Durbin-Watson statistic cannot be used to test serial correlation in errors. In such cases, t-test is used.

The autocorrelations of time series refer to the correlations of that series with its own past values.

- When autocorrelations of the error term are zero, the model can be specified correctly.
- When autocorrelations of the error term are significantly different from zero, the model cannot be specified correctly.

Example:

Suppose a sample has 59 observations and one independent variable. Then,

S.D = 1 / $\sqrt{T}$ = 1 / $\sqrt{59}$ = 0.1302
Critical value of t (at 5% significant level with df = 59 – 2 = 57) is 2.

Suppose autocorrelations of the Residual are as follows:

Lag	Autocorrelation	Standard Error	t-statistic*
1	0.0677	0.1302	0.5197
2	-0.1929	0.1302	-1.4814
3	0.0541	0.1302	0.4152
4	-0.1498	0.1302	-1.1507

* **t-statistic = Autocorrelations / Standard Error**

It can be seen from the table that none of the first four autocorrelations has t-statistic > 2 in absolute value.

Conclusion: None of these autocorrelations differ significantly from 0 thus, residuals are not serially correlated and model is specified correctly and OLS can be used to estimate the parameters and the standard errors of the parameters in the autoregressive model.

Correcting Serial Correlation in AR models: The serial correlation among the residuals in AR models can be removed by estimating an autoregressive model by adding more lags of the dependent variable as explanatory variables.

4.3 Mean Reversion

A time series shows mean reversion if it tends to move towards its mean i.e. decrease when its current value is above its mean and increase when its current value is below its mean.

- When a time series equals its mean-reverting value, then the model predicts that the value of the time series will be the same in the next period i.e. $x_{t+1} = x_t$.

Mean reverting level of $x_t = \dfrac{b_0}{1-b_1}$

- Time series will remain the same if its current value $= \dfrac{b_0}{1-b_1}$
- Time series will Increase if its current value $< \dfrac{b_0}{1-b_1}$
- Time series will Decrease if its current value $> \dfrac{b_0}{1-b_1}$

4.4 Multiperiod Forecasts and the Chain Rule of Forecasting

The **chain rule of forecasting** is a process in which a predicted value two periods ahead is estimated by first predicting the next period's value and substituting it into the equation of a predicted value two periods ahead i.e.

The one-period ahead forecast of x_t from an AR (1) model is as follows:

$$\hat{x}_{t+1} = \hat{b}_0 + \hat{b}_1 x_t$$

Two-period ahead forecast is:

$$\hat{x}_{t+2} = \hat{b}_0 + \hat{b}_1 x_{t+1}$$

NOTE:

Multiperiod forecast is more uncertain than single-period forecast because the uncertainty increases when number of periods in the forecast increase.

Example:

The one-period ahead forecast of x_t from an AR (1) model when $x_t = 0.65$ is as follows:

$$\hat{x}_{t+1} = 0.0834 + 0.8665(0.65) = 0.6466$$

Two-period ahead forecast is:

$$\hat{x}_{t+2} = 0.0834 + 0.8665(0.6466) = 0.6437$$

Practice: Example 6
Volume 1, Reading 9.

4.5 Comparing Forecast Model Performance

The accuracy of the model depends on its forecast error variance.

- *The smaller the forecast error variance, the more accurate the model will be.*

In-sample forecast errors: These are the residuals from the fitted time series model i.e. residuals within a sample period.

Out-of-sample forecast errors: These are the residuals outside the sample period. It is more important to have smaller forecast error variance (i.e. high accuracy) for out-of-sample forecasts because the future is always out of sample.

To evaluate the out-of-sample forecasting accuracy of the model, ***Root mean squared error*** (RMSE) is used. RMSE is the square root of the average squared error.

Decision Rule: *The smaller the RMSE, the more accurate the model will be.*

The **RMSE (Root Mean Squared Error)** is used as a criterion for comparing forecasting performance of different forecasting models. To accurately evaluate uncertainty of forecast, both the uncertainty related to the error term and the uncertainty related to the estimated parameters in the time-series model must be considered.

NOTE:

If the model has the lowest RMSE for in-sample data, it does not guarantee that the model will have the lowest RMSE for out-of-sample data as well.

4.6 Instability of Regression Coefficients

When the estimated regression coefficients in one period are quite different from those estimated during another period, this problem is known as instability or nonstationarity.

The estimates of regression coefficients of the time-series model can be different across different sample periods i.e. the estimates of regression coefficients using shorter sample period will be different from using longer sample periods. Thus, sample period selection is one of the important decisions in time series regression analysis.

- Using longer time periods increase statistical reliability but estimates are not stable.
- Using shorter time periods increase stability of the estimates but statistical reliability is decreased.

NOTE:

We cannot select the correct sample period for the regression analysis by simply analyzing the autocorrelations of the residuals from a time-series model. In order to select the correct sample, it is necessary that data should be Covariance Stationary.

5. RANDOM WALKS AND UNIT ROOTS

5.1 Random Walks

A. Random walk without drift: In a random walk without drift, the value of the dependent variable in one period is equal to the value of the series in the previous period plus an unpredictable random error.

$$x_t = x_{t-1} + \varepsilon_t$$

where,

$b_0 = 0$ and $b_1 = 1$.

In other words, the best predictor of the time series in the next period is its current value plus an error term.

The following conditions must hold:

1. Error term has an expected value of zero.
2. Error term has a constant variance.
3. Error term is uncorrelated with previous error terms.

- The equation of a random walk represents a special case of an AR (1) model with $b_0 = 0$ and $b_1 = 1$.
- AR (1) model cannot be used for time series with random walk because random walk has no finite mean, variance and covariance. In random walk $b_0 = 0$ and $b_1 = 1$, so $\frac{b_0}{1 - b_1} = 0 / 0 =$ undefined mean reverting level.
- A standard regression analysis cannot be used for a time series that is random walk.

Correcting Random Walk: When time series has a random walk, it must be converted to covariance-stationary time series by taking the first difference between x_t and x_{t-1} i.e. equation becomes:

$$y_t = x_t - x_{t-1} = \varepsilon_t$$

- Thus, best forecast of y_t made in period t-1 is 0. This implies that the best forecast is that the value of the current time series x_{t-1} will not change in future.

After taking the first difference, the first differential variable y_t becomes covariance stationary. It has $b_0 = 0$ and $b_1 = 0$ and mean reverting level = 0/1 = 0.

- The first differential variable y_t can now be modeled using linear regression.
- However, modeling the first differential variable y_t with an AR (1) model is not helpful to predict the future because $b_0 = 0$ and $b_1 = 0$.

Consequences of Random Walk: When the model has random walk, its R^2 will be significantly high and at the same time changes in dependent variable are unpredictable. In other words, the statistical results of the regression will be invalid.

B. Random walk with a drift: In a random walk with a drift, dependent variable increases or decreases by a constant amount in each period.

$$x_t = b_0 + x_{t-1} + \varepsilon_t$$

where,

$b_0 \neq 0$ and $b_1 = 1$.

By taking first difference,

$$y_t = x_t - x_{t-1} = b_0 + \varepsilon_t$$

NOTE:

All random walks (with & without a drift) have unit roots.

5.2 The Unit Root Test of Nonstationarity

AR (1) time series model will be covariance stationary only when the absolute value of the lag coefficients $b_1<1$. (Note that when b_1 is > 1 in absolute value, it is known as explosive root).

Detecting Random Walk: When time series has random walk, the series does not follow t-distribution and t-test will be invalid. Therefore, t-statistic cannot be used to test the presence of random walk because standard errors in an AR model are invalid if the model has a random walk. Thus, **Dickey-Fuller** test is used to detect nonstationarity:

Method 1: Examining Autocorrelations of the AR model

Stationary Time Series:

- Autocorrelations at all lags equals to zero, or
- Autocorrelations decrease rapidly to zero as the number of lags increases in the model.

Nonstationary time series:

- Autocorrelations at all lags are not equal to zero, or
- Autocorrelations do not decrease rapidly to zero as the number of lags increases in the model.

Method 2: Using Dickey-Fuller Test

Subtracting x_{t-1} from both sides of AR (1) equation we have:

$$x_t - x_{t-1} = b_0 + (b_1 - 1)\, x_{t-1} + \varepsilon_t$$

(or)

$$x_t - x_{t-1} = b_0 + g_1 x_{t-1} + \varepsilon_t$$

where,

$g_1 = (b_1 - 1)$.

- If $b_1 = 1$, then $g_1 = 0$. This implies that there is a unit root in AR (1) model.

Null Hypothesis: H_0: $g_1 = 0 \rightarrow$ time series has a unit root and is Nonstationary

Alternative Hypothesis: H_1: $g_1 < 0 \rightarrow$ time series does not have a unit root and is Stationary

- t-statistic is calculated for predicted value of g_1 and critical values of t-test are computed from Dickey-Fuller test table (these critical t-values in absolute value > than typical critical t-values).

Practice: Example 12
Volume 1, Reading 11.

6. MOVING-AVERAGE TIME SERIES MODELS

Moving average (MA) is different from AR model. MA is an average of successive observations in a time series. It has lagged values of residuals instead of lagged values of dependent variable.

6.1 Smoothing Past Values with an n-Period Moving Average

n-period moving average is used to smooth out the fluctuations in the value of a time series across different time periods.

$$\frac{x_t + x_{t-1} + x_{t-2} + \ldots + x_{t-(n-1)}}{n}$$

Drawbacks of Moving Average:

- It is biased towards large movements in the actual data.
- It is not the best predictor of the future.
- It gives equal weights to all the periods in the moving average.

Distinguishing AR time series from a MA time series:

- Autocorrelations of most AR (p) time series start large and decline gradually.
- Autocorrelations of MA (q) time series suddenly drop to 0 after the first q autocorrelations.

7. SEASONALITY IN TIME-SERIES MODELS

When a time series variable exhibit a repeating patterns at regular intervals over time, it is known as seasonality e.g. sales in Dec. > sales in Jan. A time series with seasonality also has a non-constant mean and thus is not covariance stationary.

Detecting seasonality: In case of seasonality in the data, autocorrelation in the model differs by season. For example, in case of quarterly sales data of a company, if the fourth autocorrelation of the error term differs significantly from 0 → it is a sign of seasonality in the model.

Decision Rule: When t-statistic of the fourth lag of autocorrelations of the error > critical t-value → reject null hypothesis that fourth autocorrelations is 0. Thus, there is seasonality problem.

Correcting Seasonality: This problem can be solved by adding seasonal lags in an AR model i.e. after including a seasonal lag in case of quarterly sales data, the AR model becomes:

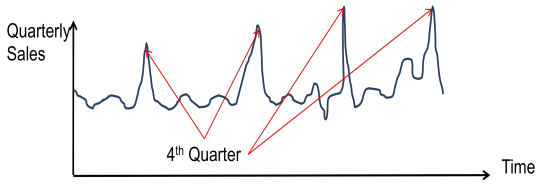
$$x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-4} + \varepsilon_t$$

In case of monthly sales data, the AR model becomes:

$$x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-12} + \varepsilon_t$$

NOTE:

R^2 of the model without seasonal lag will be less than the R^2 of the model with seasonal lag. This implies that when time series exhibit seasonality, including a seasonal lag in the model improves the accuracy of the model.



Practice: Example 15
Volume 1, Reading 9.

8. AUTOREGRESSIVE MOVING-AVERAGE MODELS (ARMA)

An ARMA model combines both autoregressive lags of the dependent variable and moving-average errors.

Drawbacks of ARMA model:

- Parameters of ARMA models are usually very unstable.
- ARMA models depend on the sample used.
- Choosing the right ARMA model is a difficult task because it is more of an art than a science.

9. AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTICITY MODELS (ARCH)

When regression model has (conditional) heteroskedasticity i.e. variance of the error in a particular time-series model in one period depends on the variance of the error in previous periods, standard errors of the regression coefficients in AR, MA or ARMA models will be incorrect, and hypothesis tests would be invalid.

ARCH model:

ARCH model must be used to test the existence of conditional heteroskedasticity. An ARCH (1) time series is the one in which the variance of the error in one period depends on size of the squared error in the previous period i.e. if a large error occurs in one period, the variance of the error in the next period will be even larger.

To test whether time series is ARCH (1), the squared residuals from a previously estimated time-series model are regressed on the constant and first lag of the squared residuals i.e.

$$\hat{\varepsilon}_t = \alpha_0 + \alpha_1 \hat{\varepsilon}_{t-1}^2 + \mu_t$$

where,

μ_t is an error term

Decision Rule: If the estimate of α_1 is statistically significantly different from zero, the time series is ARCH (1). If a time-series model has ARCH (1) errors, then the variance of the errors in period t+1 can be predicted in period t using the formula:

$$\hat{\sigma}_{t+1}^2 = \hat{\alpha}_0 + \alpha_1 \hat{\varepsilon}_t^2$$

Consequences of ARCH:

- Standard errors for the regression parameters will not be correct.
- When ARCH exists, we can predict the variance of the error terms.

Generalized least squares or other methods that correct for heteroskedasticity must be used to estimate the correct standard error of the parameters in the time-series model.

Autoregressive model versus ARCH model:

- Using AR (1) model implies that model is correctly specified.
- Using ARCH (1) implies that model can not be correctly specified due to existence of conditional heteroskedasticity in the residuals; therefore, ARCH (1) model is used to forecast variance/volatility of residuals.

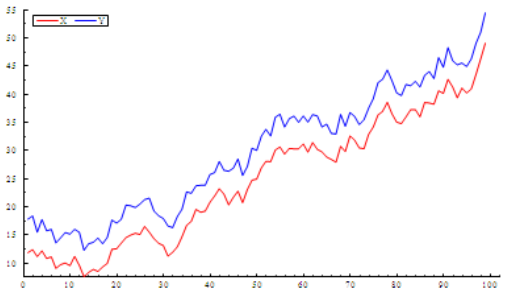
10. REGRESSIONS WITH MORE THAN ONE TIME SERIES

1. When neither of the time series (dependent & independent) has a unit root, linear regression can be used.
2. One of the two time series (i.e. either dependent or independent but not both) has a unit root, we should not use linear regression because error term in the regression would not be covariance stationary.
3. If both time series have a unit root, and the time series are not cointegrated, we cannot use linear regression.
4. If both time series have a unit root, and the time series is cointegrated, linear regression can be used. Because, when two time series are cointegrated, the error term of the regression is covariance stationary and the t-tests are reliable.

Cointegration: Two time series are cointegrated if

- A long term financial or economic relationship exists between them.
- They share a common trend i.e. two or more variables move together through time.

Two Cointegrated Time Series



NOTE:

Cointegrated regression estimates the long-term relation between the two series. Therefore, it is not the best model of the short-term relation between the two series.

Detecting Cointegration: The Engle-Granger Dickey-Fuller test can be used to determine if time series are cointegrated.

Engle and Granger Test:

1. Estimate the regression $y_t = b_0 + b_1 x_t + \varepsilon_t$
2. Unit root in the error term is tested using Dickey-fuller test but the critical values of the Engle-Granger are used.
3. If test fails to reject the null hypothesis that the error term has a unit root, then error term in the regression is not covariance stationary. This implies that two time series are not cointegrated and regression relation is spurious.
4. If test rejects the null hypothesis that the error term has a unit root, then error term in the regression is covariance stationary. This implies that two time series are cointegrated and regression results and parameters will be consistent.

NOTE:

- When the first difference is stationary, series has a single unit root. When further differences are required to make series stationary, series is referred to have multiple unit roots.
- For multiple regression model, rules and procedures for unit root and stationarity are the same as that of single regression.

12. SUGGESTED STEPS IN TIME-SERIES FORECASTING

Following is a guideline to determine an accurate model to predict a time series.

1. Select the model on the basis of objective i.e. if the objective is to predict the future behavior of a variable based on the past behavior of the same variable, use Time series model and if the objective is to predict the future behavior of a variable based on assumed casual relationship with other variables Cross sectional model should be used.

2. When time-series model is used, plot the series to detect Covariance Stationarity in the data. Trends in the time series data include:

- A linear trend
- An exponential trend
- Seasonality
- Structural change i.e. a significant shift in mean or variance of the time series during the sample period

3. When there is no seasonality or structural change found in the data, linear trend or exponential trend is appropriate to use i.e.

i. Use linear trend model when the data plot on a straight line with an upward or downward slope.
ii. Use log-linear trend model when the plot of the data exhibits a curve.
iii. Estimate the regression model.
iv. Compute the residuals
v. Use Durbin-Watson statistic to test serial correlation in the residual.

4. When serial correlation is detected in the model, AR model should be used. However, before using AR model, time series must be tested for Covariance Stationarity.

- If time series has a linear trend and covariance nonstationary; it can be transformed into covariance stationary by taking the first difference of the data.
- If time series has exponential trend and covariance nonstationary; it can be transformed into covariance stationary by taking natural log of the time series and then taking the first difference.
- If the time series exhibits structural change, two different time-series model (i.e. before & after the shift) must be estimated.
- When time series exhibits seasonality, seasonal lags must be included in the AR model.

5. When time series is converted into Covariance Stationarity, AR model can be used i.e.

- Estimate AR (1) model;
- Test serial correlation in the regression errors; if no serial correlation is found only then AR (1) model can be used. When serial correlation is detected in AR (1), then AR (2) should be used and tested for serial correlation. When no serial correlation is found, AR (2) can be used. If serial correlation is still present, order of AR is gradually increasing until all serial correlation is removed.

6. Plot the data and detect any seasonality. When seasonality is present, add seasonal lags in the model.

7. Test the presence of autoregressive conditional heteroskedasticity in the residuals of the model i.e. by using ARCH (1) model.

8. In order to determine the better forecasting model, calculate out-of-sample RMSE of each model and select the model with the lowest out-of-sample RMSE.

Practice: End of Chapter Practice Problems for Reading 9 & FinQuiz Item-set ID# 11585.

1. INTRODUCTION

There are three major probabilistic approaches or techniques that are used to assess risk:

1) Scenario analysis: It employs probabilities to a small number of possible outcomes. It helps to assess the effects of discrete risk.
2) Decision trees: It employs tree diagrams of possible outcomes. It helps to assess the effects of discrete risk.
3) Simulations: It involves generating a unique set of cash flows and value by using a number of possible outcomes from different sets of distribution. In this method, distributions of values are estimated for each parameter in the analysis (growth, market share, operating margin, beta, etc.). It helps to assess the effects of continuous risk. Hence, it is a more flexible approach than scenario analysis and decision trees.

2. SIMULATIONS

2.1 Steps in Simulation

Following are the steps used in running a simulation:

1) **Determine "probabilistic" variables:** The first step in running a simulation is determining variables for which distributions are estimated. Unlike scenario analysis and decision trees, where the number of variables and the potential outcomes associated with them are limited in number, in simulation, there is no constraint on how many variables can be used. However, it is preferred to use only those variables that have a significant impact on value.
2) **Define probability distributions for these variables:** Once variables are determined, we define probability distributions of these variables. There are three ways in which probability distributions can be defined:

   a) ***Historical data:*** Historical data can be used to determine probability distributions for variables that have a long history and reliable data over that history.
   b) ***Cross sectional data:*** Cross sectional data can be used to determine probability distributions for variables for which data on differences in those variable across existing investments that are similar to the investment being analysed is available. E.g. a distribution of pre-tax operating margins across manufacturing companies in 2014.
   c) ***Statistical distribution and parameters***: When historical and cross sectional data for a variable is insufficient or unreliable, then we can use a statistical distribution to analyse variability in the input and estimate the parameters for that distribution. E.g. using statistical distribution, we can conclude that operating margins will be distributed uniformly, with a minimum of 5% and a maximum of 10%, and that revenue growth is normally distributed with an expected value of 7% and a standard deviation of 5%.

It is difficult to determine the right distribution and the parameters for the distribution using statistical distribution due to two reasons.

i. Practically, there are few inputs that may meet stringent requirements demanded by statistical distributions, e.g. revenue growth cannot be normally distributed because the lowest value it can take on is –100%.
ii. Even if distribution has been determined, parameters are needed to be estimated.

It is important to note that for some inputs, probability distributions are discrete and for some inputs, they are continuous. Similarly, for some inputs, probability distributions are based upon historical data and for others, they are based on statistical distributions.

3) **Check for correlation across variables:** After defining probability distribution, correlations across variables must be checked. E.g. interest rates and inflation are correlated with each other i.e., high inflation is usually accompanied by high interest rates. When there is strong correlation (positive or negative) across inputs, then we can use only one of the two

inputs, preferably the one which has greater impact on value.

4) **Run the simulation:** When we run the first simulation, one outcome is drawn from each distribution and the value is computed based upon those outcomes. This process is repeated as many times as desired to get a set of values. It is important to note that marginal contribution of each simulation decreases with the increase in number of simulations.

The number of simulations that can be run is determined by the following:

i. **Number of probabilistic inputs:** If there are larger number of inputs that have probability distributions, then the required number of simulations will also be greater.
ii. **Characteristics of probability distributions:** If the distributions in an analysis are more diverse (e.g., some inputs have normal distributions, some have historical data distributions, while some have discrete) then the number of required simulations will be greater.
iii. **Range of outcomes:** The greater the potential range of outcomes on each input, the greater will be the number of simulations.

Note: Practically, it is preferred to run as many simulations as possible.

Impediments to good simulations: Following are the two constraints in running good simulations that have been eased in recent years.

1) Informational impediments: Due to lack of information, it is difficult to estimate distributions of values for each input into a valuation.
2) Computational impediments: Simulations tend to be too time and resource intensive for the typical analysis if it is run on personal computers rather than specialized computers.

2.2 An Example of a Simulation

A company, (say ABC), analyses dozens of new home improvement stores every year. It also has hundreds of stores in operation at different stages of their life cycles.

Suppose ABC is analysing a new home improvement store that will be like other traditional stores. ABC needs to make several estimates for analysing a new store, e.g. revenues at the store. Given that the ABC's store sizes are similar across locations, the firm can get an estimate of the expected revenues by looking at revenues at their existing stores.

To run a simulation of the ABC's store's cash flows and value, we will make the following assumptions:

- *Base revenues:* The estimate of the base year's revenues will be taken as a base. We will assume that revenue will be normally distributed with an expected value of $44 million and a standard deviation of $10 million.
- *Pre-tax operating margin:* The pre-tax operating margin is assumed to be uniformly distributed with a minimum value of 6% and a maximum value of 12%, with an expected value of 9%. Non-operating expenses are anticipated to be $1.5 million a year.
- *Revenue growth:* A slightly modified version of the actual distribution of historical real GDP changes can be used as the distribution of future changes in real GDP. Suppose, the average real GDP growth over the period is 3%; during worst year a drop in real GDP of more than 8% was observed and during best year, an increase of more than 8% was observed. The expected annual growth rate in revenues is the sum of the expected inflation rate and the growth rate in real GDP. Assume that the expected inflation rate is 2%.

The store is expected to generate cash flows for 10 years and there is no expected salvage value from the store closure.

The cost of capital for the ABC is 10% and the tax rate is 40%.

We can compute the value of this store to the ABC, based entirely upon the expected values of each variable:

Expected base-year revenue = $44 million

Expected base-year after-tax cash flow = (Revenue × Pre tax margin – Non-operating expenses)(1 – Tax rate) = (44 × 0.09 – 1.5)(1 – 0.4) = $1.476 million

Expected growth rate = GDP growth rate + Expected inflation = 3% + 2% = 5%

Value of store =

$$CF(1+g)\frac{\left(1-\frac{(1+g)^n}{(1+r)^n}\right)}{r-g} = 1.476(1.05)\frac{\left(1-\frac{1.05^{10}}{1.10^{10}}\right)}{0.10-0.05} = \$11.53 \text{ million}$$

Risk-adjusted value for this store = $11.53 million

Suppose, a simulation is run 10,000 times, based upon the probability distributions for each of the inputs. The key statistics on the values obtained across the 10,000 runs are summarized below:

- Average value across the simulations is $11.67 million;
- The median value is $10.90 million;
- Lowest value across all runs is −$5.05 million;
- Highest value is $39.42 million;
- Standard deviation in values is $5.96 million.

2.3 Use in Decision Making

1) **Better input estimation:** Ideally in simulations, an analyst examines both historical and cross sectional data on each input variable before deciding which distribution to use and the parameters of the distribution.
2) **Simulation yields a distribution for expected value rather than a point estimate:** Simulation generates expected value of $11.67 million for the store as well as a distribution for expected value as it estimates standard deviation of $5.96 million in that value and a breakdown of the values, by percentile.

Advantages of Simulations:

- Simulations yield better estimates of expected value than conventional risk-adjusted value models.
- Simulations lead to better decisions because they provide estimates of the expected value and the distribution in that value.

2.4 Simulations with Constraints

1) **Book Value Constraints:** There are two types of constraints on book value of equity that may demand risk hedging.
   i. ***Regulatory Capital Restrictions:*** Financial service firms (i.e. banks and insurance companies) are required to maintain book equity as a fraction of loans or other assets at or above a floor ratio specified by the authorities. Value at risk, or VAR, is a measure used by financial service firms to understand the potential risks in their investments and to evaluate the likelihood of a catastrophic outcome. Simulations approach can be used to simulate the values of investments under a variety of scenarios in order to identify the possibility of falling below the regulatory ratios as well as hedging against this event occurring.
   ii. ***Negative Book Value for Equity:*** In some countries, a negative book value of equity can create substantial costs for the firms and its investors. E.g. in Europe, firms with negative book values of equity are required to raise fresh equity capital to bring their book values above zero; in some countries in Asia, firms with negative book values of equity are not allowed to pay dividends. Simulations can be used to assess the probability of a negative book value for equity and to hedge against it.
2) **Earnings and Cash Flow Constraints:** Earnings and cash flow constraints can be either internally or externally imposed. E.g. negative consequences of reporting a loss or not meeting analysis estimates of earnings, loan covenants (i.e. interest rate on loan) can be related to earnings outcomes. For such constraints, simulations can be used to both assess the likelihood of violations of these constraints and to analyse the effect of risk hedging products on this likelihood.
3) **Market Value Constraints:** Simulations can be used to quantify the likelihood of distress and its impact on expected cash flows and discount rates as well as to build in the cost of indirect bankruptcy costs into valuation by comparing the value of a business to its outstanding claims in all possible scenarios (rather than just the most likely one).

2.5 Issues

Following are some key issues associated with using simulations in risk assessment:

1) **Garbage in, garbage out:** The distributions chosen for the inputs in simulation should be based upon analysis and data, rather than guesswork, in order to have meaningful results. Simulations may yield great-looking output even when the inputs are selected randomly. Simulations also require having an adequate knowledge of statistical distributions and their characteristics.
2) **Real data may not fit distributions:** In real world, the data rarely follows the stringent requirements of statistical distributions. This implies that if we use probability distributions for any data that does not resemble the true distribution underlying an input variable, it will give misleading results.
3) **Non-stationary distributions:** Shifts in market structure may lead to change in distributions (i.e. non-stationary distributions) either by changing the form of the distribution or by changing the parameters of the distribution. This implies that mean and variance estimated from historical data for an input that is normally distributed may change for the next period if there is change in market structure. Hence, it is preferred to use forward looking probability distributions.
4) **Changing correlation across inputs:** Correlation across input variables can be modelled into simulations only if the correlations remain stable and predictable. If the correlations between input

variables change over time, it becomes far more difficult to model them.

2.6 Risk-Adjusted Value and Simulations

In simulations, the cash flows generated are expected cash flows which are not adjusted for risk. Hence, they should be discounted using a risk-adjusted rate rather than risk-free rate. The standard deviation in values from a simulation can be used as a measure of investment or asset risk. If standard deviation is used as a measure of risk for making investment decision, it is not appropriate to use a risk-adjusted discount rate as it will result in a double counting of risk.

Suppose, we have to choose between two assets, both of which are valued using simulations and risk-adjusted discount rates. The result of simulations is as follows:

Asset	Risk-Adjusted Discount Rate	Simulation Expected Value	Simulation Standard Deviati
A	12%	$100	15%
B	15%	$100	21%

- Asset B is considered to be riskier due to its greater standard deviation and a higher discount rate is used to compute value. If Asset B is rejected because of its higher standard deviation, then we would be penalizing it twice. Hence, the correct way is to run simulation using the risk-free rate as the discount rate for both assets.
- It is important to understand that if selection decision regarding assets is made on the basis of their standard deviation in simulated values, it is assumed that in investment decision making, total risk matters rather than focusing on only the non-diversifiable risk. This implies that an asset with high standard deviation in simulated values may result in little additional risk when added to a portfolio compared to considering it on stand-alone basis because much of its risk can be diversified away.
- The stock which has less volatile value distribution may be considered a better investment than another stock with a more volatile distribution.

3. AN OVERALL ASSESSMENT OF PROBABILISTIC RISK ASSESSMENT APPROACHES

3.1 Comparing the Approaches

Decision regarding which probabilistic approach to use for assessing risk depends upon how an analyst plan to use the output and what types of risk are faced by him:

1) Selective versus full risk analysis:

- In scenario analysis, we can analyse limited scenarios (e.g. the best case, the most likely case, and the worst case) and therefore, we cannot complete assessment of all possible outcomes from risky investments or assets. In contrast, in decision trees and simulations, all possible outcomes can be considered.
  - In decision trees, all possible outcomes can be captured by converting continuous risk into a manageable set of possible outcomes.
  - In simulations, all possible outcomes can be captured by using probability distributions.
  - In scenario analysis, the sum of the probabilities of the scenarios can be less than one, whereas the sum of the probabilities of outcomes in decision trees and simulations must be equal to one. This implies that in decision trees and simulations, expected values across outcomes can be estimated using the probabilities as weights, and these expected values are comparable to the single estimate risk-adjusted values calculated using discounted cash flow and relative valuation models.

2) Type of risk: Scenario analysis and decision trees are used to assess the impact of discrete risk whereas simulations are used to assess the impact of continuous risks. This implies that when risks occur concurrently, then scenario analysis is easier to use. When risks is sequential (i.e. occur in phases), decision trees are preferred to use.

3) Correlation across risks: In scenario analysis, correlations can be incorporated into the analysis subjectively by creating scenarios, e.g. the high (low) interest rate scenario will also include slower (higher) economic growth. In simulations, correlated risks can be explicitly modelled. However, it is difficult to model correlated risks in decision trees.

Risk Type and Probabilistic Approaches			
Discrete / Continuous	**Correlated / Independent**	**Sequential / Concurrent**	**Risk Approach**
Discrete	Independent	Sequential	Decision Tree
Discrete	Correlated	Concurrent	Scenario Analysis
Continuous	Either	Either	Simulations

4) **Quality of the information:** Simulations are preferred to use when there is substantial historical and cross sectional data available that can be used to generate probability distributions and parameters. Decision trees are appropriate to use when risks can be assessed either using past data or population characteristics because in decision trees we need estimates of the probabilities of the outcomes at each chance node. Hence, mostly scenario analysis is used when assessing new and unpredictable risks.

3.2 Complement or Replacement for Risk-Adjusted Value

- Both decision trees and simulations are approaches that can be used as either, complements to or substitutes for risk-adjusted value. In contrast, Scenario analysis will always be a complement to risk-adjusted value, since it does not capture all possible outcomes.
- Decision trees, simulations, and scenario analysis use expected rather than risk-adjusted cash flows and the risk-adjusted discount rate.
- In all three approaches, the risk-adjusted discount rate can be changed for different outcomes because all of these three approaches provide a range for estimated value and a measure of variability (in terms of value at the end nodes in a decision tree or as a standard deviation in value in a simulation). It is important to note that it is inappropriate to discount cash flows of risky investments at a risk-adjusted rate (in simulations and decision trees) and then reject them on the basis of their high variability.

3.3 In Practice

With ease in data availability and computing power, the use of probabilistic approaches has become more common. Because of this, simulations can now be implemented in a variety of new markets as discussed below.

1) **Deregulated electricity markets:** With increasing number of deregulations in electricity markets, companies involved in the business of buying and selling electricity have started using simulation models to quantify the changes in demand and supply of power, and the resulting price volatility in order to determine how much should be spent on building new power plants and how to use the excess capacity in these plants.
2) **Commodity companies:** Companies in commodity businesses (e.g. oil and precious metals) have started using probabilistic approaches to examine how much they should bid for new sources for these commodities, rather than making decision on a single best estimate of the future price.
3) **Technology companies:** Simulations and scenario analyses are now being used to model the effects of the entry and diffusion of new technologies on revenues and earnings.