

# What Is Fintech

## ➤ Basic concept of Fintech ✓

- ✗ • Broadly refers to technology – driven innovation occurring in the financial services industry ( in its broadest sense );
- ✗ • Narrowly refers to technological innovation in the design and delivery of financial services and products ( for the purpose of this reading ).

## ➤ The stages in development of Fintech

- **Early form:** Data processing and the automation of routine tasks;
- **Then followed system:** Decision-making applications based on complex machine-learning logic, where computer programs are able to "learn" how to complete tasks over time.

1024

B  
KB      txt  
MB      ~~GB~~

GB      128 MB

TB      ~~HRD~~ HD

P B

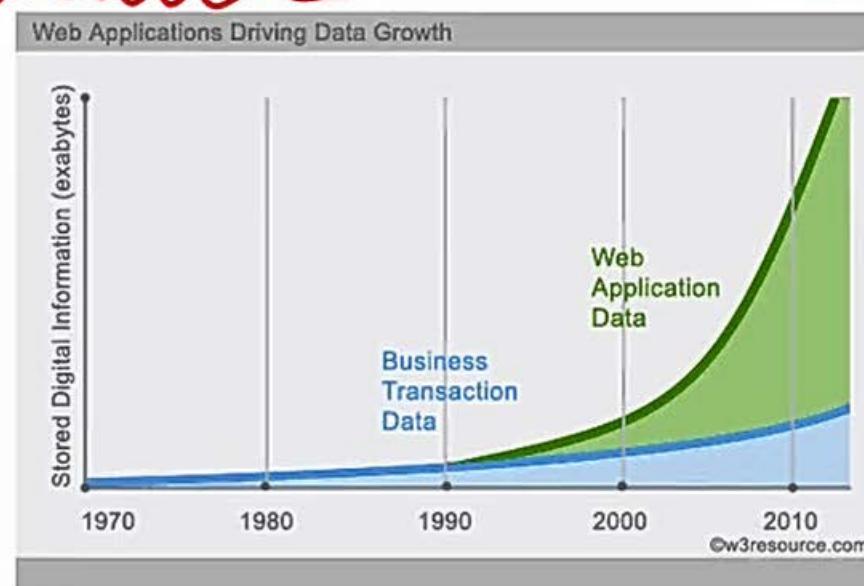
EB

# ◆ What Is Fintech

## ➤ Analysis of large datasets

- In addition to growing amounts of traditional data, massive amounts of alternative data generated from non-traditional data sources

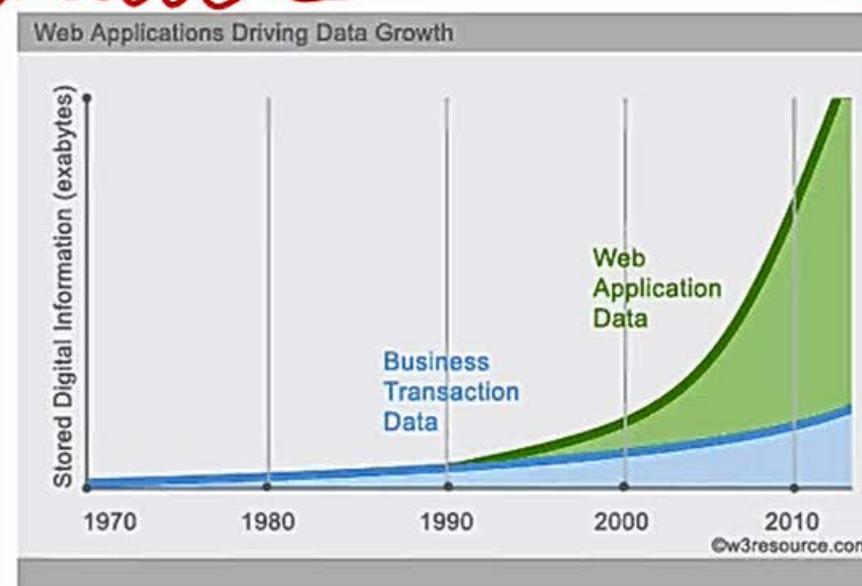
- ✓ **Traditional data source** : security prices, corporate financial statements, and economic indicators
- ✓ **Non-traditional data source** : social media sensor networks



# ◆ What Is Fintech

## ➤ Analysis of large datasets

- In addition to growing amounts of traditional data, massive amounts of alternative data generated from non-traditional data sources
  - ✓ **Traditional data source** : security prices, corporate financial statements, and economic indicators
  - ✓ **Non-traditional data source** : social media sensor networks



# What Is Fintech

## ➤ Analytical tools

- Artificial Intelligence (AI) – computer systems capable of performing tasks that previously required human intelligence.

## ➤ Automated trading

- Computer algorithms or automated trading applications may provide a number of benefits to investors

- ✓ more efficient trading
- ✓ lower transaction costs
- ✓ anonymity 匿名
- ✓ greater access to market liquidity

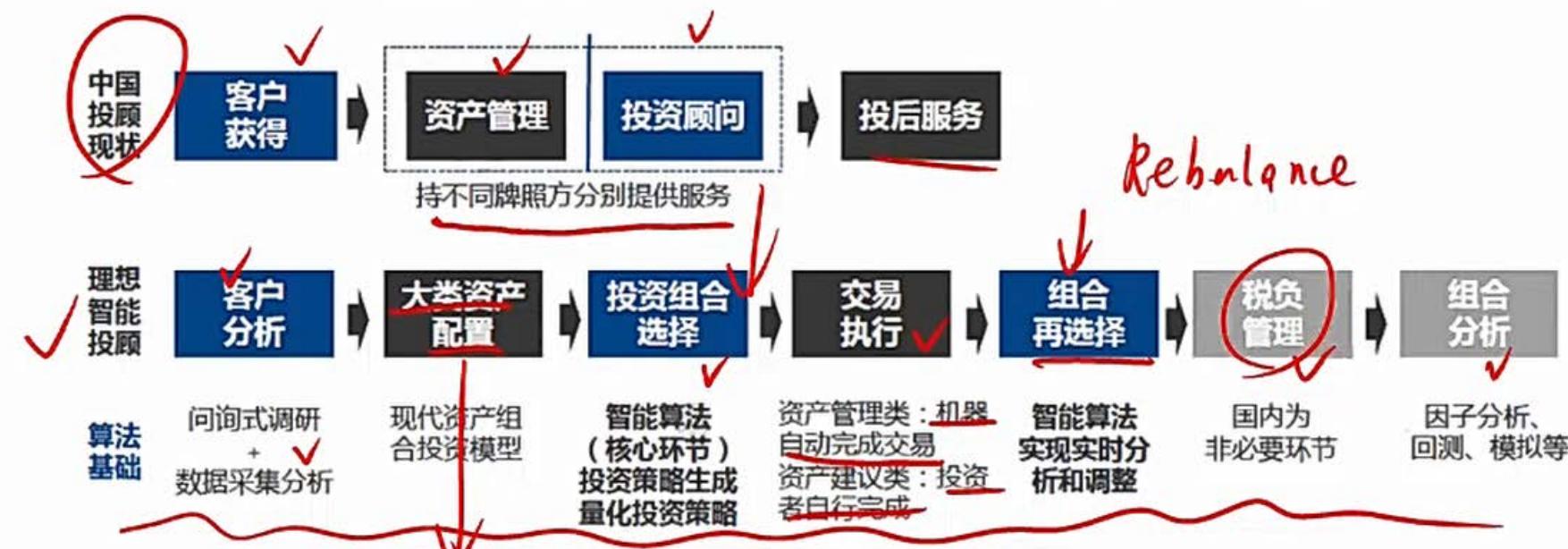
交易量 ↑

# ◆ What Is Fintech

智能投顾

## ➤ Automated advice

- Robo-advisers or automated personal wealth management services
- To provide investment services to retail investors at lower cost

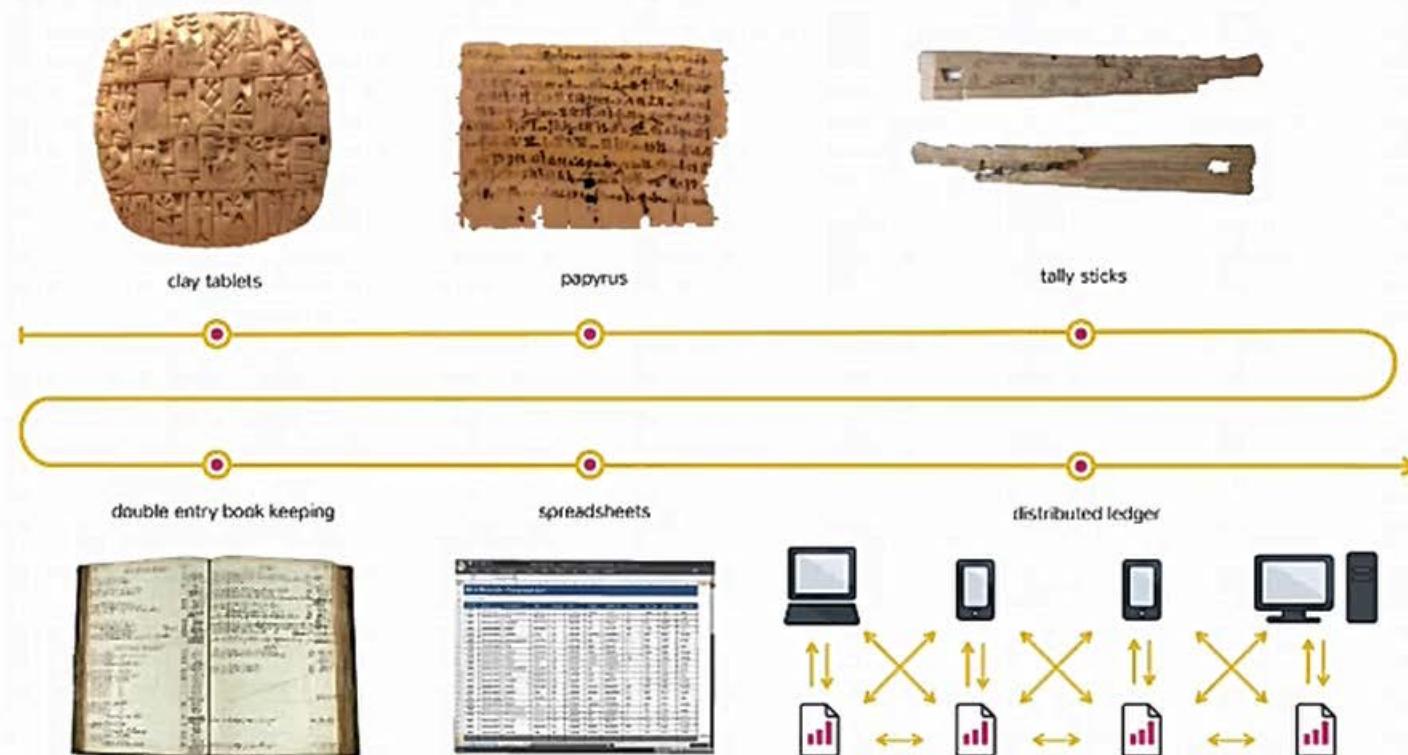


\* 资料来源：《中国智能投顾市场发展趋势研究报告》，慧辰资讯TMT互联网研究部

CML/EF

## ➤ Financial record keeping

- New technology, such as Distributed Ledger Technology ( DLT ), may provide secure ways to track ownership of financial assets on a peer-to-peer ( P2P ) basis, such as Bitcoin.



# ◆ What Is Fintech

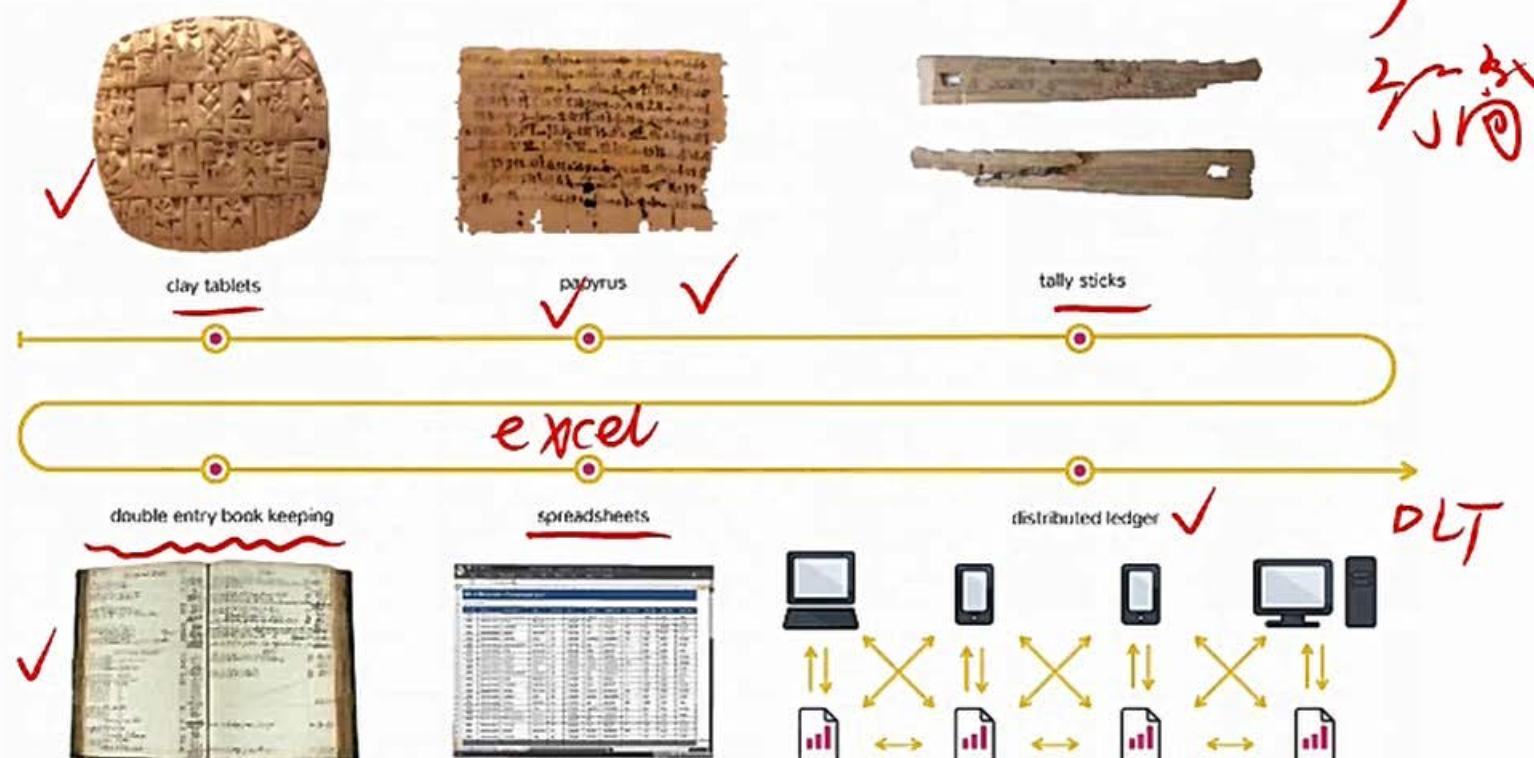
## ➤ Financial record keeping

- New technology, such as Distributed Ledger Technology ( DLT ), may provide secure ways to track ownership of financial assets on a peer-to-peer ( P2P ) basis, such as Bitcoin.

多人同時更新

數據實時同步

簡



# Example

- A correct description of fintech is that it:
- A. is driven by rapid growth in data and related technological advances.
  - B. increases the need for intermediaries.
  - C. is at its most advanced state using systems that follow specified rules and instructions.

Answer:

A is correct. Drivers of fintech include extremely rapid growth in data (including their quantity, types, sources, and quality) and technological advances enabling the capture and extraction of information from it.

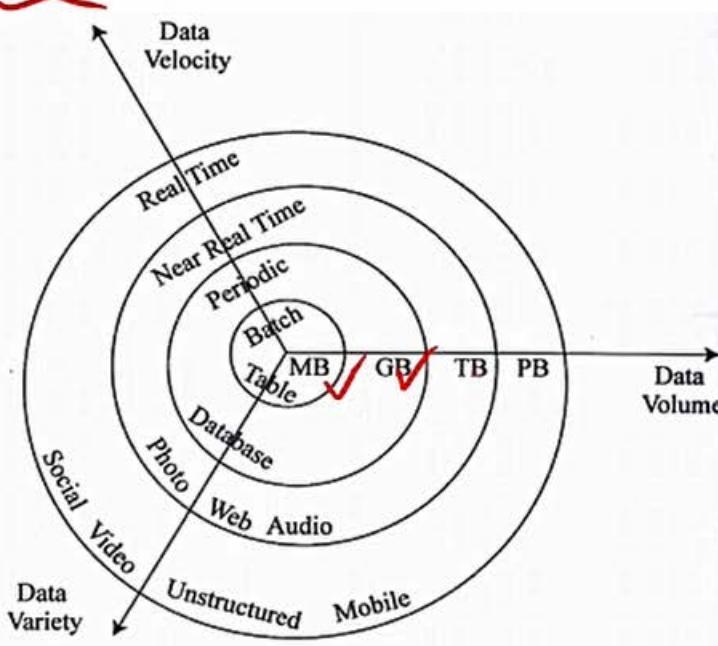
# Big Data ✓

## ➤ Definition

- The term **Big Data** refers to **the vast amount** of data being generated by industry, governments, individuals, and electronic devices, including data generated from **traditional sources** as well as **non-traditional data types** ( also known as **alternative data** )

## ➤ The characteristics of Big Data

- Volume ( very large )
- Velocity ( real-time or near-real-time )
- Variety (mainly unstructured )



3V

LEVEL I V4 R43



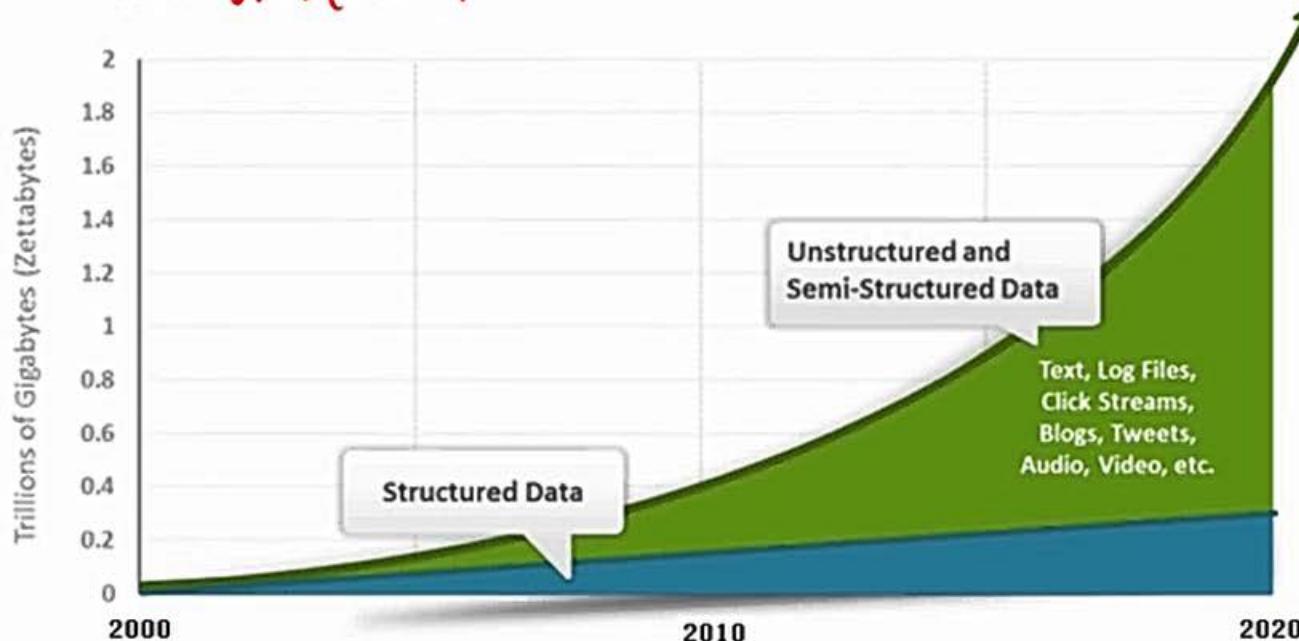
# Big Data Structured Query language

## ➤ Structured, semi-structured and unstructured data

~~traditional~~ Structured data : SQL tables or CSV files

- Semi-structured data : HTML code
- Unstructured data : video message, blogs, WeChat messages

~~non-traditional~~



# Big Data

## ➤ Three main sources of alternative data

- individuals
- business processes: Including direct sales information, such as credit card data, as well as corporate exhaust.
- Sensors: Sensor data are collected from such devices as smart phones, cameras, RFID chips, and satellites that are usually connected to computers via wireless networks.

Individuals	Business Processes	Sensors
Social media	Transaction data	Satellites
News, reviews	Corporate data	Geolocation
Web searches, personal data		Internet of Things
		Other sensors

NFC

物联网

# Big Data

## ➤ Big Data challenges

- Big Data poses several challenges when it is used in investment analysis, including **the quality, volume, and appropriateness of the data.**
- The data must be sourced, cleansed, and organized before analysis can occur. This process can be **extremely difficult** with alternative data owing to the unstructured characteristics of the data involved.



# Data Science

## ➤ Data processing methods

- Capture
  - ✓ how the data are collected and transformed into a format that can be used by the analytical process.
- Curation
  - ✓ Data curation refers to the process of ensuring data quality and accuracy through a data cleaning exercise.
- Storage
  - ✓ Data storage refers to how the data will be recorded, archived, and accessed and the underlying database design.
- Search
  - ✓ Search refers to how to query data.
- Transfer
  - ✓ Transfer refers to how the data will move from the underlying data source or storage location to the underlying analytical tool.



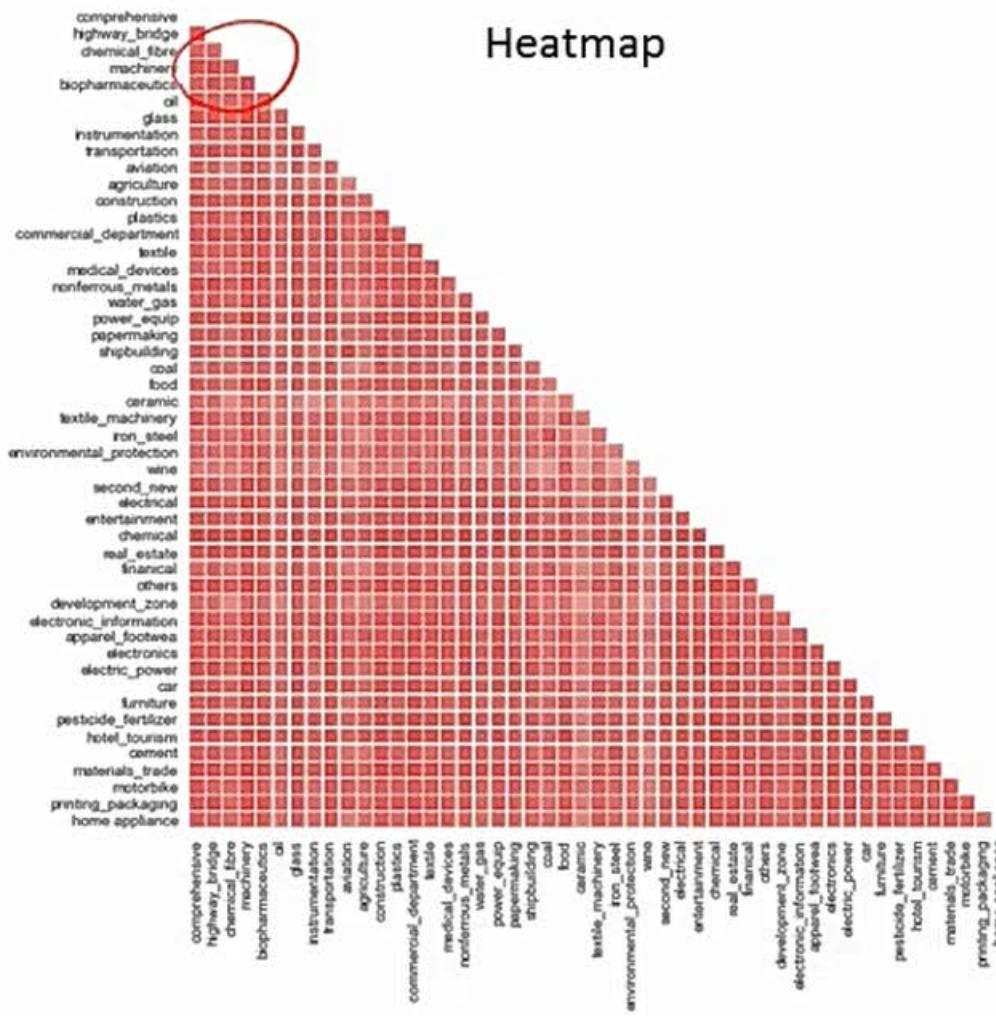
# Data Science

## ➤ Data visualization

- Visualization refers to how the data will be formatted, displayed, and summarized in graphical form.
- Traditional structured data can be visualized using tables, charts, and trends, whereas non-traditional unstructured data require new techniques of data visualization.

# Data Science

## ➤ Data visualization



1

✓

10  
0.9  
0.8  
0.7  
0.6  
0.5

# Example

➤ A characteristic of Big Data is that:

- A. One of its traditional sources is business processes. X
- B. It involves formats with diverse types of structures. Variety
- C. Real-time communication of it is uncommon due to vast content.

Answer:

B is correct. Big Data is collected from many different sources and is a variety of formats, including structured data ( e.g., SQL tables or CSV files ), semi-structured data ( e.g., HTML code ), and unstructured data ( e.g., video messages ).

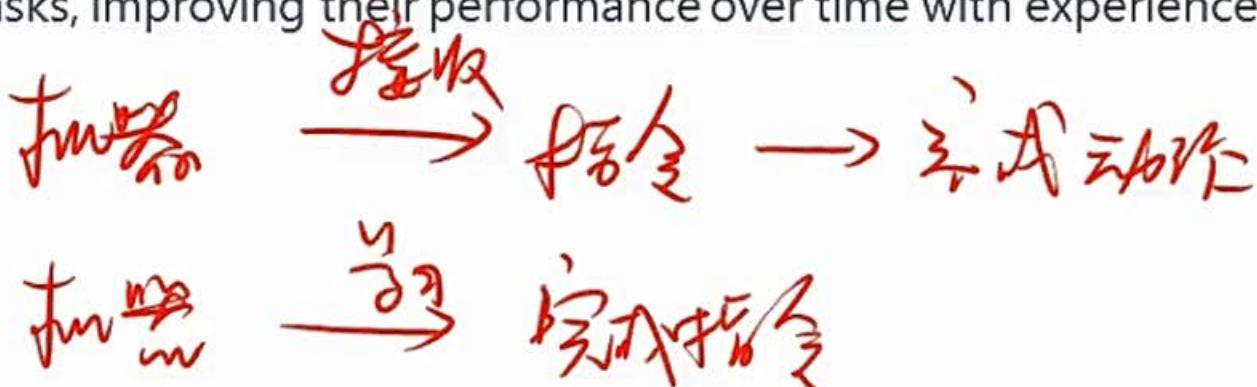
# Advanced Analytical Tools

## ➤ Artificial intelligence

- Artificial intelligence computer systems are capable of performing tasks that have traditionally required human intelligence. This is often accomplished through the use of "if then" rules.

## ➤ Machine learning ( ML )

- Machine learning ( ML ) is a technology that has grown out of the wider AI field.
- ML algorithms are computer programs that are able to **"learn"** how to complete tasks, improving their performance over time with experience.



# Advanced Analytical Tools

## ➤ How machine learning works?

- Dataset can be split into a **training dataset and validation dataset** ( evaluation dataset )
  - ✓ The training dataset allows the algorithm to identify relationships between inputs and outputs based on historical patterns in the data.
  - ✓ These relationships are then tested on the validation dataset.
- ML still **required human judgement** in understanding data and choosing the right analytic techniques.
- Errors may arise from **overfitting** and **underfitted**
  - ✓ Overfitting: make too much use of the data
  - ✓ Underfitted: make too little use of the data
- ✓ In addition, ML techniques can appear to be opaque or **"black box"** approaches, which arrive at outcomes that **may not be entirely understood or explainable.**

# Advanced Analytical Tools

## ➤ Types of machine learning



标注

- Supervised learning

- ✓ Computers learn to model relationships based on labeled training data.
- ✓ Trying to group companies into peer groups based on their industries.

无监督

- Unsupervised learning

- ✓ Computers are not given labeled data but instead are given only data from which the algorithm seeks to describe the data and their structure.
- ✓ Trying to group companies into peer groups based on their characteristics rather than using standard sector or other acknowledged criteria.

- More examples

反洗钱 / 金融  
垃圾邮件 / 语音识别

- ✓ Identify whether it is money laundering is unsupervised learning
- ✓ spam mail classification is unsupervised learning

预测

➤ **In the use of machine learning ( ML ):**

- A. Some techniques are termed "black box" due to data biases. X
- B. Human judgment is not needed because algorithms continuously learn from data. X
- C. Training data can be learned too precisely, resulting in inaccurate predictions when used with different datasets. overfit.

Answer:

**C is correct.** Overfitting occurs when the ML model learns the input and target dataset too precisely. In this case, the model has been "over trained" on the data and is treating noise in the data as true parameters. An ML model that has been overfitted is not able to accurately predict outcomes using a different dataset and may be too complex.

# Selected Applications of Fintech

## ➤ Text analytics

- Computer programs that analyze and derive meaning typically from large, unstructured text- or voice-based datasets, which include
  - ✓ company filings, written reports, quarterly earnings calls, social media, email, internet postings, and surveys

## ➤ An important application of text analytics is natural language processing ( NLP ). ✓

- A computer programs to analyze and interpret human language.
- Applications include **translation, speech recognition, text mining, sentiment analysis, and topic analysis.**
- Models using NLP analysis may incorporate **non-traditional information** to evaluate what people are saying such as their preferences, opinions, likes, or dislikes – in an attempt to **identify trends** and short-term indicators about a company, a stock, or an economic event that might have a bearing on future performance.

# Selected Applications of Fintech

## ➤ Robo-advisers services

- Robo-advisory services provide investment solutions through online platforms, reducing the need for direct interaction with financial advisers.
- Regulations governing robo-advisers services, such as
  - ✓ SEC(Securities and Exchange Commission) in the United States(registered)
  - ✓ FCA(Financial Conduct Authority) in the United kingdom
  - ✓ ASIC(Australian Securities and Investments Commission) In Australia(license)

# Selected Applications of Fintech

## ➤ Robo-advisers services( cont'd )

- Current robo-advisory services include automated asset allocation, trade execution, portfolio optimization, tax-loss harvesting, and rebalancing for investor portfolios.
- Two types of wealth management services dominate the robo-advice sector
  - ✓ Fully automated digital wealth managers
    - ◆ Including direct deposits, periodic rebalancing, and dividend reinvestment options
  - ✓ Adviser-assisted digital wealth managers
    - ◆ Involving a more holistic analysis of a client's assets and liabilities

# Selected Applications of Fintech

## ➤ Robo-advisers services( cont'd )

- Current robo-advisory services include automated asset allocation, trade execution, portfolio optimization, tax-loss harvesting, and rebalancing for investor portfolios.
- Two types of wealth management services dominate the robo-advice sector
  - ✓ Fully automated digital wealth managers
    - ◆ Including direct deposits, periodic rebalancing, and dividend reinvestment options
  - ✓ Adviser-assisted digital wealth managers
    - ◆ Involving a more holistic analysis of a client's assets and liabilities

partially  
L Assist Asset manager

# Selected Applications of Fintech

## ➤ Robo-advisers services

- The **characteristics** of robo-advisers' analyses and recommendations
  - ✓ Most following a **passive**, or fairly conservative **investment approach**
  - ✓ Typically having **low fees and low account minimums**
  - ✓ Robo-advisers can reach **underserved populations**
- **Criticism** of robo-advisers
  - ✓ It may not always be completely transparent why a robo-adviser chooses to make a recommendation or take a trading action.
  - ✓ The growth of the complexity and size of an investor's portfolio makes a team of human advisers likely to endure.

black-box

# Selected Applications of Fintech

## ➤ Risk analysis

- As mandated by regulators worldwide, the global investment industry has undertaken major steps in stress testing and risk assessment that involve the analysis of vast amounts of quantitative and qualitative risk data.
  - ✓ Stress tests may also take qualitative information into consideration, such as capital planning procedures, expected business plan changes, business model sustainability, and operational risk.
  - ✓ There is increasing interest in monitoring risk in real time.
  - ✓ ML techniques may be used to help assess data quality.
  - ✓ The backtesting simulations in portfolio risk management are often computationally intense and may be facilitated through the use of advanced AI-based techniques.

# Selected Applications of Fintech

## ➤ Algorithmic trading

II

组合交易

- Algorithmic trading is the computerized **buying and selling of financial instruments**, in accordance with pre-specified rules and guidelines.
- Algorithmic trading is often used to execute large institutional orders, **slicing orders into smaller pieces** and executing across different exchanges and trading venues. ✓
- Many benefits provided by algorithmic trading
  - ✓ Including **speed** of execution, anonymity, and lower transaction costs
  - ✓ Over the course of a day, algorithms may **continuously update and revise** their execution strategy on the basis of changing prices, volumes, and market volatility.
- **High-frequency trading** ( HFT ) is a form of algorithmic trading that will define when to trade, what to trade and how to trade, which will be discussed in #Portfolio Management#.

## Example-1

➤ **Text Analytics is appropriate for application to:**

- A. Economic trend analysis. ✓
- B. Large, structured datasets. ✗
- C. Public but not private information. ✗

Answer:

A is correct. Through the Text Analytics application of natural language processing ( NLP ), models using NLP analysis may incorporate non-traditional information to evaluate what people are saying – via their preferences, opinions, likes, or dislikes – in the attempt to identify trends and short-term indicators about a company, a stock, or an economic event that might have a bearing on future performance.

## Example-2

➤ In providing investment services, robo-advisers are most likely to:

- A. Rely on their cost effectiveness to pursue active strategies. *Passive*
- B. Offer fairly conservative advice as easily accessible guidance.
- C. Be free from regulation when acting as fully-automated wealth managers. *X*

Answer:

B is correct. Research suggests that robo-advisers tend to offer fairly conservative advice, providing a cost-effective and easily accessible form of financial guidance to underserved populations, such as the mass affluent and mass market segments.

# Distributed Ledger Technology

## ➤ Introduction

- Distributed ledger technology — technology based on a distributed ledger — represents a fintech development that offers potential improvements in the area of financial record keeping.
- A distributed ledger is a type of database that may be shared among entities in a network.  
✓ 账本公开 / 共享
- In a distributed ledger, entries are recorded, stored, and distributed across a network of participants so that each participant has a matching copy of the digital database.

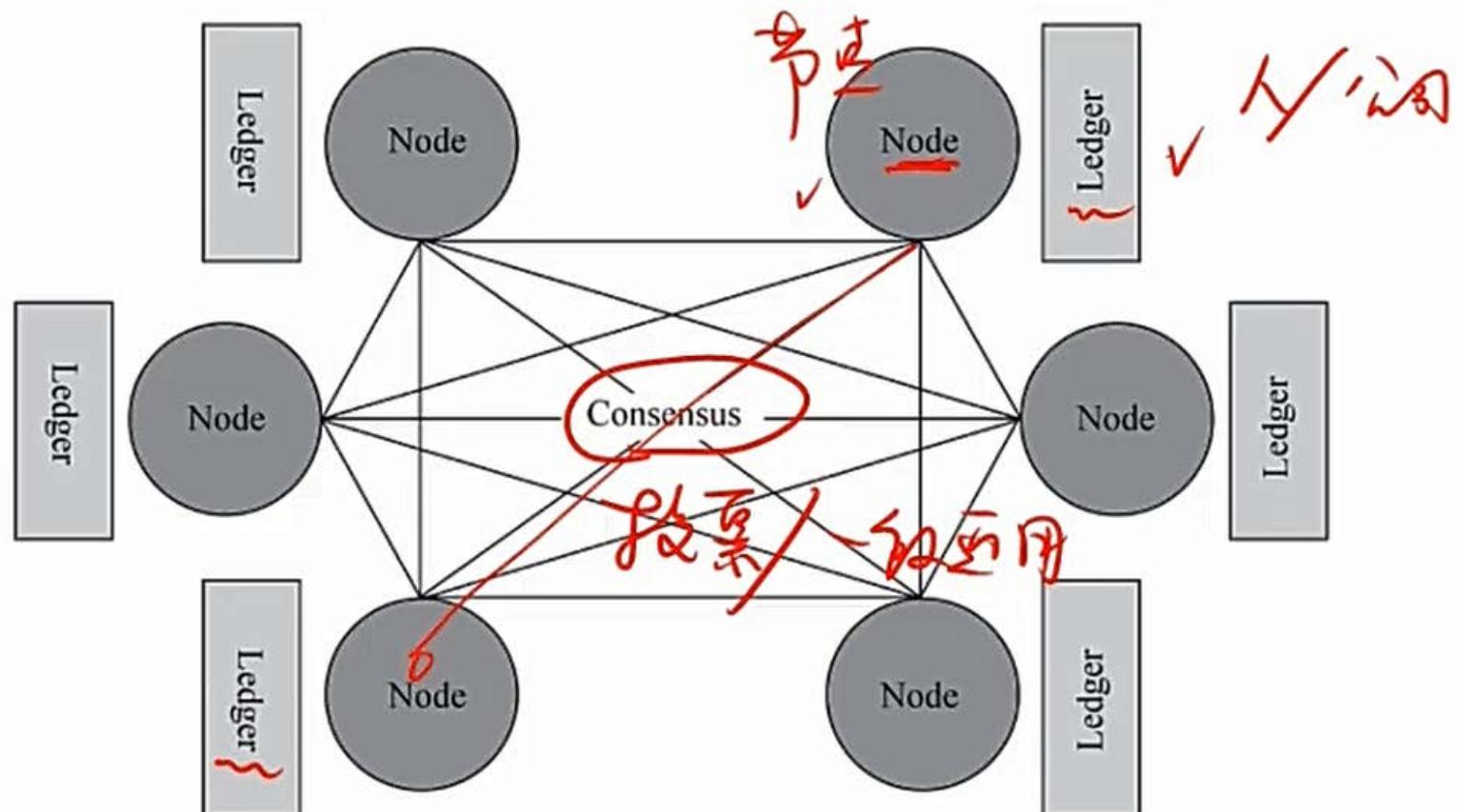


# Distributed Ledger Technology

共识  
Consensus

## > Distributed Ledger Network Setup

- Basic elements of a DLT network include a **digital ledger**, a **consensus mechanism** used to confirm new entries, and a participant **network**.



# Distributed Ledger Technology



## ➤ Consensus mechanism

- ✓ The consensus mechanism is the process by which the computer entities ( or nodes ) in a network **agree on a common state of the ledger.**
- ✓ Consensus generally involves two steps: transaction validation and agreement on ledger update by network parties.

## ➤ Features of DLT include the use of **cryptography**

- An algorithmic process to encrypt data, making the data unusable if received by unauthorized parties — which enables a high level of network security and database integrity.

## ➤ DLT has the potential to accommodate "**Smart contracts**"

- Computer programs that self-execute on the basis of pre-specified terms
- ✓ automatic execution of contingent claims for derivatives
- ✓ instantaneous transfer of collateral in the event of default.

Block chain

DLT

# Distributed Ledger Technology

## ➤ Blockchain

- A type of digital ledger in which information, such as changes in ownership, is recorded sequentially within blocks that are then linked or "chained" together and secured using cryptographic methods.
- steps involved in adding a transaction to a blockchain distributed ledger.
  - ✓ Transaction takes place between buyer and seller.
  - ✓ Transaction is broadcast to the network of computers (nodes).
  - ✓ Nodes validate the transaction details and parties to the transaction.
  - ✓ Once verified, the transaction is combined with other transactions to form a new block (of predetermined size) of data for the ledger.
  - ✓ This block of data is then added or linked (using a cryptographic process) to the previous block(s) containing data. chain
  - ✓ Transaction is considered complete and ledger has been updated.

# Distributed Ledger Technology

- DLT can take the form of permissionless and permissioned networks.

- Permissionless networks

✓ Permissionless networks are open to any user who wishes to make a transaction, and all users within the network can see all transactions that exist on the blockchain.

✓ The main benefit of a permissionless network is that it does not depend on a centralized authority to confirm or deny the validity of transactions, because this takes place through the consensus mechanism.

✓ A well-known example of an application of blockchain technology using an open, permissionless network is **bitcoin**.

# Distributed Ledger Technology

- DLT can take the form of permissionless and permissioned networks. ( cont'd )

- Permissioned networks

✓ In permissioned network, network members may be restricted from participating in certain network activities. ( from adding transactions to viewing transactions with limited or selected details of the transaction)



\* 资料来源：铅笔 - 信息技术行业分布式账本技术：超区块链接

# Applications of Distributed Ledger Technology

## ➤ Cryptocurrencies, also known as a digital currency

- Most issued cryptocurrencies use a decentralized distributed ledger to record and verify all digital currency transactions.
- Cryptocurrencies have not traditionally been government backed or regulated.
- Many cryptocurrencies have a self-imposed limit on the total amount of currency they may issue.
- It provides an attractive means of raising capital.
  - ✓ an ICO is an unregulated process whereby companies sell their crypto tokens to investors in exchange for fiat money or for another agreed upon cryptocurrency.

## ➤ Tokenization

*I don't fit in*

- Through tokenization, the process of representing ownership rights to physical assets on a blockchain or distributed ledger, DLT has the potential to streamline this process by creating a single, digital record of ownership with which to verify ownership title and authenticity, including all historical activity.

↑ taken way  
→

186. ...

③ 打

人

# Applications of Distributed Ledger Technology

## ➤ Tokenization

- Through tokenization, the process of representing ownership rights to physical assets on a blockchain or distributed ledger, DLT has the potential to streamline this process by creating a single, digital record of ownership with which to verify ownership title and authenticity, including all historical activity.

I don't like it

## ➤ Post-trade clearing and settlement ✓

- DLT has the ability to streamline existing post-trade processes by providing near-real-time trade verification, reconciliation, and settlement, thereby reducing the complexity, time, and costs associated with processing transactions.

## ➤ Compliance

一致投票

- DLT-based compliance may better support shared information, communications, and transparency within and between firms, exchanges, custodians, and regulators.

## Example-1

- A benefit of distributed ledger technology ( DLT ) favoring its use by the investment industry is its:
- A. Scalability of underlying systems.
  - B. Ease of integration with existing systems.
  - C. Streamlining of current post-trade processes.

✓ Answer:

C is correct. DLT has the potential to streamline the existing, often complex and labor intensive post-trade processes in securities markets by providing close to real-time trade verification, reconciliation, and settlement, thereby reducing related complexity, time, and costs.

## Example-2

➤ What is a distributed ledger technology ( DLT ) application suited for physical assets?

- A. Tokenization
- B. Cryptocurrencies ✓
- C. Permissioned networks

Answer:

1. A

A is correct. Through tokenization—the process of representing ownership rights to physical assets on a blockchain or distributed ledger – DLT has the potential to streamline this rights process by creating a single, digital record of ownership with which to verify ownership title and authenticity, including all historical activity.



# Sample Covariance and Correlation

### ➤ Covariance:

- Covariance measures how one random variable moves with another random variable. — a measure of **linear** association.
  - Sample covariance:

$$\underline{Cov}(X, Y) = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})/(n - 1)$$

- $-\infty < \text{Cov}(X, Y) < +\infty$

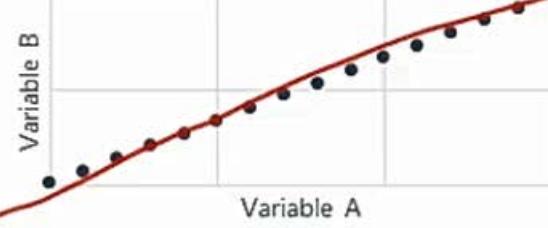
## ➤ Correlation:

- The correlation coefficient measures the **direction** and **extent** of linear association between two variables
  - Sample correlation:

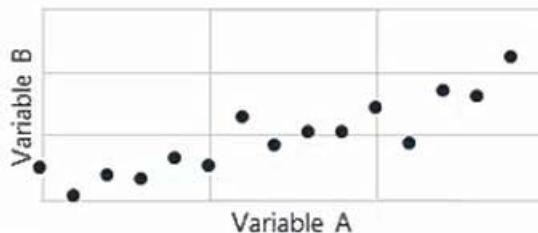
$$r = \frac{Cov(X, Y)}{\sqrt{S_x S_y}}$$

# ◆ Interpretations of Correlation Coefficients

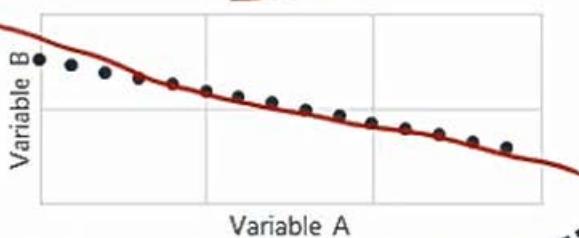
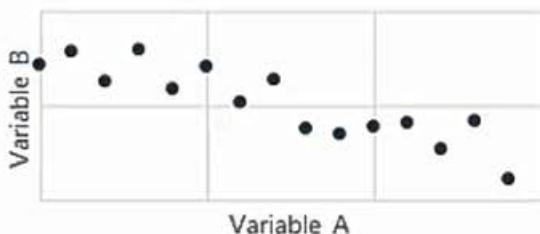
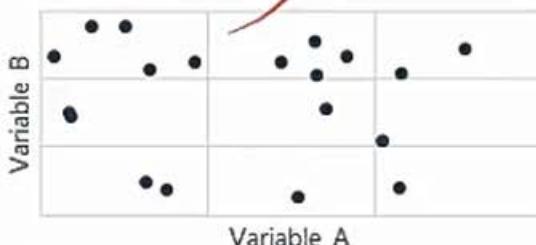
- A **scatter plots** is a graph that shows the relationship between the observations for two data series in two dimensions.



$r=0$



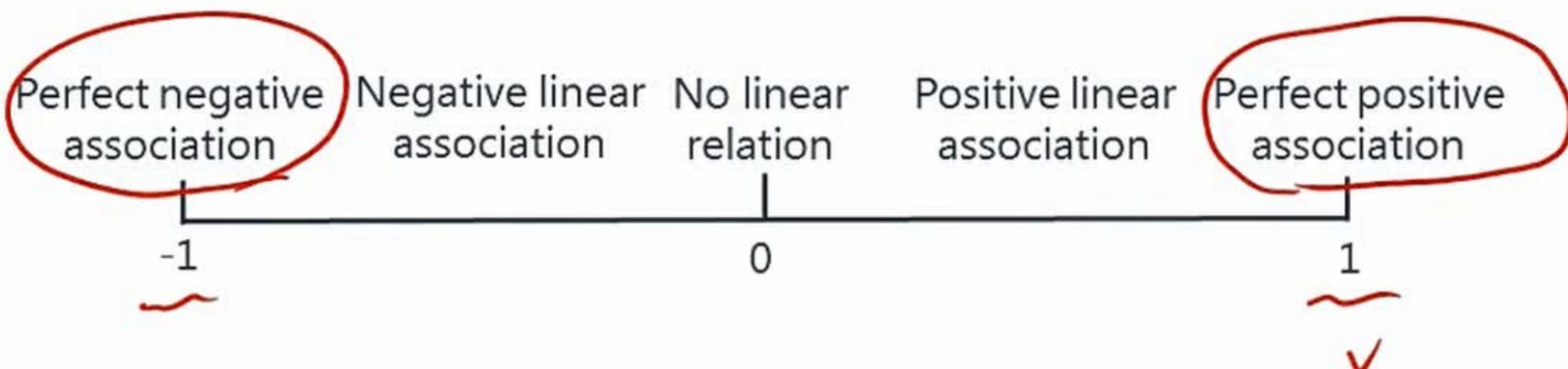
$0 < r < 1$





## Interpretations of Correlation Coefficients

- The correlation coefficient is a measure of **linear association**.
  - It is a simple number with no unit of measurement attached, so the correlation coefficient is much easier to explain than the covariance.

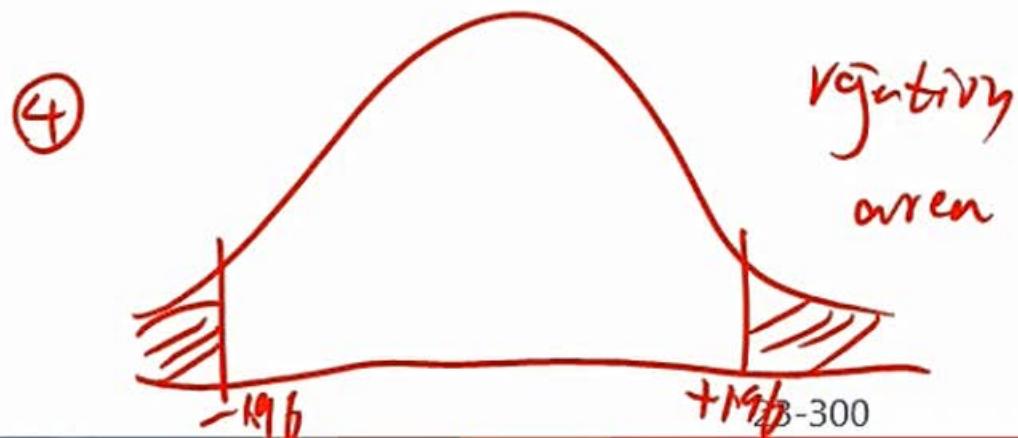


①  $\mu = 170 \rightarrow H_0$   $\mu$

$\mu \neq 170 \rightarrow H_A$

②  $\frac{\bar{X} - 170}{SE} = t$

③  $\alpha = 5\%$   $N(10000)$   $\rightarrow 1.96$   
 $\alpha = 1\%$   $\rightarrow 2.58$



④  $t = 5$   
 $t > 1.96 \quad \mu \neq 170$   
 $\rightarrow \text{reject } H_0$

# Significance Test of The Correlation

- Test whether the correlation between the population of two random variables is significantly different from zero.

- Step 1:  $H_0: \rho=0; H_a: \rho \neq 0$  (Two-tailed test)
- Step 2: Calculate the test statistic: ?

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

$$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}}, df = n - 2$$

$$t = \frac{r - 0}{SE}$$

- Step 3: Decision rule: reject  $H_0$  if  $|t| > +t_{\text{critical}}$
- Step 4: Draw a conclusion that the correlation between the population of two variables is **significantly different from zero** if  $H_0$  is rejected.

## Example

- The covariance between X and Y is 16. The standard deviation of X is 4 and the standard deviation of Y is 8. The sample size is 20. Test the significance of the correlation coefficient at the 5% significance level.

~~$$r = \frac{16}{4 \times 8} = 0.5$$~~

### Solution:

- The sample correlation coefficient  $r = 16/(4 \times 8) = 0.5$ . The t-statistic can be computed as:

(2)

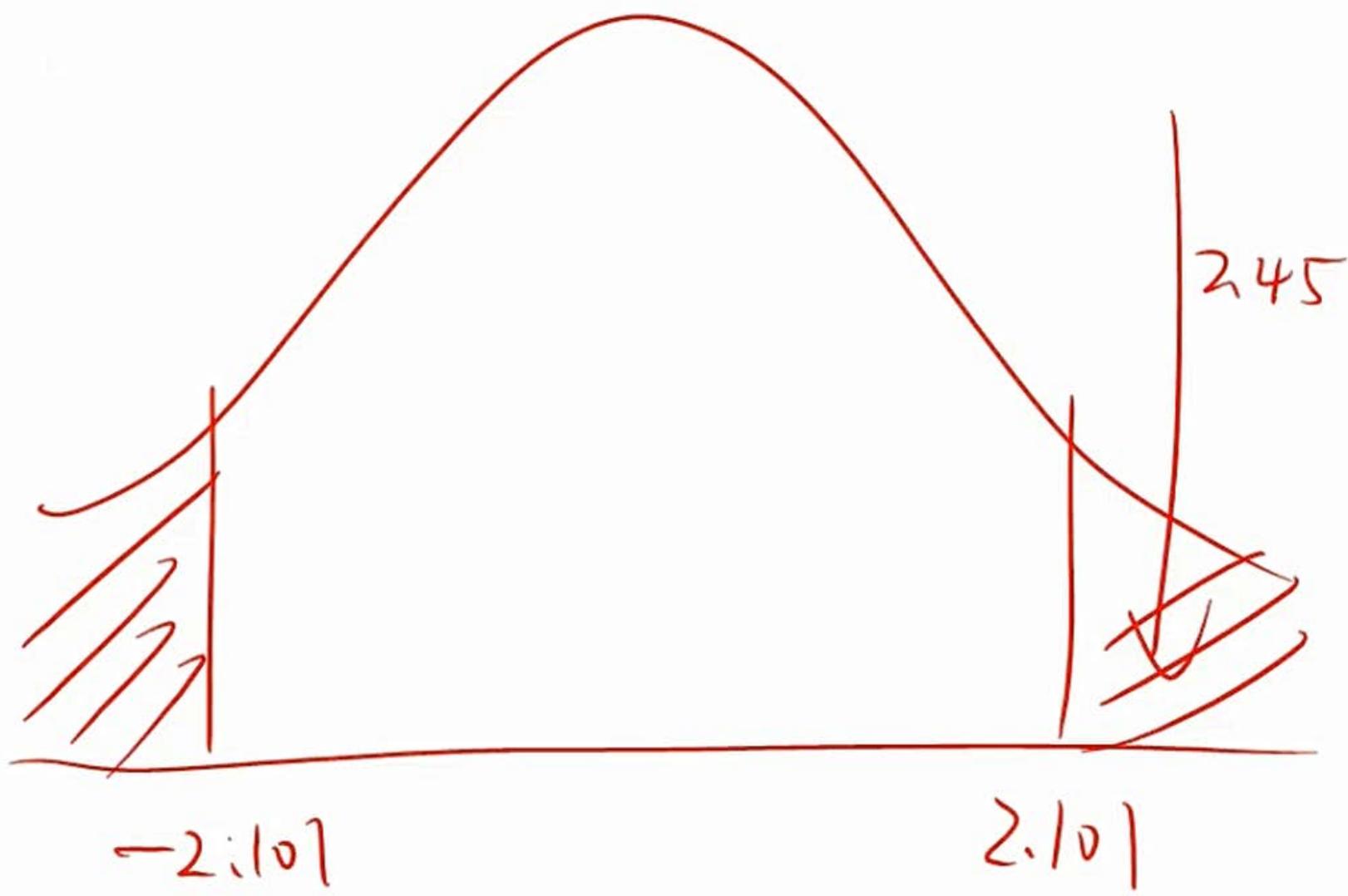
$$t = 0.5 \times \frac{\sqrt{20-2}}{\sqrt{1-0.25}} = 2.45$$

$$\left. \begin{array}{l} H_0: \rho \leq 0 \\ H_a: \rho > 0 \end{array} \right\}$$

$$\frac{0.5 - 0}{\sqrt{1 - 0.5^2}}$$

The critical t-value for  $\alpha=5\%$ , two-tailed test with  $df=18$  is 2.101.

Since the test statistic of 2.45 is larger than the critical value of 2.101, we have sufficient evidence to **reject the null hypothesis**. So we can say that the correlation coefficient between X and Y is significantly different from zero.



# ◆ Significance Test of The Correlation

- Test whether the correlation between the population of two random variables is significantly different from zero.

- Step 1:  $H_0: \rho=0$ ;  $H_a: \rho \neq 0$  (Two-tailed test)
- Step 2: Calculate the test statistic:

$$t = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}}, \text{ df} = n-2$$

$$SE = \sqrt{\frac{r^2}{n-2}}$$

$\text{df} = n-2$

- Step 3: Decision rule: reject  $H_0$  if  $|t| > +t_{\text{critical}}$
- Step 4: Draw a conclusion that the correlation between the population of two variables is **significantly different from zero** if  $H_0$  is rejected.



# Limitations to Correlation Analysis

## ➤ Three limits of correlation analysis

- Outliers
- Spurious correlation
- Nonlinear relationships

$$r \neq 0 \quad \checkmark$$

# ◆ Limitations to Correlation Analysis

## ➤ Outliers



- Outliers are small numbers of observations at either extreme (small or large) of a sample.
- As a general rule, we must determine whether a computed sample correlation changes greatly by removing a few outliers.

✓ Outliers contain information about the two variables' relationship

◆ Should be included in the correlation analysis

✓ Outliers contain no information

◆ Should be excluded.

# Limitations to Correlation Analysis

## ➤ Spurious correlation

偽

- Correlations can be spurious in the sense of misleadingly pointing towards associations between variables.

- ① ✓ Correlation between two variables that reflects chance relationships in a particular data set
- ② ✓ Correlation induced by a calculation that mixes each of two variables with a third (Two variables that are uncorrelated may be correlated if divided by a third variable.)
- ③ ✓ Correlation between two variables arising not from a direct relation between them but from their relation to a third variable  
(Height may be positively correlated with the extent of a person's vocabulary)

之入之出之

# ◆ Limitations to Correlation Analysis

## ➤ Nonlinear relationships

- Two variables can have a strong nonlinear relation and still have a very low correlation.
  - ✓ For example, two variables could have a nonlinear relationship such as  $Y = (1-X)^3$  and the correlation coefficient would be close to zero, which is a limitation of correlation analysis.

$r=0 \rightarrow \text{No linear relationship}$

# The Basics of Simple Linear Regression

- Linear regression with one independent variable, sometimes called **simple linear regression**, models the relationship between two variables as a straight line.
- Linear regression with one regressor allows you to
  - use linear regression to summarize the relationship, if this relationship between two variables is linear,
  - use one variable to make predictions about another.

# Simple Linear Regression

- The simple linear regression model

$$Y_i = b_0 + b_1 X_i + \varepsilon_i, i = 1, \dots, n$$

- **Linear regression** assumes a linear relation between the dependent and the independent variables.
  - **The dependent variable,  $Y$**  is the variable whose variation about its mean is to be explained by the regression.
  - **The independent variable,  $X$**  is the variable used to explain the dependent variable in a regression.
  - **Regression coefficients,  $b_0$**  is intercept term of the regression,  **$b_1$**  is slope coefficient of the regression, regression coefficient.
  - **The error term,  $\varepsilon_i$**  is the portion of the dependent variable that is not explained by the independent variable(s) in the regression.

# ◆ Calculation of Regression Coefficients

- How does linear regression estimate  $b_0$  and  $b_1$ ?
  - Computes a line that best fits the observations
  - The **estimated intercept coefficient** ( $\hat{b}_0$ ) is interpreted as the value of Y when X is equal to zero.
  - The **estimated slope coefficient** ( $\hat{b}_1$ ) defines the sensitivity of Y to a change in X .The estimated slope coefficient ( $\hat{b}_1$ ) equals covariance divided by variance of X.
- **Example of interpretation of estimated coefficients**
  - An estimated slope coefficient of 2 would indicate that the dependent variable will change two units for every 1 unit change in the independent variable.
  - The intercept term of 2% can be interpreted to mean that the independent variable is zero, the dependent variable is 2%.

$$\left( \sum_i \xi_i \right) \frac{\partial}{\partial \xi_i}$$

$$\sum_i \sum_i^2 f_{\alpha, i}$$

$$\sum_i \xi_i^4 \frac{\partial}{\partial \xi_i}$$

# Calculation of Regression Coefficients

## ➤ Ordinary least squares (OLS)

~~最小二乘法~~



- OLS estimation is a process that estimates the population parameters  $B_i$  with corresponding values for  $b_i$  which
  - ✓ minimize the sum of squared vertical distances between the observations and the regression line (also called residuals or error terms).
- the OLS sample coefficients are those that:

$$b_1 = \frac{Cov(X, Y)}{Var(X)} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- The estimated intercept coefficient ( $\hat{b}_0$ ) : because the point  $(\bar{X}, \bar{Y})$  lies on the regression line. we can solve  $\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$ .

$$b_1 = \frac{\text{Cov}(x, y)}{S^2(x)}$$

$$\beta = \frac{\text{Cov}(z, m)}{S^2(m)}$$

## Example: Calculate A Regression Coefficient

- Bouvier Co. is a Canadian company that sells forestry products to several Pacific Rim customers. Bouvier's sales are very sensitive to exchange rates. The following table shows recent annual sales (in millions of Canadian dollars) and the average exchange rate for the year (expressed as the units of foreign currency needed to buy one Canadian dollar).

Year i	$X_i = \text{Exchange Rate}$	$Y_i = \text{Sales}$
1	0.40	20
2	0.36	25
3	0.42	16
4	0.31	30
5	0.33	35
6	0.34	30

- Calculate the intercept and coefficient for an estimated linear regression with the exchange rate as the independent variable and sales as the dependent variable.

## Example: Calculate A Regression Coefficient

- The sample mean of the exchange rate is:

$$\bar{X} = \sum_{i=1}^n X_i / n = 2.16 / 6 = 0.36$$

- The sample mean of sales is:

$$\bar{Y} = \sum_{i=1}^n Y_i / n = 156 / 6 = 26$$

- We want to estimate a regression equation of the form  $Y_i = b_0 + b_1 X_i + \varepsilon_i$ . The estimates of the slope coefficient and the intercept are

✓  $\hat{b}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{-1.39}{0.009} = -154.44$

✓  $\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} = 26 - (-154.44)(0.36) = 26 + 55.6 = 81.6$

- So the regression equation is  $Y_i = 81.6 - 154.444X_i$

*{ $\Sigma x_i$  ( $\Sigma x_i^2$ ),  $y$ )*

*$S^2(X)$*

*$S^2(Y^2)$*

*$Y^-$*

# The Assumptions of the Linear Regression

## The six classic normal linear regression model assumptions

- The relationship between the dependent variable,  $Y$ , and the independent variable,  $X$  is linear in the parameters  $b_0$  and  $b_1$ . However, the requirement does not exclude  $X$  from being raised to a power other than 1 ( $X^2$ ).

$\sum_{X=2} \checkmark Y_i = b_0 e^{b_1 X_i} + \varepsilon_i$  is nonlinear in  $b_1$ , so we could not apply the linear regression model to it.

$$Y = b_0 \cdot e^{b_1 X_2}$$

$\sum_{X=3}$

$\checkmark$  Even if the dependent variable is nonlinear,  $\underline{Y_i = b_0 + b_1 x_i^2 + \varepsilon_i}$ , however, linear regression can still be used to estimate.

- 2 ● The independent variable,  $X$ , is not random, with the **exception** that  $X$  is random but also uncorrelated with the error term.

- 3 ● The expected value of the error term is zero (i.e.,  $E(\varepsilon_i) = 0$ )

$$y = b_0 + b_1 x$$

- 4 ● The variance of the error term is constant (i.e., the error terms are homoskedastic)

- 5 ● The error term is uncorrelated across observations (i.e.,  $E(\varepsilon_i \varepsilon_j) = 0$  for all  $i \neq j$ )
- The error term is normally distributed.

$$y = b_0 + e^{b_1} \cdot x^2$$

①  $y \neq x$  之間的關係

2種類

$$x=2 \quad y = b_0 + e^{b_1} \cdot 4$$

②  $y \neq x$  之間的關係 -

2/3

$$b_0 = 1 \quad b_1 = 1 \quad y = 1 + e^{x^2}$$

# ◆ Standard Error of Estimate

- **Standard Error of Estimate (SEE)** gives some indication of how well a given linear regression model captures the relationship between the dependent and independent variable.
  - SEE is low if the regression is very strong and high if the relationship is weak.
- The formula of SEE

$$SEE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n \varepsilon_i^2}{n-2}}$$

- The SEE formula looks like the formula for computing a standard deviation, except that  $n-2$  appears in the denominator instead of  $n-1$ .
- In fact, the SEE is the standard deviation of the error term because the degree of freedom of the error is  $n-2$  to indicate the fact that the sample includes  $n$  observations with two estimated parameters,  $\hat{b}_0$  and  $\hat{b}_1$ .

$$\frac{\sum (E_i - \bar{E}_i)^2}{df}$$

df

$$\frac{\sum (x_i - \bar{x})^2}{df}$$

$$= \frac{\sum x_i^2}{df}$$

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k + \epsilon$$

$$n - k - 1$$

# Standard Error of Estimate

- Standard Error of Estimate (SEE) gives some indication of how well a given linear regression model captures the relationship between the dependent and independent variable.

SEE is low if the regression is very strong and high if the relationship is weak.

- The formula of SEE

SEE error

$$\text{SEE} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n \varepsilon_i^2}{n-2}} \quad \checkmark \quad df = n-2$$

- The SEE formula looks like the formula for computing a standard deviation, except that n-2 appears in the denominator instead of n-1. (n-k-1)
- In fact, the SEE is the standard deviation of the error term because the degree of freedom of the error is n-2 to indicate the fact that the sample includes n observations with two estimated parameters,  $\hat{b}_0$  and  $\hat{b}_1$ .

## Coefficient of Determination ( $R^2$ )

- Coefficient of determination ( $R^2$ ) measures the **fraction** of the total variation in the dependent variable that is explained by the independent variable.
  - $0 \leq R^2 \leq 1$
  - more intuitive than SEE, a  $R^2$  of 0.8250 means the independent variable explains approximately 82.5 percent of the variation in the dependent variable.



# Coefficient of Determination ( $R^2$ )

## ➤ How to calculate $R^2$ ?

- For simple linear regression only,  $R^2$  is equal to the squared correlation coefficient.

- For simple and multiple linear regression,

$$R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

$$= \frac{\text{total variation} - \text{unexplained variation}}{\text{total variation}}$$

$$= 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

$$= 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

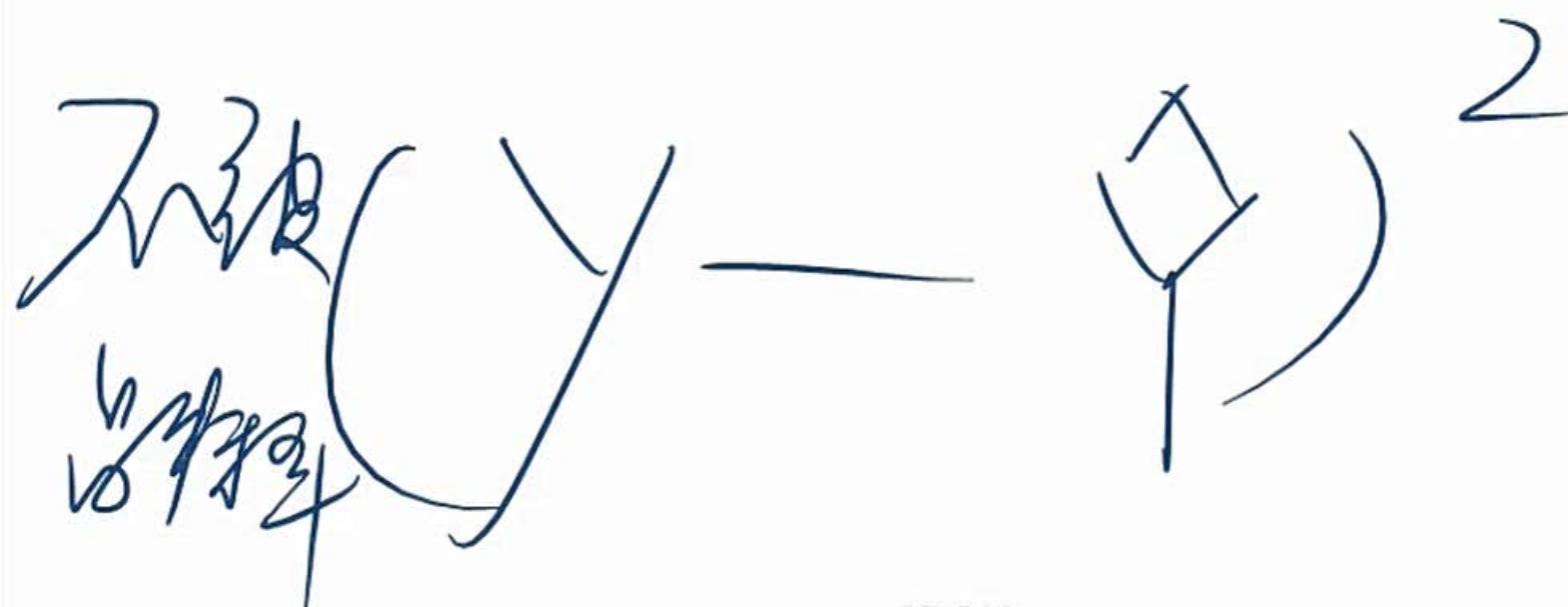
- For report purpose, regression programs also report **multiple R**, which is the correlation between the actual values and the forecast values of  $Y$ . The coefficient of determination is the square of multiple  $R$ .

$$R^2 = r^2$$

已知  $(y - \bar{y})^2$



已被  
解釋



$$1 - \frac{\sum (Y_i - \hat{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

不被

是

$$= R^2$$

# Coefficient of Determination ( $R^2$ )

## ➤ How to calculate $R^2$ ?

- For simple linear regression only,  $R^2$  is equal to the squared correlation coefficient.

- For simple and multiple linear regression,

$$R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

$$= \frac{\text{total variation} - \text{unexplained variation}}{\text{total variation}}$$

$$= 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

$$= 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \rightarrow 2$$

$$R^2 = r^2$$

- 67

$$R^2 = 0.45$$

- For report purpose, regression programs also report **multiple R**, which is the correlation between the actual values and the forecast values of  $Y$ . The coefficient of determination is the square of multiple  $R$ .

wave triple R =

$$\sqrt{R^2}$$

$$\text{R}^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

For report purpose, regression programs also report multiple R, which is the correlation between the actual values and the forecast values of Y. The coefficient of determination is the square of multiple R.

# Coefficient of Determination ( $R^2$ )

## ➤ The Different between the $R^2$ , multiple R and Correlation Coefficient

- The correlation coefficient indicates the sign of the relationship between two variables, whereas the coefficient of determination does not. 只看去向
- Multiple R is the correlation between the actual values and the forecast values of Y.  
(只看去向)
  - ✓ It is the square root of  $R^2$  and always positive.
  - ✓ It can be the same the correlation between dependent and independent variable only in case of a simple linear regression and a positive slope coefficient.
  - ✓ It can also apply to multiple regression.

$R^2$

- The coefficient of determination can apply to an equation with several independent variables, and it implies a explanatory power, while the correlation coefficient only applies to two variables and does not imply explanatory between the variables.

$R^2$

as |

as |

wwwUTRPT R

0.9

0.9

r

-0.9  
—  
3

0.  
3

$$t = \frac{b_1 - 0}{SE}$$

in

$SE(b_1)$

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

$$OLS \rightarrow \hat{b}_1 = 2$$

$$SE(\hat{b}_1) = 0,5$$

$$t = \frac{2 - 0}{0,5} = 4$$

# Regression Coefficient Confidence Interval

- Regression coefficient confidence interval

$$\hat{b}_1 \pm t_c S_{\hat{b}_1}$$

- If the confidence interval with a given degree of confidence does not include the hypothesized value, the null is rejected, and the coefficient is said to be statistically different from hypothesized value.
- $S_{\hat{b}_1}$  is the standard error of the estimated coefficient.
  - Stronger regression results (usually lower SEE or higher R<sup>2</sup>) lead to **smaller** standard error of an estimated coefficient  $S_{\hat{b}_1}$  and tighter confidence intervals.

# Hypothesis Testing about Regression Coefficient

## ➤ Significance test for regression coefficient

- $H_0: b_1 = 0$
- Test Statistic:

$$S\bar{E}(b_1)$$

$$t = \frac{\hat{b}_1 - 0}{S\hat{b}_1} = \frac{\hat{b}_1}{S\hat{b}_1}, \text{ df=n-2}$$

- Decision rule: reject  $H_0$  if  $|t| > + t_{\text{critical}}$
- Rejection of the null means that the slope coefficient is significantly different from zero.

# Hypothesis Testing about Regression Coefficient

## Hypothesis testing about regression coefficient

- $H_0: b_1 = \text{hypothesized value of } b_1$

$$\underline{b_1 = 2}$$

- Test Statistic:

$$t = \frac{\hat{b}_1 - \text{hypothesized value of } b_1}{s_{\hat{b}_1}}, \text{ df} = n-2 \quad (n+1)$$

- Decision rule:** reject  $H_0$  if  $|t| > + t_{\text{critical}}$
- Rejection of the null means that the slope coefficient is significantly different from the hypothesized value of  $b_1$ .

$$t = \frac{\hat{b}_1 - 2}{S E(\hat{b}_1)}$$

# Regression Coefficient Confidence Interval

- Regression coefficient confidence interval

$$\hat{b}_1 \pm t_c S_{\hat{b}_1}$$

- If the confidence interval with a given degree of confidence does not include the hypothesized value, the null is rejected, and the coefficient is said to be statistically different from hypothesized value.

- $S_{\hat{b}_1}$  is the standard error of the estimated coefficient.

- Stronger regression results (usually lower SEE or higher R<sup>2</sup>) lead to **smaller** standard error of an estimated coefficient  $S_{\hat{b}_1}$  and tighter confidence intervals.

$n > 30$ , 背

$n < 30$

$df = n - k - 1$

切勿  
切勿

# ◆ Analysis of Variance (ANOVA) Table

- Analysis of variance (ANOVA) is developed to test the **overall significance of the model.**
- **How it work?**
  - it is a statistical procedure for **dividing the total variability of a variable into components** that can be attributed to two sources, one from the regression model and the other from the model error.
  - In regression analysis, we use ANOVA to determine the usefulness of the independent variable or variables in explaining variation in the dependent variable.
- An important statistical test conducted in ANOVA is the **F-test**.

$R^2$

In practise



## ◆ Analysis of Variance (ANOVA) Table

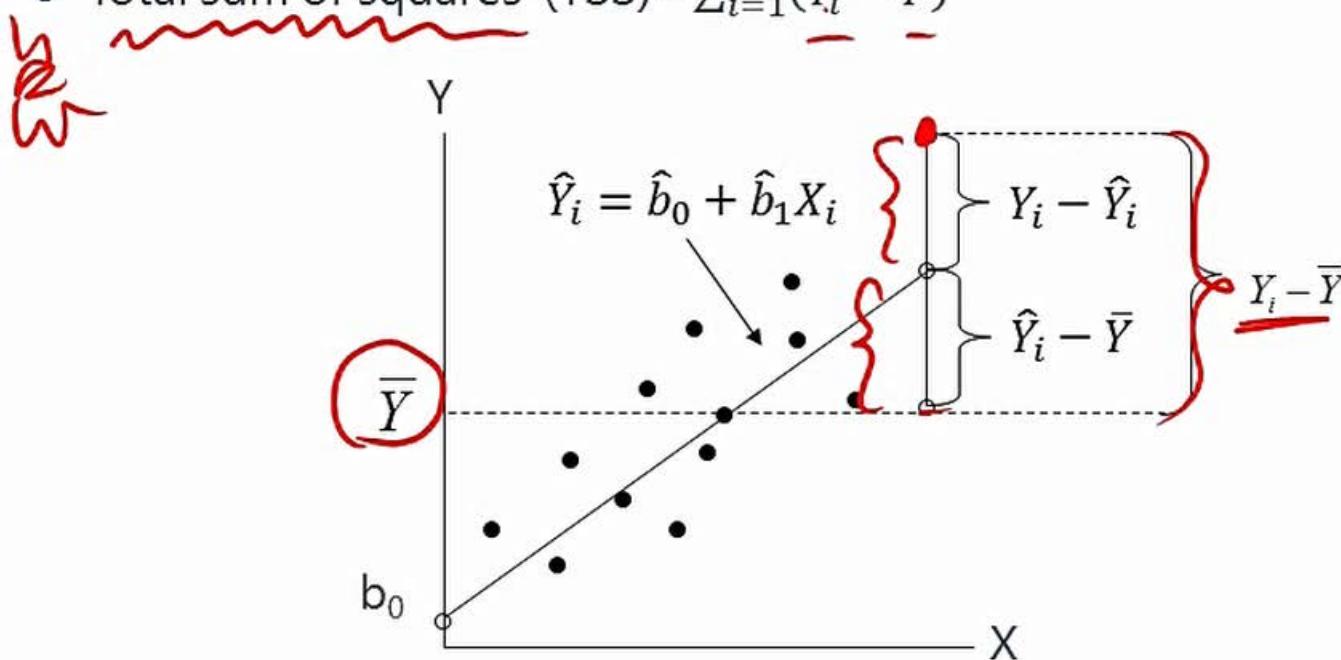
## ➤ Elements of ANOVA Table

- The sum of squared errors or residuals (SSE) =  $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
  - The regression sum of squares (RSS) =  $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
  - Total sum of squares (TSS) =  $\sum_{i=1}^n (Y_i - \bar{Y})^2$

unexplained

卷之四

行  
來



## ➤ ANOVA table

	df	SS	MSS	F
✓ Regression	k=1	RSS ✓	MSR=RSS/k	MSR/MSE
Residual/ Error	<u>n-k-1</u>	<u>SSE</u> ✓	MSE=SSE/(n-2)	
Total	<u>n-1</u>	SST ✓	-	

➤ **The F-statistic tests** whether all the slope coefficients in a linear regression are equal to 0. In a regression with one independent variable, this is a test of the null hypothesis  $H_0: b_1 = 0$  against the alternative hypothesis  $H_a: b_1 \neq 0$ .

- $F = \frac{RSS/1}{SSE/(n-2)}$  with degree of freedoms equal to 1 and n-2.
- In simple regression, the *F*-test duplicates the *t*-test for the significance of the slope coefficient. However, this relation is not true for regressions with two or more slope coefficients.

# ◆ ANOVA Table

➤ ANOVA table

	df	SS	MSS	F
Regression	k=1	RSS	MSR=RSS/k	MSR/MSE
Error	n-2	SSE	MSE=SSE/(n-2)	
Total	n-1	SST		

- The **F-statistic tests** whether all the slope coefficients in a linear regression are equal to 0. In a regression with one independent variable, this is a test of the null hypothesis  $H_0: b_1 = 0$  against the alternative hypothesis  $H_a: b_1 \neq 0$ .

- $F = \frac{RSS/1}{SSE/(n-2)}$  with degree of freedoms equal to 1 and n-2.
- In simple regression, the *F*-test duplicates the *t*-test for the significance of the slope coefficient. However, this relation is not true for regressions with two or more slope coefficients.

$$\text{RSS} \approx k^2 = k$$

~~$$R^2 = 1 - \frac{SSE}{SST} = \frac{RSS}{SST}$$~~

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

1.  $H_0$  Significant  $H_A$
2.  $F = \frac{\frac{V}{RSS/1}}{\frac{V}{SSE/(n-2)}} \quad (\text{单}) = \frac{MSE}{SSE}$
3.  $\hat{y}_j$

- In simple regression, the F-test ~~duplicates~~ the t-test for the significance of the slope coefficient. However, this relation is not true for regressions with two or more slope coefficients.

# ◆ Analysis of Variance (ANOVA) Table

- Analysis of variance (ANOVA) is developed to test the **overall significance of the model.** *R<sup>2</sup>*
- **How it work?**
  - it is a statistical procedure for **dividing the total variability of a variable into components** that can be attributed to two sources, one from the regression model and the other from the model error.  
~~SST~~ ~~SS<sub>E</sub>~~ ~~R<sub>S</sub>~~
  - In regression analysis, we use ANOVA to determine the usefulness of the independent variable or variables in explaining variation in the dependent variable.
- An important statistical test conducted in ANOVA is the **F-test.**

*In practise*

$X$ 是否显著者

$$\rightarrow t\text{-test} \quad \frac{b_1 - b}{SE(b_1)}$$



而  $b_1$  与  $X$  是否显著

$$\left\{ \begin{array}{ll} \text{单元} & t\text{-test} \\ \text{多元} & \rightarrow R^2(X) \end{array} \right.$$



$$F(\text{test}) = \frac{MSR}{MSE}$$

$$df(K, n-K-1)$$

- ~~4~~ • In simple regression, the F-test ~~duplicates~~ the t-test for the significance of the slope coefficient. However, this relation is not true for regressions with two or more slope coefficients.

$$F = \frac{t^2}{\text{error}}$$

## Example

- An analyst ran a regression and got the following result:

	Coefficient	t-statistic	p-value
Intercept	-0.5	-0.91	0.18
Slope	2	20.00	<0.001

ANOVA Table	df	SS	MSS	F
Regression	1	8000	8000	400
Error	100	2000	20	
Total	101	10000	-	

- What is the standard error of estimate?
- What is the result of the slope coefficient significance test?
- What is the result of the sample correlation?
- What is the 95% confidence interval of the slope coefficient?
- What is the result of F test?

$$2 \pm 1.96 \times 0.1$$

$$= 2 \pm 0.196 \\ [1.9804, 2.0196]$$

## Example

$$t = \frac{2-0}{SE(b_1)} = 20$$

Y

- An analyst ran a regression and got the following result:

	Coefficient	t-statistic	p-value
Intercept	-0.5 ✓	-0.91	0.18
Slope	2 ✓	20.00	<0.001 ✓

2 ~ 0.5

ANOVA Table	df	SS	MSS	E
Regression	1. ✓	8000 ✓	8000	400
Error	100 ✓	2000	20	
Total	101 ✓	10000	-	

- n=102
- What is the standard error of estimate?  $= \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{2000}{100}} = \sqrt{20}$
  - What is the result of the slope coefficient significance test?
  - What is the result of the sample correlation?  $R^2 = 0.8 + \sqrt{0.8}$
  - What is the 95% confidence interval of the slope coefficient?
  - What is the result of F test?

# Predicted Value of the Dependent Variable

- Two sources of uncertainty when using the regression model and the estimated parameters to make a prediction.

- The error term itself contains uncertainty
- Uncertainty in the estimated parameters

- Point estimate

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X$$

- Confidence interval estimate

$$\hat{Y} \pm (t_c \times s_f)$$

$$\hat{b}_1 \pm t_c \cdot SE(\hat{b}_1)$$

$$s_f = SEE(\hat{Y})$$

$t_c$  = the critical t-value with  $df=n-2$

$s_f$  = the standard error of the forecast

$$s_f = SEE \times \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)S_X^2}} = SEE \times \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X_i - \bar{X})^2}}$$



# ◆ Limitations of Regression Analysis

- Regression relations can change over time, just as correlations can.
  - Parameter instability: the problem or issue of population regression parameters that have changed over time.
- Public knowledge of regression relationships may negate their future usefulness.
  - For example, an analyst discovers that stocks with a certain characteristic have had historically very high returns. If other analysts discover and act upon this relationship, then the prices of stocks with that characteristic will be bid up and the relation no longer holds in the future.
- If the regression assumptions are violated, hypothesis tests and predictions based on linear regression will not be valid.

1.  $\hat{b}_0, \hat{b}_1 \rightarrow$  interpret, calculate

$$\left\{ \begin{array}{l} \hat{b}_1 = \frac{\text{Cov}(x, y)}{\sigma^2(x)} \\ \hat{b}_0 = \bar{y} - \hat{b}_1 \times \bar{x} \end{array} \right.$$

2. Six Assumptions
- $y$  与  $x$  之间有线性关系
  - $x$  是 not Random
  - $E(\varepsilon_i) = 0$
  - $\sigma^2(\varepsilon_i) = \text{常数}$
  - $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$
  - $\varepsilon_i \sim N(0, \sigma^2)$

3.  $SEE = \sqrt{\frac{SSE}{n-2}}$  越小越好

$$\left\{ \begin{array}{l} R^2 = 1 - \frac{SSE}{SST} = \frac{RSS}{SST} \\ R^2 = r^2 \text{ (相关系数)} \end{array} \right.$$

越大越好

Comparison .  $R^2$  / multiple  $R$  / regress

#### 4. Significance test of slope

~~$t = \hat{b}_1 - b \over SE(\hat{b}_1)$~~

$$CI: \hat{b}_1 \pm t_{\alpha/2} \cdot SE(\hat{b}_1)$$

# 5. ANOVA

$\downarrow$

$R^2$

$$F = \frac{\frac{RSS/1}{SSE/n-2}}{\frac{RSS/K}{SSE/(n-K-1)}} = \frac{MSR}{MSE}$$

df  $(K, n-K-1)$

$\downarrow$  单元

F-test  $\Leftrightarrow t$  test

$$(F = t^2)$$

b. forecast

$$\hat{Y} = b_0 + b_1 \cdot X$$

$$(\hat{Y} \pm t_{\alpha} \cdot \text{SEE}_{\hat{Y}})$$

$$\xrightarrow{\quad} f(\text{SEE})$$

- Kenneth asked several questions about regression analysis. Which of the choices provided is the best answer to each of McCoin's questions?

### Regression Statistics

R-squared ✓ 0.7436

Standard error ~~of estimation~~ 0.0213 SEE

Observations 24

ANOVA	df	SS	MSS	F	Significance F
Regression	1	0.029	0.029000	63.81	0
Residual	22	0.010	0.000455		<i>p-value</i>
Total	23	0.040			

	Coefficients	Standard Error	t-Statistic	p-Value
Intercept	0.077	0.007	11.328	0
Slope	0.826	0.103	7.988	0

# Example

$R^2$

1. What is the value of the coefficient of determination?

A. 0.8261.

B. 0.7436.

C. 0.8623.

$$\frac{0.629}{0.04} = \frac{0.63}{0.04} = 25\%$$

2. Suppose that you deleted several of the observations that had small residual values. If you re-estimated the regression equation using this reduced sample, what would likely happen to the standard error of the estimate and the  $R$ -squared?

Standard Error of the Estimate

A. Decrease

B. Decrease

C. Increase

$R$ -Squared

Decrease

Increase

Decrease

## Example

3. What is the correlation between  $X$  and  $Y$ ?

- A. -0.7436.
- B. 0.7436.
- C. 0.8623.

$$\sqrt{R^2}$$

4. Is the relationship between independent variable and dependent variable significant at the 5 percent level?

- A. No, because the  $R$ -squared is greater than 0.05.
- B. ~~No, because the p-values of the intercept and slope are less than 0.05.~~
- C. ~~Yes,~~ Yes, because the  $p$ -values for  $F$  and  $t$  for the slope coefficient are less than 0.05.

Yes

P

t



# The Basics of Multiple Regression

- **Multiple linear regression** allows us to determine the effect of more than one independent variable on a particular dependent variable.
- The multiple linear regression model

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + \varepsilon \quad \checkmark$$

- $X_i$  = the  $i$ th observation of the independent variable,  $i=1, 2, \dots, k$
- $b_0$  = the intercept of the equation
- $b_1, \dots, b_k$  = the slope coefficients for each of the independent variables

- Predicted value of the dependent variable

$$\hat{Y} = \hat{b}_0 + \hat{b}_1\hat{X}_1 + \hat{b}_2\hat{X}_2 + \dots + \hat{b}_k\hat{X}_k$$

# Interpreting The Multiple Regression Results

- The slope coefficients in a multiple regression are known as **partial regression coefficients**.
  - In normal cases, a partial regression coefficient measures the expected change in the dependent variable for a 1-unit increase in an independent variable, holding all the other independent variables constant.
  - A partial regression coefficient in a **log-log regression model** measures the expected proportional changes in the dependent variable for a 1-unit proportional changes in the independent variable, holding all the other independent variables constant.
    - ✓  $In(Y) = b_0 + b_1 \ln(X_1) + b_2 \ln(X_2) + \dots + b_k \ln(X_k) + \varepsilon$
- The **intercept coefficient** is interpreted as the value of Y when  $X_1, X_2, \dots, X_k$  are all equal to zero.

$$\ln(\alpha x) = x\%$$



$$b_1 \ln(\alpha y) \rightarrow y\%$$

# ◆ Interpreting The Multiple Regression Results

- The slope coefficients in a multiple regression are known as **partial regression coefficients**.

- In normal cases, a partial regression coefficient measures the expected change in the dependent variable for a 1-unit increase in an independent variable, holding all the other independent variables constant.

*X增加 1 -> Y增加 b<sub>1</sub>*

- A partial regression coefficient in **a log-log regression model** measures the expected proportional changes in the dependent variable for a 1-unit proportional changes in the independent variable, holding all the other independent variables constant.

*X增加 1%, Y增加 b<sub>1</sub> %*

$$\checkmark \ln(Y) = b_0 + b_1 \ln(X_1) + b_2 \ln(X_2) + \dots + b_k \ln(X_k) + \varepsilon$$

- The **intercept coefficient** is interpreted as the value of Y when  $X_1, X_2, \dots, X_k$  are all equal to zero.

*↓*

*b<sub>0</sub> %*

# Multiple Regression Assumptions

## ➤ The assumptions of the multiple linear regression

- The relationship between the dependent variable,  $Y$ , and the independent variables,  $X_1, X_2, \dots, X_k$ , is linear
- The independent variables are not random. And no exact linear relation exists between two or more of the independent variables X与X之间无强线性关系
- The expected value of the error term, conditioned on the independent variables, is zero:  $E(\varepsilon_i | X_1, X_2, \dots, X_k) = 0$  零均值
- The variance of the error term is the same for all observations (The error terms are homoskedastic)
- The error term is uncorrelated across observations ( $E(\varepsilon_i \varepsilon_j) = 0$  for all  $i \neq j$ )
- The error term is normally distributed.

# Hypothesis Testing about Regression Coefficient

## ➤ Significance test for a regression coefficient

- $H_0: b_j = 0$  ✓
- Test statistic:  $t = \frac{\hat{b}_j - 0}{S_{\hat{b}_j}}$  df = n-k-1

## ➤ p-value: the smallest significance level for which the null hypothesis can be rejected

- Reject  $H_0$  if p-value <  $\alpha$
- Fail to reject  $H_0$  if p-value >  $\alpha$

## ➤ Regression coefficient confidence interval

- $\hat{b}_j \pm (t_c \times S_{\hat{b}_j})$

# Regression Coefficient F-test

- How to test the regression's **overall significance?**
  - If none of the independent variables in a regression model helps explain the dependent variable, **the slope coefficients should all equal 0.**
- In a multiple regression, however, we **cannot** test the null hypothesis that *all* slope coefficients equal 0 based on **t-tests that each individual slope coefficient equals 0**, because the individual tests do not account for the **effects of interactions among the independent variables.**
- To test the null hypothesis that all of the slope coefficients in the multiple regression model are jointly equal to 0, we must use an **F-test**.
  - The F-statistic measures **how well the regression equation explains the variation in the dependent variable.**

# Regression Coefficient F-test

## Define hypothesis:

- $H_0: b_1 = b_2 = b_3 = \dots = b_k = 0$
- $H_a: \text{at least one } b_j \neq 0 \text{ (for } j = 1, 2, \dots, k)$

原假设

备择假设

## F-statistic:

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

- $k, n-k-1$  are the degrees of freedom for an F-test (there are  $k+1$  estimated coefficients in multiple regression)

## Decision rule

- Reject  $H_0$  : if F-statistic  $> F_\alpha(k, n-k-1)$

## Note that we use a "one-tailed" F-test.

# ◆ Analysis of Variance (ANOVA)

- Finding F-Statistics in an ANOVA Table

	df	SS	MSS	F	Significance F
Regression	k	RSS	$MSR = RSS/k$	$MSR/MSE$	P-value
Residual	$n - k - 1$	SSE	$MSE = SSE/(n - k - 1)$		
Total	$n - 1$	SST	-		

- Standard error of estimate

$$SEE = \sqrt{\frac{SSE}{n - k - 1}} = \sqrt{MSE}$$

# Adjusted R<sup>2</sup>

## ➤ R<sup>2</sup> and adjusted R<sup>2</sup>

- In a multiple linear regression, R<sup>2</sup> is less appropriate as a measure of whether a regression model fits the data well (goodness of fit).
  - ✓ We can increase R<sup>2</sup> simply by including many additional independent variables that explain even a slight amount of the previously unexplained variation, even if the amount they explain is not statistically significant.

- Function of adjusted R<sup>2</sup>

$$\bar{R}^2 = 1 - \frac{\text{SSE}_{(n-k-1)}}{\text{SST}_{(n-1)}} = 1 - \left[ \left( \frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

- ✓ adjusted for degrees of freedom
- ✓ if k>=1, R<sup>2</sup> is strictly greater than adjusted R<sup>2</sup>
- ✓ adjusted R<sup>2</sup> may be less than zero
- ✓ a high adjusted R<sup>2</sup> does not necessarily mean the correct choice of variables

# Adjusted R<sup>2</sup>

## R<sup>2</sup> and adjusted R<sup>2</sup>



$$\bar{R}^2$$

- In a multiple linear regression, R<sup>2</sup> is less appropriate as a measure of whether a regression model fits the data well (goodness of fit).
  - We can increase R<sup>2</sup> simply by including many additional independent variables that explain even a slight amount of the previously unexplained variation, even if the amount they explain is not statistically significant.
- Function of adjusted R<sup>2</sup>

very few / little

$$\bar{R}^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} = 1 - \left[ \left( \frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

## ➤ $R^2$ and adjusted $R^2$

- In a multiple linear regression,  $R^2$  is less appropriate as a measure of whether a regression model fits the data well (goodness of fit).
  - ✓ We can increase  $R^2$  simply by including many additional independent variables that explain even a slight amount of the previously unexplained variation, even if the amount they explain is not statistically significant.

- Function of adjusted  $R^2$

$$= 1 - \left( \frac{SSE}{SST} \right) \times \frac{n-1}{n-k-1}$$

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SST}{(n-1)}} = 1 - \left[ \left( \frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

$$\overbrace{\quad}^{?} \quad \bar{R}^2 < R^2$$

- ✓ adjusted for degrees of freedom
- ✓ if  $k >= 1$ ,  $R^2$  is strictly greater than adjusted  $R^2$
- ✓ adjusted  $R^2$  may be less than zero
- ✓ a high adjusted  $R^2$  does not necessarily mean the correct choice of variables

significant.

- Function of adjusted  $R^2$

$$= 1 - \left| \frac{SSE}{SST} \right| \times \left( \frac{n-1}{n-k-1} \right)$$

✓

$\frac{SSE}{RSS}$



$$\bar{R}^2 = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SST}{(n-1)}} = 1 - \left[ \left( \frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

for penalty

✓

$\frac{1}{3} R^2$

- ✓ adjusted for degrees of freedom
- ✓ if  $k >= 1$ ,  $R^2$  is strictly greater than adjusted  $R^2$   $\xrightarrow{?} \bar{R}^2 < R^2$
- ✓ adjusted  $R^2$  may be less than zero
- ✓ a high adjusted  $R^2$  does not necessarily mean the correct choice of variables

GOLDEN FUTURE

multiple R =

A<sup>2</sup>R | R<sup>2</sup>

# Dummy Variables

- Often, financial analysts need to use qualitative variables as independent variables in a regression. One type of qualitative variable, called a dummy variable 定性变量
  - Takes on a value of 1 if a particular condition is true and 0 if that condition is false
  - If we want to distinguish among  $n$  categories, we need  $n - 1$  dummy variables
- Not all qualitative variables are simple dummy variables.
  - For example, in a trinomial choice model (a model with three choices), a qualitative variable might have the value 0, 1, or 2.

# Dummy Variables

➤ Illustrates the use of dummy variables

- ✓ ●  $\text{Return}_t = b_0 + b_1 \text{Jan}_t + b_2 \text{Feb}_t + \dots + b_{11} \text{Nov}_t + \varepsilon_t$ 
  - ✓  $\text{Return}_t$  = a monthly observation of returns
  - ✓  $\text{Jan}_t = 1$  if period t is in the January,  $\text{Jan}_t = 0$  otherwise
  - ✓  $\text{Feb}_t = 1$  if period t is in the February,  $\text{Feb}_t = 0$  otherwise
  - ✓ ...
  - ✓  $\text{Nov}_t = \text{Oct}_t = \dots = \text{Jan}_t = 0$  if period t is in the December.
- ✗ ● The intercept,  $b_0$ , measures the average return for stocks in December because there is no dummy variable for December.
- ✗ ● Each of the estimated coefficients for the dummy variables shows the estimated difference between returns in that month and returns for December.
  - ✓ The result of significance test for each slope coefficient indicates the significance of the difference between returns in that month and returns for December.

## ➤ Three Multiple Regression Assumption Violations

- Heteroskedasticity
- Serial correlation (autocorrelation)
- Multicollinearity

异方差

序列自相关

多重共线性

① 什么概念

② 什么影响

③ 如何检测

④ 如何修正

# ◆ Multiple Regression Assumption Violations

## ➤ Heteroskedasticity

- Heteroskedasticity refers to the situation that the variance of the error term is not constant. (i.e., the error terms are not homoskedastic)
- **Unconditional heteroskedasticity** occurs when heteroskedasticity of the error variance is not correlated with the independent variables in the multiple regression. It creates no major problems for statistical inference.
- **Conditional heteroskedasticity** is heteroskedasticity in the error variance that is correlated with (conditional on) the values of the independent variables in the regression.
  - ✓ Conditional heteroskedasticity causes the most problems for statistical inference.

## ➤ Heteroskedasticity

- Heteroskedasticity refers to the situation that the variance of the error term is not constant. (i.e., the error terms are not homoskedastic)
- Unconditional heteroskedasticity occurs when heteroskedasticity of the error variance is not correlated with the independent variables in the multiple regression. It creates no major problems for statistical inference.
- Conditional heteroskedasticity is heteroskedasticity in the error variance that is correlated with (conditional on) the values of the independent variables in the regression.
  - ✓ Conditional heteroskedasticity causes the most problems for statistical inference.

$$\text{Cov}(\varepsilon, X) \neq 0$$

— — —

# Multiple Regression Assumption Violations

## Effect of Heteroskedasticity on Regression Analysis

- Not affect the consistency of regression parameter estimators ( $\hat{b}_j$ )
  - ✓ Consistency: the larger the number of sample, the lower probability of error.
- Heteroskedasticity introduces bias into estimators of the standard error of regression coefficients.
  - ✓ t-tests for the significance of individual regression coefficients are unreliable.
    - ◆ In regression with financial data, the estimated standard errors are most likely to be underestimated and the t-statistics are inflated.
  - ✓ The F-test for the overall significance of the regression is unreliable.

$$A_t = \frac{b_1 - \sigma}{\text{ST}(b_1)}$$

✓

$$A_t = \frac{\hat{b}_t - \sigma}{\text{SE}(\hat{b}_t)}$$

不稳

more easier to reject  
 $P(I) \uparrow$

# Multiple Regression Assumption Violations

## Effect of Heteroskedasticity on Regression Analysis

- Not affect the consistency of regression parameter estimators ( $\hat{b}_j$ )  
 ✓ Consistency: the larger the number of sample, the lower probability of error.
- Heteroskedasticity introduces bias into estimators of the standard error of regression coefficients.  $SSE$   $S_{\hat{b}_1}$   $S_{\hat{b}_2}$  --  
 ✓ t-tests for the significance of individual regression coefficients are unreliable.  
 ◆ In regression with financial data, the estimated standard errors are most likely to be underestimated and the t-statistics are inflated.  
 ✓ The F-test for the overall significance of the regression is unreliable.

$SE(\hat{b}_j) \downarrow \rightarrow t \uparrow \rightarrow$  more likely to reject  $\rightarrow p(t) \uparrow$

# Multiple Regression Assumption Violations

## ➤ Detecting heteroskedasticity

- Two methods to detect heteroskedasticity

- ✓ Residual scatter plots (residual vs. independent variable)
- ✓ The Breusch-Pagan  $\chi^2$  test

殘差圖

Bp test

Bp↑  
↓

reject H<sub>0</sub>  
↓

有異方差

◆ H<sub>0</sub>: No heteroskedasticity, one-tailed test

◆ Chi-square test:  $BP = n \times R_{\text{residual}}^2$ , df=k

$$\sum^2 = C_0 + 4X + \zeta$$

△ □ Tips: Regress squared residuals with independent variable.  
X, and  $R_{\text{residual}}^2$  is the coefficient of determination.

◆ Decision rule: The H<sub>0</sub> is rejected if BP test statistic is large.

→  $s_{\hat{\beta}_1}$   $s_{\hat{\beta}_2}$

## ➤ Correcting heteroskedasticity

- Computing robust standard errors, to correct the standard error of estimated coefficients, (a.k.a, White-corrected standard error)
- Generalized least squares, modify the equation to eliminate heteroskedasticity

# Multiple Regression Assumption Violations

## ➤ Serial correlation (or autocorrelation)

- Regression errors are correlated across observations.
- Serial correlation is often found in **time series data**



✓ **Positive serial correlation** is serial correlation in which a positive error for one observation increases the chance of a positive error for another observation, in other word,  $\text{Cov}(\varepsilon_i, \varepsilon_{i+1}) > 0$ .

◆ In examining positive serial correlation, we make the common assumption that serial correlation takes the form of **first-order serial correlation**, or serial correlation between adjacent observations.

✓ **Negative serial correlation** is less seen in financial datas.

◆ In examining positive serial correlation, we make the common

$$\text{Cov}(\varepsilon_i, \varepsilon_{i+1}) > 0$$

assumption that serial correlation takes the form of first-order serial correlation, or serial correlation between adjacent observations.

Second order

$$\text{Cov}(\varepsilon_i, \varepsilon_{i+2}) > 0$$

# Multiple Regression Assumption Violations

## ➤ Effect of serial correlation on regression analysis

### ● Positive serial correlation → Type I error & F-test unreliable

- ✓ Not affect the consistency of estimated regression coefficients.
- ✓ F-statistic to test for overall significance of the regression may be inflated because the mean squared error will tend to underestimate the population error variance.
- ✓ Standard errors for the regression coefficient are artificially small → the estimated t-statistics to be overestimated → the prob. of type I error increased.

$\underline{SE(\hat{b}_1)} \downarrow \rightarrow \underline{t \uparrow} \rightarrow$  more likely to reject  
 $\rightarrow P(I)$

# ◆ Multiple Regression Assumption Violations

## ➤ Detecting serial correlation

- Durbin-Watson test

✓  $H_0$ : No serial correlation

$$DW = \frac{\sum_{t=2}^T (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^T \hat{\epsilon}_t^2} \approx 2(1 - r)$$

Dw test

✓

$$r \rightarrow P(\epsilon_t, \epsilon_{t+1})$$

✓ Decision rule

Reject  $H_0$ ,  
conclude  
positive serial  
correlation

$\left\{ \begin{array}{l} r \rightarrow 1 \quad DW \rightarrow 0 \\ r \rightarrow -1 \quad DW \rightarrow 4 \\ r \rightarrow 0 \quad DW \rightarrow 2 \end{array} \right.$   
 Inconclusive      Do not  
reject  $H_0$       Inconclusive

correlation  
 coefficient  
 Reject  $H_0$ ,  
conclude  
negative serial  
correlation

0

$d_L$

$d_U$

$4-d_U$

$4-d_L$

4

1  
2

## ➤ Detecting serial correlation

- Durbin-Watson test

✓  $H_0$ : No serial correlation

Dw test

✓  
 $r \rightarrow P(\varepsilon_t, \varepsilon_{t+1})$   
 first order

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

✓ Decision rule

Reject  $H_0$ ,

conclude

positive serial  
correlation

$r \rightarrow 1 \quad DW \rightarrow 0$

$r \rightarrow -1 \quad DW \rightarrow 4$

Inconclusive

Do not  
reject  $H_0$

correlation

coincident

Reject  $H_0$ ,

conclude

negative serial  
correlation

0

$d_L$

$d_U$

4 -  $d_U$

4 -  $d_L$

4

## ➤ Detecting serial correlation

- Durbin-Watson test

✓  $H_0$ : No serial correlation

~~3.5~~

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

✓ Decision rule

Reject  $H_0$ ,  
conclude  
positive serial  
correlation

Inconclusive

1.9

Do not  
reject  $H_0$

$2 \times 2 - 1.8$

Reject  $H_0$ ,  
conclude  
negative serial  
correlation

0

$d_L$

$4 - d_U$

$4 - d_L$

4

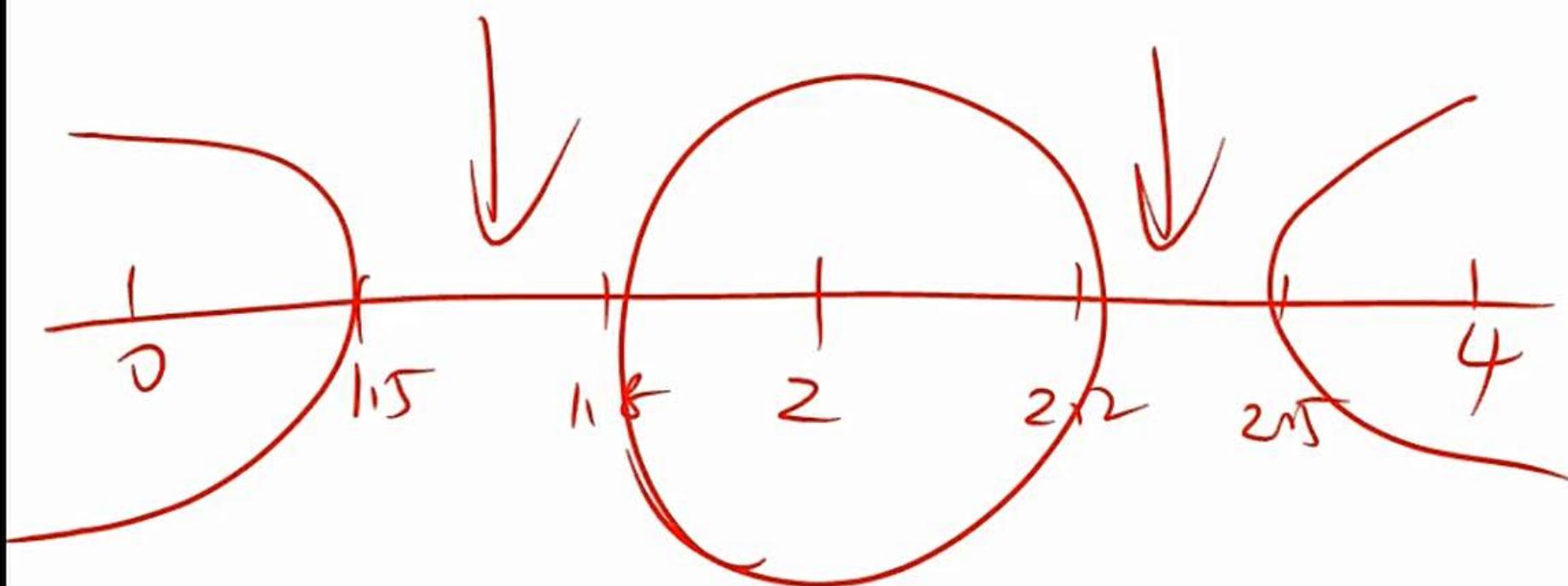
1.8

2.2

2.4

2

$$r \rightarrow D_w = 2(r)$$



$$d_L = 1.5$$

$$d_u = 1.8$$

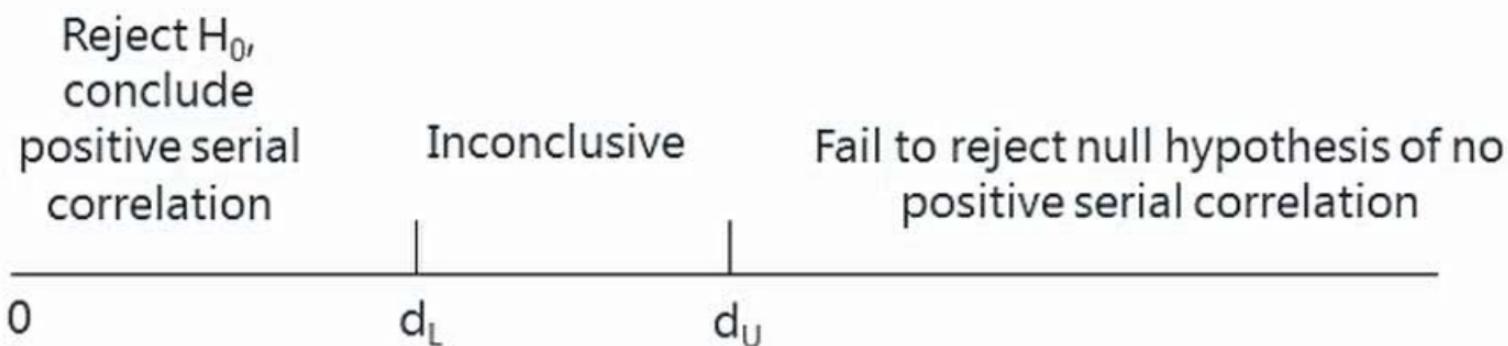
## ➤ Detecting positive serial correlation

- Durbin-Watson test

- ✓  $H_0$ : No **positive** serial correlation

- ✓  $DW \approx 2 \times (1 - r)$

- ✓ Decision rule



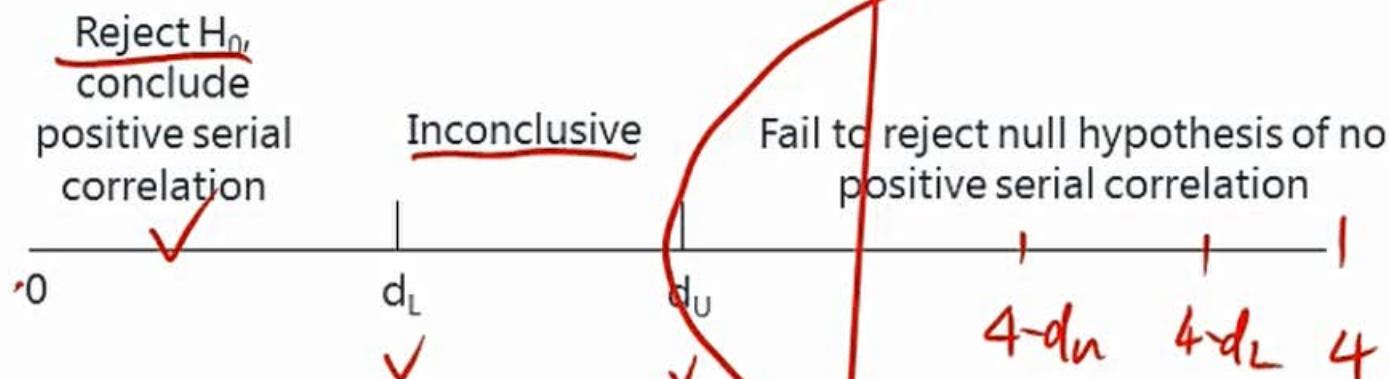
➤ Detecting positive serial correlation

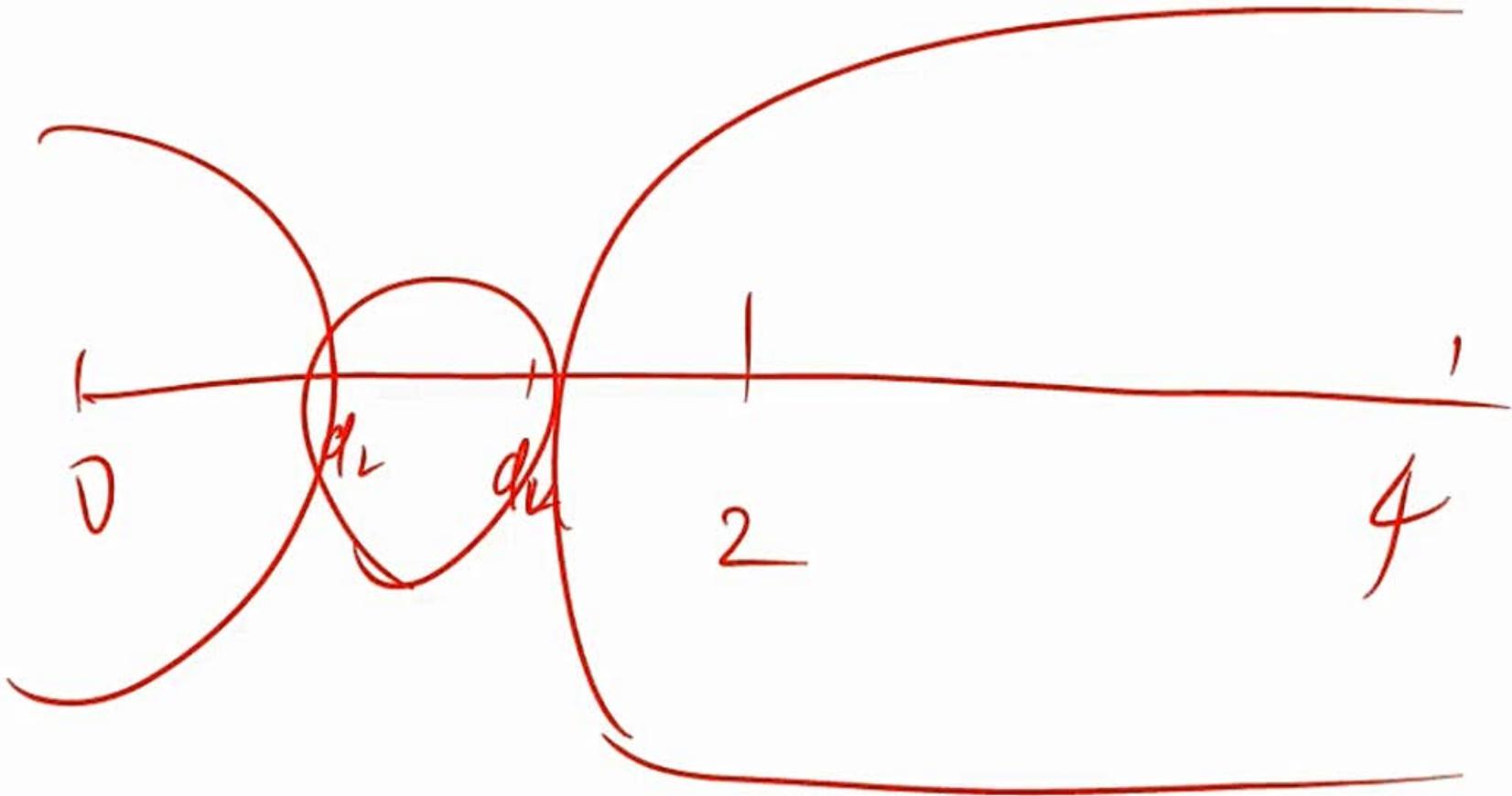
- Durbin-Watson test

✓  $H_0$ : No positive serial correlation

✓  $DW \approx 2 \times (1 - r)$

✓ Decision rule





# Multiple Regression Assumption Violations

## ➤ Methods to correct serial correlation

- Adjust the coefficient standard errors for the linear regression parameter estimates to account for the serial correlation (Recommended)
  - ✓ Hansen method or Newey-West method to adjust coefficient standard error.(a.k.a. robust standard errors)
  - ✓ They simultaneously correct for conditional heteroskedasticity
    - ◆ However, the two methods are not used when regression are not serially correlated.

Hasen → 新森 ✓  
→ 布法羅 ✓

## ➤ Multicollinearity

- Multicollinearity occurs when two or more independent variables (or combinations of independent variables) are **highly correlated** with each other.
- In practice, multicollinearity is often a matter of degree rather than of absence or presence, because approximate linear relationships among financial variables are common.

## ➤ Effect of multicollinearity on regression analysis

- Not affect the consistency of coefficient estimates  $\hat{b}_j$
- The estimates become extremely **imprecise** and **unreliable**, practically impossible to distinguish the individual impacts of the independent variables on the dependent variables
- Introduces **bias** into estimators of the standard error of regression coefficients.
  - ✓ **Inflated** standard errors for the regression coefficients → the estimated t-statistics to be underestimated → the little power of rejection

$SE(\hat{b}_j) \uparrow \rightarrow t \downarrow \rightarrow$  less likely to reject  
 $\rightarrow P(\text{II}) \uparrow$

# Multiple Regression Assumption Violations

## ➤ Two methods to detect multicollinearity

✓ 3种方法

- **Classic method:** A high  $R^2$  (and significant F-statistic) even though the t-statistics on the estimated slope coefficients are not significant.

- ✓ Insignificant t-statistics reflect inflated standard errors
- ✓ Although low t-statistics, a high  $R^2$  would reflect the overall significance of the regression

- **Occasionally suggested method:** Using the magnitude of pairwise correlations among the independent variables to assess multicollinearity.

- ✓ high pairwise correlations among the independent variables can usually indicate multicollinearity
- ✓ The method is not adequate or precise

$r(x_i, x_j)$

## ➤ Methods to correct multicollinearity

- Excluding one or more of the regression variables.

方法

# Summary of Assumption Violations

Assumption violation	Impact	Detection	Solution
Conditional Heteroskedasticity ✓	Incorrect standard errors	<ul style="list-style-type: none"> <li>✓ Residual scatter plots ✓</li> <li>✓ Breusch-Pagen <math>\chi^2</math>-test (<math>BP = n \times R^2_{\text{residual}}</math>)</li> </ul>	<ul style="list-style-type: none"> <li>✓ Robust standard errors (White)</li> <li>✓ Generalized least squares</li> </ul>
<del>Positive</del> Serial correlation	Incorrect standard errors	<ul style="list-style-type: none"> <li>✓ Residual scatter plots ✓</li> <li>✓ Durbin-Watson test (<math>DW \approx 2 \times (1 - r)</math>)</li> </ul>	<ul style="list-style-type: none"> <li>✓ Hansen method or Newey-West method</li> </ul>
Multicollinearity	high $R^2$ and low t-statistics  ↓ & SET	<ul style="list-style-type: none"> <li>✓ t-tests: fail to reject <math>H_0</math>; F-test: reject <math>H_0</math>; <math>R^2</math> is high</li> <li>✓ High magnitude of pairwise correlations</li> </ul>	<ul style="list-style-type: none"> <li>✓ Remove one or more independent variables</li> </ul>



# Model Misspecification

## ➤ Model Misspecification

- Model misspecification refers to the incorrect set of variables included in the regression and or the incorrect regression equation's functional form.
- Model misspecifications will cause the estimated regression coefficients to be **inconsistent**, leading to invalidation of statistical inference using OLS
  - ✓ However, the occurrence of heteroskedasticity, serial correlation (autocorrelation) and multicollinearity does not change the consistency of the estimated regression coefficients.

# Model Misspecification

## ➤ The principle of model specification

- The model should be grounded in cogent **economic reasoning**.
- The functional form chosen for the variables in the regression should be appropriate given the **nature of the variables**.
- The model should be **parsimonious**.
- The model should be **examined for violations** of regression assumptions before being accepted.
- The model should be tested and be found useful **out of sample** before being accepted.

# ◆ Model Misspecification

## ➤ The principle of model specification

- The model should be grounded in cogent **economic reasoning**.
- The functional form chosen for the variables in the regression should be appropriate given the **nature of the variables**.
- The model should be parsimonious. 精简 / 小统
- The model should be examined for violations of regression assumptions before being accepted.
- The model should be tested and be found useful out of sample before being accepted.

大选至简



test - in sample

validate - out of sample

➤ Three types of model misspecification

- Misspecified functional form
- Time-Series Misspecification
- Other Types of Time-Series Misspecification



## ➤ Misspecified functional form.

- One or more important variables could be omitted from regression.
  - ✓ e.g. If the true regression model is  $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i$ , but we estimate the model  $Y_i = a_0 + a_1 X_{1i} + \varepsilon_i$ , however,  $X_{2i}$  is a very important indicator.
- One or more of the regression variables may need to be transformed before estimating the regression.
  - ✓ e.g. log-log regression
- The regression model pools data from different samples that should not be pooled.
  - ✓ e.g. Represent the relationship between two financial variables at two different time periods



# Model Misspecification

- **Time-Series Misspecification (Regressors that are correlated with the error term)**
  - Including lagged dependent variables as independent variables in regressions with serially correlated errors.  
✓  $x_t = b_0 + b_1x_{t1} + b_2x_{t2} + \dots + b_px_{tp} + \varepsilon_t$
  - Including a function of a dependent variable as an independent variable, sometimes as a result of the incorrect dating of variables.  
✓  $x_t = b_0 + b_1f(x_t) + \varepsilon_t$
  - Independent variables are measured with error, thus our estimated model violates the assumption that the error term is uncorrelated with the independent variable.

$$\hat{Y}_t = b_0 + b_1 Y_{t-1} + X_t + \epsilon$$

$$b_1 f(Y_{t-1})$$

$$\underline{Y}_t = b_0 + b_1 (Y_{t-1} \times 2^{-1}) + X_t$$

## ➤ Time-Series Misspecification (Regressors that are correlated with the error term)

- Including lagged dependent variables as independent variables in regressions with serially correlated errors.  
✓ ✓  $x_t = b_0 + b_1x_{t1} + b_2x_{t2} + \dots + b_px_{tp} + \varepsilon_t$
- Including a function of a dependent variable as an independent variable, sometimes as a result of the incorrect dating of variables.  
✓  $\underbrace{x_t}_{\text{f}} = b_0 + b_1f(x_t) + \varepsilon_t$
- Independent variables are measured with error, thus our estimated model violates the assumption that the error term is uncorrelated with the independent variable.

$$y = b_0 + b_1 X + \varepsilon$$

$$y = b_0 + b_1(x + \frac{1}{2}\varepsilon) + \varepsilon$$

$$x + \frac{1}{2}\varepsilon$$

$$\varepsilon$$

- Independent variables are measured with error, thus our estimated model violates the assumption that the error term is uncorrelated with the independent variable.

## ➤ Other Types of Time-Series Misspecification (Nonstationarity)

- Relations among time series with trends
  - ✓ for example, the relation between consumption and GDP
- Relations among time series that may be random walks
  - ✓ time series for which the best predictor of next period's value is this period's value). Exchange rates are often random walks.

## Qualitative Dependent Variables

- **Qualitative dependent variables** are dummy variables used as dependent variables instead of as independent variables.
  - **Probit and logit models** estimate the probability of a discrete outcome given the values of the independent variables used to explain that outcome.
    - ✓ Both models must be estimated using **maximum likelihood methods**
  - **Discriminant models** yields a linear function, similar to a regression equation, which can then be used to create an overall score.
    - ✓ Altman's Z-Score

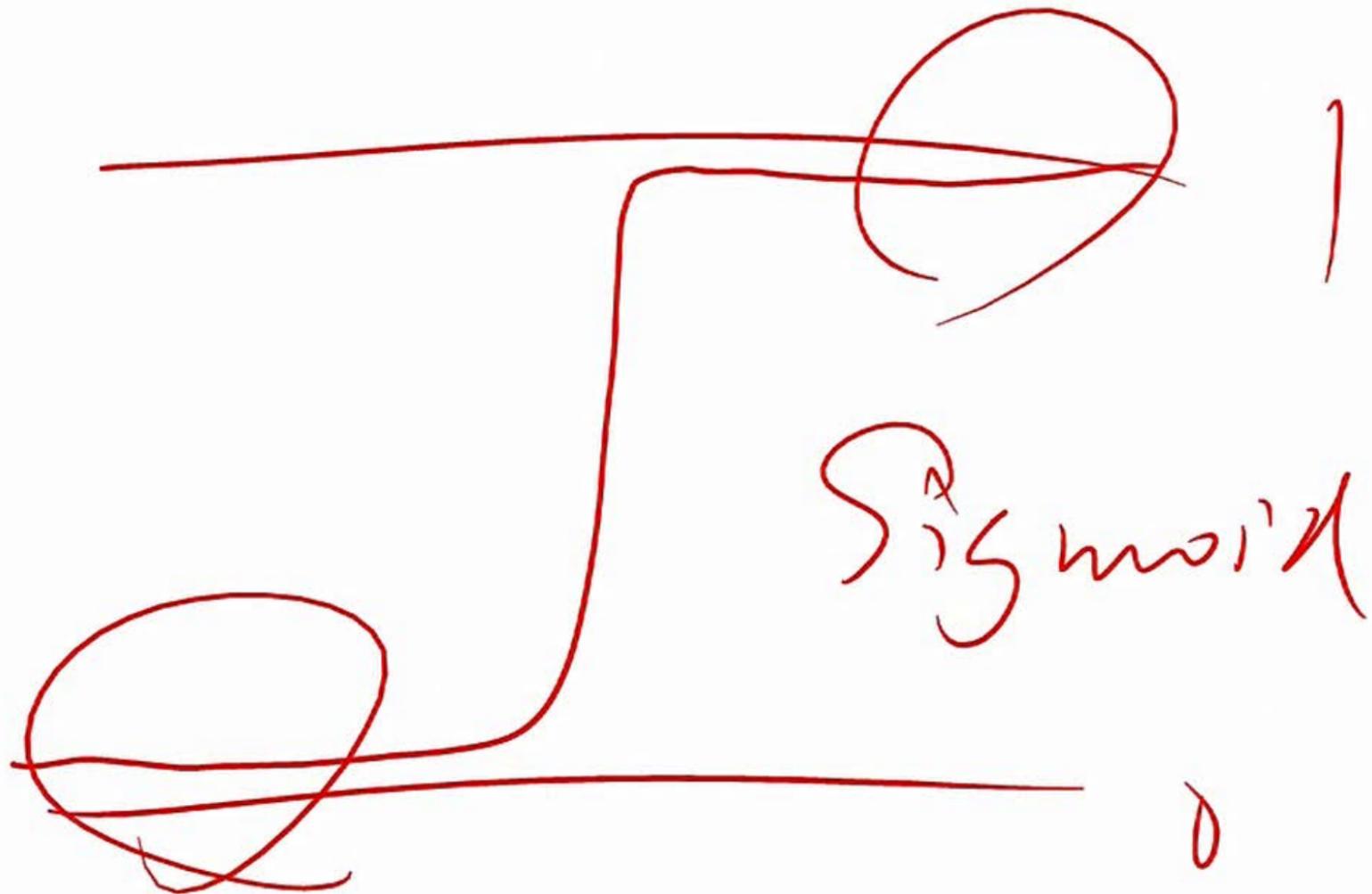
$$Y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon$$

$$\text{Y} \sim \text{both } x_1 + b_2 x_2 + h$$

$\downarrow$

$$Y(-\infty, +\infty)$$

$$\frac{1}{1+e^{-Y}} \quad (0, 1)$$



- **Probit and logit models** estimate the probability of a discrete outcome given the values of the independent variables used to explain that outcome.

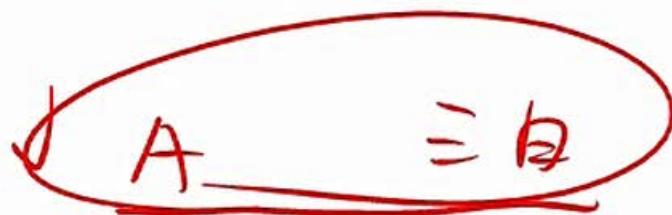
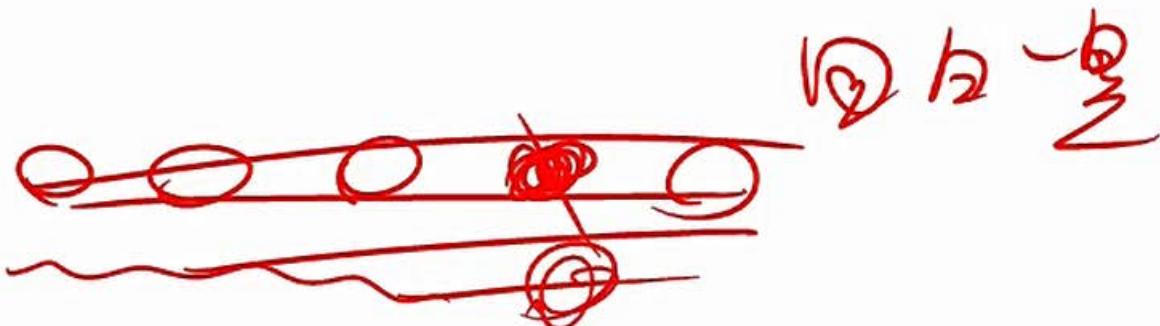
✓ Both models must be estimated using **maximum likelihood methods**

$\longleftrightarrow \rightarrow LS$

○ ○ ○

皇 白

五次



$$C_5^4 \times P_B^4 \times P_B^1$$

$$(C_5^4 \times (10\%))^4 \times (6\%)^1$$

B 三白一皇

C -白二皇

D 三皇

- Discriminant models yields a linear function, similar to a regression equation, which can then be used to create an overall score.

✓ Altman's Z-Score

# Credit Analysis

## Z-score

$$Z = 1.2 A + 1.4 B + 3.3 C + 0.6 D + 1.0 E$$

Where:

$$\underline{A = WC / TA}$$

$$B = RE / TA$$

$$C = EBIT / TA$$

---

$$D = MV \text{ of Equity} / BV \text{ of Debt}$$

$$E = Revenue / TA$$

- If  $Z < 1.81 \rightarrow$  Bankruptcy

1.  $b_1, b_2 \rightarrow$  真实参数
2. Log-Log-Regression
3. Assumptions ( $X_i \neq X_j \geq 1$  的限制条件忽略)
4. Dummy Variable  $\{v\}$  n个变量
5. Violation test ③ 个 Instrumental consistency
6. Misspecification 影响  $\nearrow$
7. Finite Quantitative dependent variable

- Hansen is developing a regression model to predict the initial return for IPOs.

**Exhibit 1. Hansen's Regression Results Dependent Variable: IPO Initial Return (Expressed in Decimal Form, i.e., 1% = 0.01)**

Variable	Coefficient ( $b_j$ )	Standard Error	t-Statistic
Intercept	0.0477	0.0019	25.11
Underwriter rank	0.0150	0.0049	3.06
Pre-offer price adjustment	0.4350	0.0202	21.53
Offer size	-0.0009	0.0011	-0.82
Fraction retained	0.0500	0.0260	1.92

**Exhibit 2. Selected ANOVA Results for Hansen's Regression**

	Degrees of Freedom (df)	Sum of Squares (SS)
Regression	4	51.433
Residual	1,720	91.436
Total	1,724	142.869

He believes that for each 1 percent increase in pre-offer price adjustment, the initial return will increase by less than 0.5 percent, holding other variables constant.

Before applying his model, Hansen asks a colleague, Phil Chang, to review its specification and results. After examining the model, Chang concludes that the model suffers from two problems: 1) conditional heteroskedasticity, and 2) omitted variable bias. Chang makes the following statements:

- **Statement 1:** "Conditional heteroskedasticity will result in consistent coefficient estimates, but both the  $t$ -statistics and  $F$ -statistic will be biased, resulting in false inferences."
- **Statement 2:** "If an omitted variable is correlated with variables already included in the model, coefficient estimates will be biased and inconsistent and standard errors will also be inconsistent."

1. The 95 percent confidence interval for the regression coefficient for the pre-offer price adjustment is closest to:

- A. 0.156 to 0.714.
- B. 0.395 to 0.475.
- C. 0.402 to 0.468.

# Example

$$0.4350 + 1.96 \times 6.0252$$

- Hansen is developing a regression model to predict the initial return for IPOs.

**Exhibit 1. Hansen's Regression Results Dependent Variable: IPO Initial Return (Expressed in Decimal Form, i.e., 1% = 0.01)**

Variable	Coefficient ( $b_j$ )	Standard Error	t-Statistic
Intercept	0.0477	0.0019	25.11
Underwriter rank	0.0150	0.0049	3.06
Pre-offer price adjustment	<u>0.4350</u>	0.0202	21.53
Offer size	-0.0009	0.0011	-0.82
Fraction retained	0.0500	0.0260	1.92

2. The *most* appropriate null hypothesis and the *most* appropriate conclusion regarding Hansen's belief about the magnitude of the initial return relative to that of the pre-offer price adjustment (reflected by the coefficient  $b_j$ ) are:

Null Hypothesis	Conclusion about $b_j$ (0.05 Level of Significance)
-----------------	---

A.  $H_0: b_j = 0.5$       Reject  $H_0$

B.  $H_0: b_j \geq 0.5$       Fail to reject  $H_0$

C.  $H_0: b_j \geq 0.5$       Reject  $H_0$

## Example

He believes that for each 1 percent increase in pre-offer price adjustment,  
the initial return will increase by less than 0.5 percent, holding other  
variables constant.

$$b_1 < 0.5\%$$

2. The *most* appropriate null hypothesis and the *most* appropriate conclusion regarding Hansen's belief about the magnitude of the initial return relative to that of the pre-offer price adjustment (reflected by the coefficient  $b_j$ ) are:

Null Hypothesis	Conclusion about $b_j$ (0.05 Level of Significance)
-----------------	---

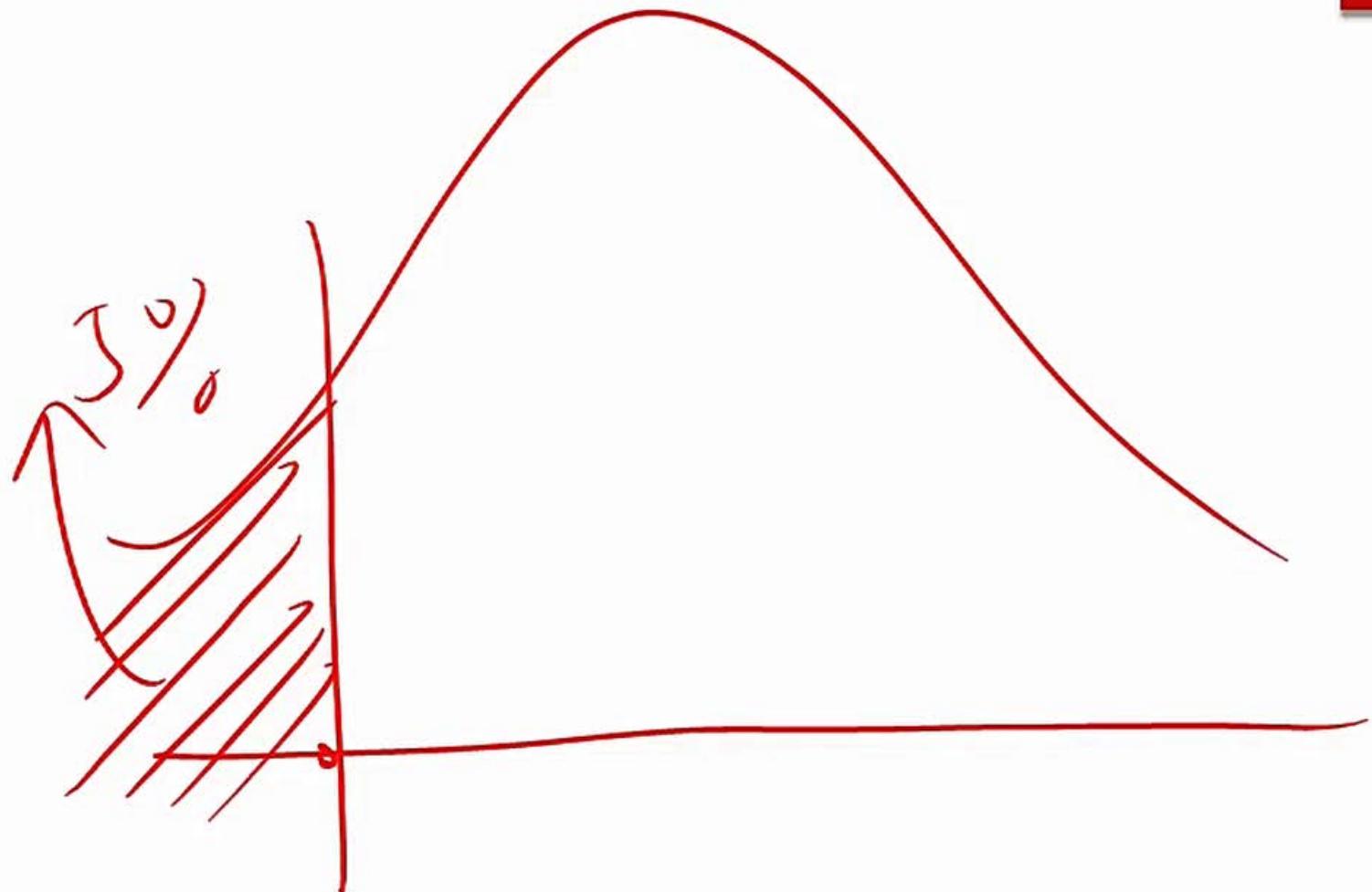
A.  $H_0: b_j = 0.5$       Reject  $H_0$

B.  $H_0: b_j \geq 0.5$  ✓      Fail to reject  $H_0$

C.  $H_0: b_j \geq 0.5$  ✓      Reject  $H_0$

$$t = \frac{14350 - 0'5}{1212} = -\frac{0'610}{1020}$$

$\rightarrow -3$



-1,65



1. The 95 percent confidence interval for the regression coefficient for the pre-offer price adjustment is closest to:

- A. 0.156 to 0.714.
- B. 0.395 to 0.475.
- C. 0.402 to 0.468.

$$t = \frac{0.4350 - 0.15}{0.202} = \frac{0.2850}{0.202}$$

$\rightarrow -3$

2. The *most* appropriate null hypothesis and the *most* appropriate conclusion regarding Hansen's belief about the magnitude of the initial return relative to that of the pre-offer price adjustment (reflected by the coefficient  $b_j$ ) are:

Null Hypothesis	Conclusion about $b_j$ (0.05 Level of Significance)
-----------------	---

- A.  $H_0: b_j = 0.5$  Reject  $H_0$
- B.  $H_0: b_j \geq 0.5$  ✓ Fail to reject  $H_0$
- C.  $H_0: b_j \geq 0.5$  ✓ Reject  $H_0$  ✓

He believes that for each 1 percent increase in pre-offer price adjustment, the initial return will increase by less than 0.5 percent, holding other variables constant.

$$b_1 < 0.5 \times$$

Before applying his model, Hansen asks a colleague, Phil Chang, to review its specification and results. After examining the model, Chang concludes that the model suffers from two problems: 1) conditional heteroskedasticity, and 2) omitted variable bias. Chang makes the following statements:

- **Statement 1:** "Conditional heteroskedasticity will result in consistent coefficient estimates, but both the  $t$ -statistics and  $F$ -statistic will be biased, resulting in false inferences." ✓
- **Statement 2:** "If an omitted variable is correlated with variables already included in the model, coefficient estimates will be biased and inconsistent and standard errors will also be inconsistent." ✓

## ► Question

3. Is Chang's Statement 1 correct?
- A. Yes.
  - B. No, because the model's  $F$ -statistic will not be biased.
  - C. No, because the model's  $t$ -statistics will not be biased.
4. Is Chang's Statement 2 correct?
- A. Yes.
  - B. No, because the model's coefficient estimates will be unbiased.
  - C. No, because the model's coefficient estimates will be consistent.

## Question

5. Hansen is concerned about the possible presence of multicollinearity in the regression. He states that adding a new independent variable that is highly correlated with one or more independent variables already in the regression model, has three potential consequences, which one is incorrect

- A. The  $R^2$  is expected to decline.
- B. The regression coefficient estimates can become imprecise and unreliable.
- C. The standard errors for some or all of the regression coefficients will become inflated.

$$\begin{array}{c} SE(b_i) \uparrow \rightarrow t \downarrow \\ \rightarrow \text{less likely} \end{array}$$

# Major Focuses of Data Analytics

- The term **Big Data**, refers to the vast amount of data including data generated from traditional sources as well as non-traditional data types.
  - A comprehensive **data analysis** is helpful for in understanding big data and machine learning.
- Data analytics generally has one of six focuses
  - Measuring correlations and understanding relationships
  - Making predictions ✓ → ↗
  - Making causal inferences → ↗ ↗
    - ✓ Causal inference focuses on establishing that a change in an independent variable causes a change in the dependent variable.
  - { ● Classifying data into distinct categories 3类
  - Sorting data into clusters with similar characteristics 集羣
  - Reducing the dimension of data or simply reduce independent variables

# What Is Machine Learning

## ➤ Definition

- A computer program that continuously modifying itself and improving the accuracy of prediction by learning from errors and experience.

## ➤ Machine learning vocabulary

- In regression analysis

✓ Y variable known as the dependent variable

✓ X variables are known as independent variables or explanatory variables

- In machine learning

✓ Y variable is called the target variable or tag variable

✓ X variables are called features

◆ Curating a dataset of features for ML processing is known as feature engineering by machine learning practitioners.

# ◆ Types of Machine Learning

## ➤ Supervised learning

- Supervised learning is machine learning that makes use of labeled training data.
- Typical data analysis tasks associated with supervised learning are classification and prediction.
  - ✓ Take credit card for example, the ML program is given processed transactions labeled ( tagged ) "fraudulent" or "non-fraudulent" and uses them to train a model in predicting fraud more accurately in new transactions.

# ◆ Types of Machine Learning

## ➤ Unsupervised learning

- Unsupervised learning is machine learning that does not make use of labeled training data.
- **Clustering** is an example of data analytics to which unsupervised learning is applied.
  - ✓ For example, we may take different firms' financial statement data and use an unsupervised ML program to cluster firms into groups based on their attributes.
  - ✓ Each cluster will contain firms that have greater overall similarity to each other than they do to firms in other clusters.
  - ✓ **Birds of a feather flock together.**

↑  
— 互  
— 互  
— 互

## Example-1

➤ Which of the following best describes machine learning? Machine learning:

- A. is a type of computer algorithm.
- B. is a set of computer-driven approaches that can be used to extract information from Big Data.
- C. is a set of computer-driven approaches adapted to extracting information from structured data.

**Answer:**

B is correct. A major application of machine learning is extracting information from Big Data. Choice A is not correct because although algorithms are used in machine learning, machine learning itself is not best described as a type of computer algorithm.

## Example-2

➤ Which of the following statements is most accurate? Machine learning:

- A. contrasts with human learning in relation to measuring performance on specific tasks.
- B. takes place when a computer program is programmed to perform specific tasks.
- C. takes place when a computer improves performance in a specific class of tasks as experience increases.

**Answer:**

C is correct. ML takes place when a computer improves performance in a specific class of tasks as experience increases.

## Example-3

➤ Which of the following statements is most accurate? When attempting to place data into groups based on their inherent similarities and differences:

- A. an unsupervised ML algorithm is used.
- B. an ML algorithm that is given tagged data is used.
- C. an ML algorithm that is given tagged data and untagged data is used.

### Answer:

A is correct. The beginning of the statement that must be completed is a description of clustering. Unsupervised ML algorithms are used in clustering.

## Example-4

➤ Which of the following statements concerning supervised learning best distinguishes it from unsupervised learning? Supervised learning involves:

- A. training on labeled data.
- B. training on unlabeled data.
- C. learning from unlabeled data.

**Answer:**

A is correct. Supervised learning computer programs are given labeled data in training, in contrast to unsupervised learning computer programs.



# Machine Learning Algorithms

- A top-level description of a limited selection of important models and procedures

Supervised Learning			Unsupervised Learning		
Penalized Regression	CART	Random Forests	Neural Networks	Clustering Algorithms	Dimension Reduction

# ◆ Supervised Learning

- Supervised learning can be divided into two categories: **regression** and **classification**
- In the context of supervised learning, the distinction between regression and classification is determined by the nature of the Y variable.
  - If the Y variable is continuous, then the task is one of **regression**
    - ✓ **penalized regression (belongs to linear regression)**
  - If the Y variable is categorical or ordinal, then it is a classification problem.
    - ✓ **classification and regression trees ( CART )**
    - ✓ **random forests**
    - ✓ **neural networks**

Regression

非

线性

单  
多  
 $y = b_0 + b_1 x^2$

多项式

$$\max(x, 0)$$

# ◆ Penalized Regression

- As with multiple linear regression, penalized regression and many other forms of linear regression are special cases of the **generalized linear model (GLM)**.
  - ✓ GLM refers to a flexible specification linear regression in which the modeler, can express preferences for model parsimony **by choice of parameters**.
- **Penalized regression** could be described as a technique of **regularization**.
  - ✓ A method that tamps down statistical variability in high-dimensional estimation or prediction problems.
- **Objectivity** : minimize errors together with an penalty term to reduce the problem of overfitting as well as remaining much prediction power.
  - In one popular type of penalized regression is LASSO ( least absolute shrinkage and selection operator).

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots$$

SEE

Penalized form

$$f(c_k)$$

$$[SEE + f(c_k)]_j$$

# ◆ Penalized Regression

- As with multiple linear regression, penalized regression and many other forms of linear regression are special cases of the **generalized linear model (GLM)**.
  - ✓ GLM refers to a flexible specification linear regression in which the modeler, can express preferences for model parsimony **by choice of parameters.**
- **Penalized regression** could be described as a technique of **regularization**.
  - ✓ A method that tamps down statistical variability in high-dimensional estimation or prediction problems.
- **Objectivity** : **minimize errors together with a penalty term** to reduce the problem of overfitting as well as retaining much prediction power.
  - In one popular type of penalized regression is **LASSO** ( least absolute shrinkage and selection operator).

$$y = 2 + 3x_1 + 4x_2 + \underline{\epsilon}$$

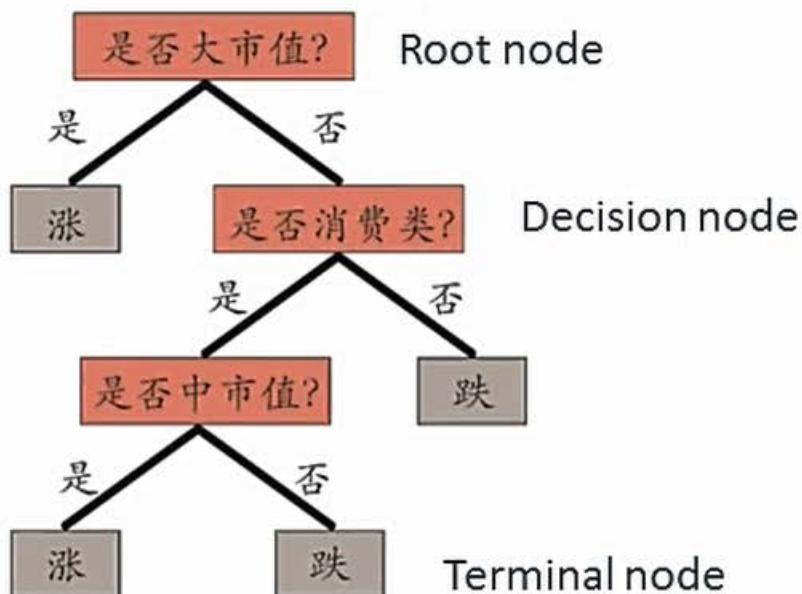
$$y = 2 + 3x + 4\underline{x_2} + \underline{\epsilon}$$

→ **Objectivity** : minimize errors together with a penalty term to reduce the problem of overfitting as well as retaining much prediction power.

- In one popular type of penalized regression is LASSO ( least absolute shrinkage and selection operator).

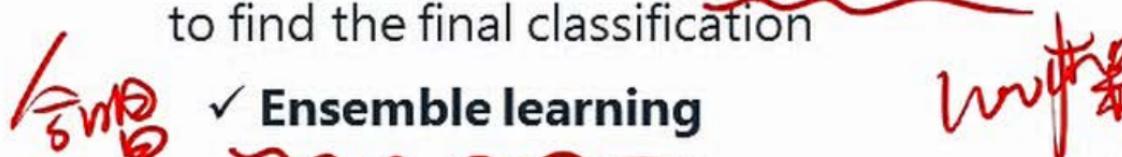
# Classification And Regression Trees

- **CART** is a common supervised ML technique that can be applied to predict either a classification problem(producing a classification tree), or a regression problem(producing a regression tree).
  - Most commonly, CART is applied where the target is binary.
- **Classification Tree ( Simple Case )**



# Random Forests

- A random forest classifier is a collection of classification trees.
  - Rather than use just one classification tree, we build several, based on random selection of features data set(variables).
  - Each tree is, therefore, slightly different from the others.
  - The last step involves a **majority vote**(also called the wisdom of crowds) to find the final classification

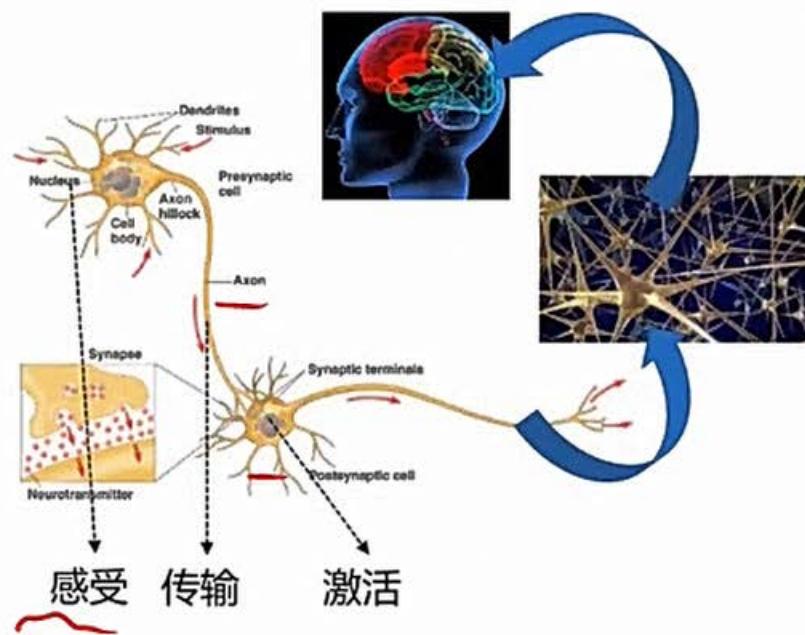


- The process involved in random forests tends to protect against **overfitting** with smaller training data set. It also reduces the **ratio of noise to signal** because errors cancel out across the collection of classification trees.

- ~~Ensemble learning~~ ✓ **Ensemble learning**
- The process involved in random forests tends to protect against overfitting with smaller training data set. It also reduces the ratio of noise to signal because errors cancel out across the collection of classification trees.

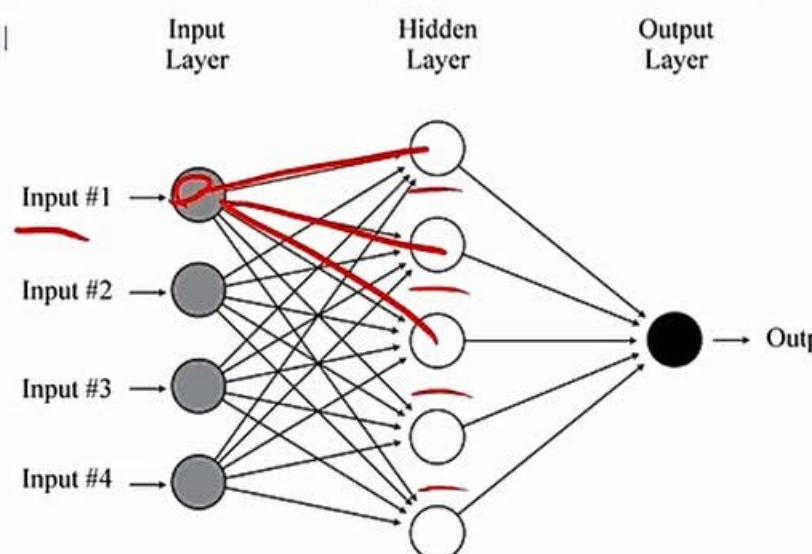
# Neural Networks

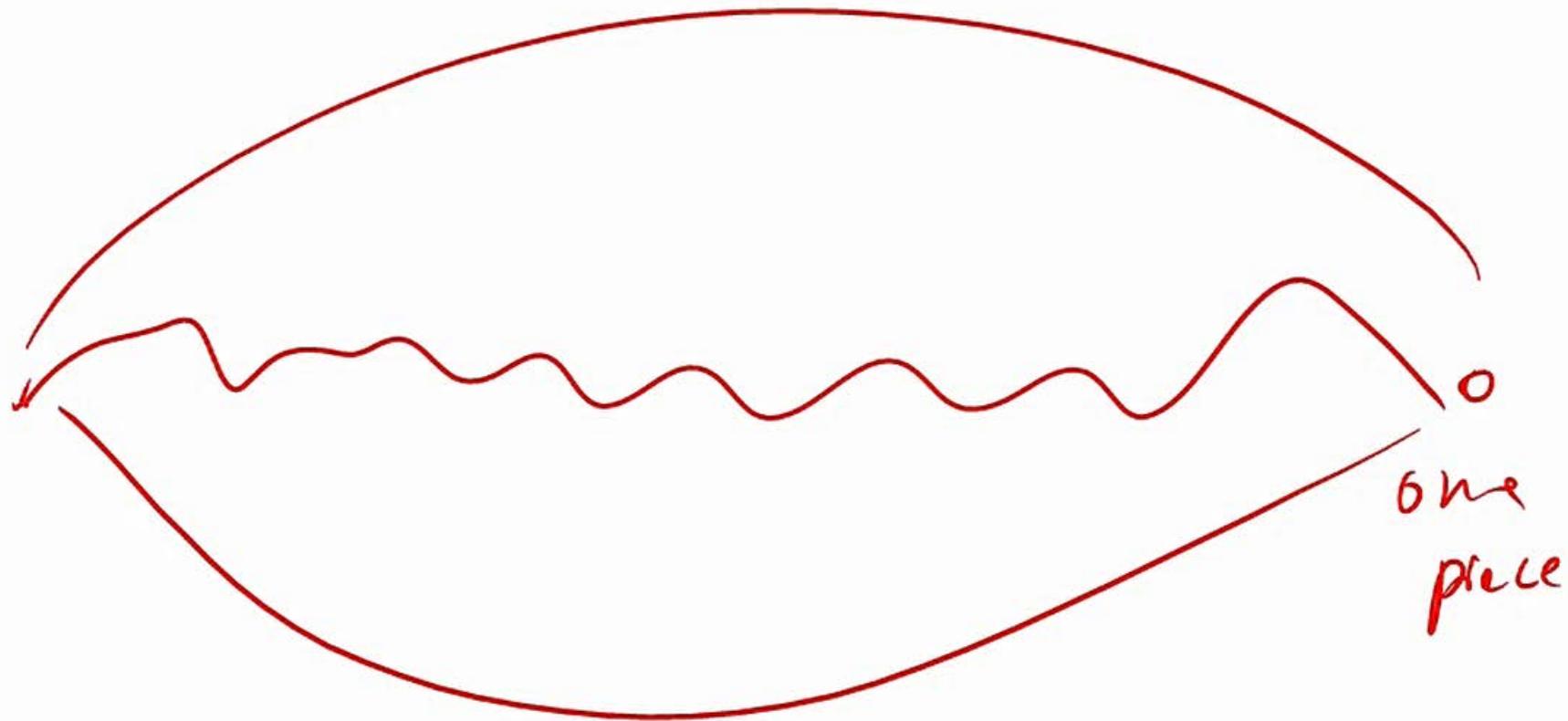
- **Neural networks** (also called artificial neural networks, or **ANNs**) have been successfully applied to a variety of tasks characterized by non-linearities and interactions among variables.



# ◆ Neural Networks

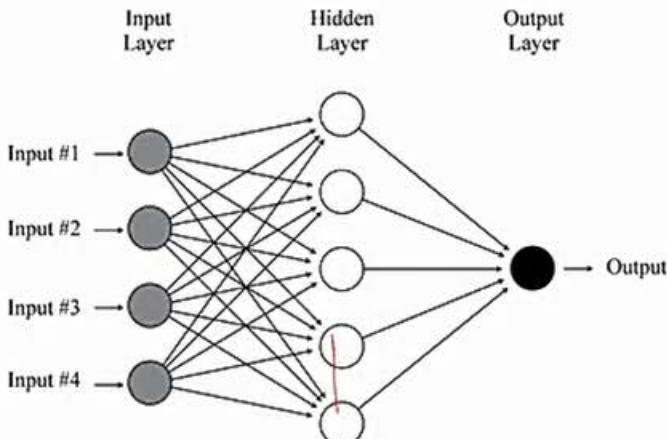
- Neural networks have three types of layers
  - an input layer
  - hidden layers
  - an output layer
- There are four nodes in input layer(four features), five nodes in the single hidden layer(five ways of transmitting data) and one node in output layer(one predict result)——these numbers are called **hyperparameter**, each node applies ai





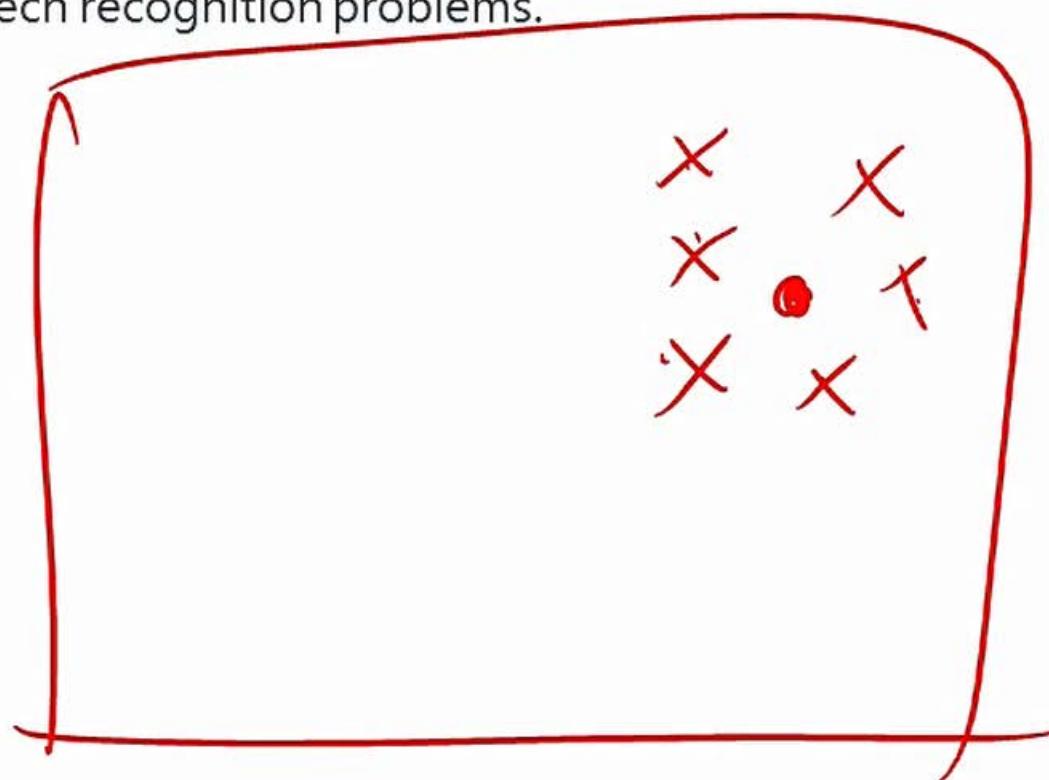
# ◆ Neural Networks

- Neural networks have three types of layers
    - an input layer
    - hidden layers
    - an output layer
  - There are four nodes in input layer(four features), five nodes in the single hidden layer(five ways of transmitting data) and one node in output layer(one predict result)—these numbers are called hyperparameter, each node applies an activation function.
4. (5) 1



# ◆ Deep Learning Nets

- Neural networks with many hidden layers ( often more than 20 ) – known as **deep learning nets ( DLNs )** – are the backbone of the artificial intelligence revolution.
  - ✓ DLNs have been shown to be useful in general for image, pattern and speech recognition problems.





# Unsupervised Learning

- It is analysis on the X variables, and there is no Y target variable set.
- Many algorithms of this type of learning address **clustering** and **dimension reduction**.

Supervised Learning			Unsupervised Learning		
Penalized Regression	CART	Random Forests	Neural Networks	Clustering Algorithms	Dimension Reduction

# Clustering Algorithms

- Approaches to clustering are often placed into one of two groups: **bottom-up clustering and top-down clustering.**

- **Bottom-up clustering**

- ✓ Bottom-up clustering starts with each observation being its own cluster and then progressively groups the observations into larger, non-overlapping clusters.
- ✓ **K-means algorithm**

easy

---

- **Top-down clustering**

- ✓ With top-down clustering, we start with all observations belonging to **a single cluster**, which is then progressively partitioned into smaller and smaller clusters.

hard

# Dimension Reduction

↓ XHRK72

## ➤ In which cases may dimension reduction be necessary?

- When there are many features in a dataset, the model becomes unnecessarily complex and “noisy”.

## ➤ One long-established statistical method for dimension reduction is **principal component analysis ( PCA )**.

PCA top 3

- PCA is used to summarize or reduce highly correlated features of data into a few main, uncorrelated **composite variables** (a composite variable is a variable that combines two or more variables that are statistically strongly related to each other ).

- ✓ The first principal component is the most volatile : It represents the most important factor for explaining the volatility in the data.
- ✓ Each subsequent principal component extracts remaining volatility, subject to the constraint that it is uncorrelated with the preceding principal components.

## Example

1. As used in supervised machine learning, regression problems involve:

- A. binary target variables.
- B. continuous target variables.
- C. categorical target variables.

Y

**Answer:**

B is correct. When the target variable is binary or categorical, the problem is a classification problem rather than a regression problem.

2. Which of the following best describes penalized regression? Penalized regression:

- ~~A.~~ is unrelated to multiple linear regression.
- ~~B.~~ involves a penalty term for the sum of squared residuals.
- C. is a category of general linear models that is used when the number of independent variables is a concern.

**Answer:**

C is correct.

$$\text{SEE} = f(\text{SSE}, n-k-1, \text{Penalty})$$

## Example

### 3. CART is best described as a type of:

- A. unsupervised ML. X
- B. a clustering algorithm based on decision trees. X
- C. a supervised ML algorithm that accounts for non-linear relationships among the features. ✓

**Answer:**

C is correct.

### 4. Neural networks are best described as an ML technique for learning:

- A. exactly modeled on the human nervous system. X
- B. based on layers of nodes when the relationships among the features are usually non-linear. ✓
- C. based on a tree structure of nodes when the relationships among the features are non-linear. X

**Answer:**

B is correct.

**CART**

## > Example

### 5. Clustering is best described as a technique in which:

- A. the grouping of observations is unsupervised.
- B. features are grouped into clusters by a top-down algorithm X
- C. observations are classified according to predetermined labels. X

#### Answer:

A is correct. Choice B is not the best choice because clustering algorithms can be either bottom up or top down.

### 6. Dimension reduction techniques are best described as means to reduce a set of features:

- A. to a manageable size.
- B. to a manageable size while controlling for variation in the data. T<sub>2</sub>
- C. to a manageable size while retaining as much of the variation in the data as possible. T<sub>1</sub>

#### Answer:

C is correct.

# ◆ Supervised Machine Learning: Training

- In a simplified view, the process to train ML models involves following steps:

1. Specify the ML technique/algorithm
2. Specify the associated hyperparameters, include the number of training cycles
3. Divide data into
  - ✓ A training sample ✓
  - ✓ A validation sample ✓
4. Evaluate learning with performance measure P, using the validation sample, and adjust ("tune") the hyperparameters
5. Repeat the training cycle the specified number of times or until the required performance level ( e.g., level of accuracy ) is obtained

- The output or artifact created by the training process is the "ML model".

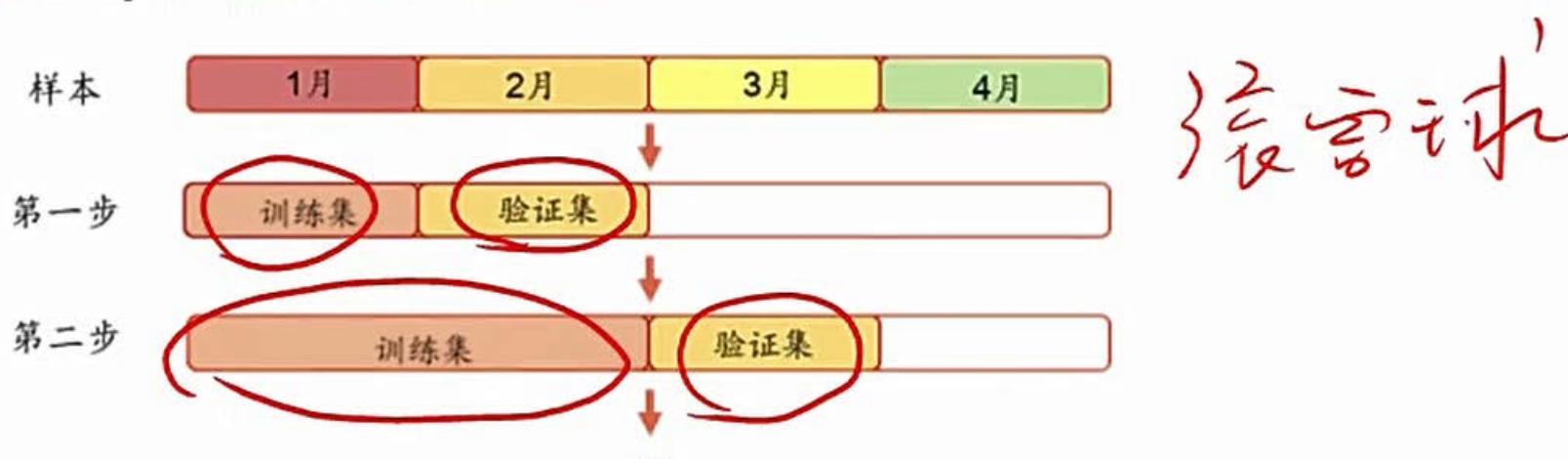
# ◆ Supervised Machine Learning: Training

- **Cross-validation is intended to control for bias in training data and is completely standard; no model validation is complete without cross-validation.**

中文

- Intuitively, the bigger the dataset, the less cross-validation is needed.
- With smaller datasets, a specific split of the data into training and validation samples may be biased, and cross-validation is a simple and effective way to address that concern.

- **Example of cross-validation**



① correlation sig. test

$Y \leftarrow CPA II$

$Y, X$  は  
变量

→ 变量  $-X$

$$P \rightarrow t = \frac{r - v}{\sqrt{\frac{1 - r^2}{n - 2}}} \text{ df}(n - 2)$$



② model

$$y = b_0 + b_1 x + \epsilon$$

③  $b_0, b_1$  OLS

④ Assumption

- ⑤ Analysis      SEE  
                  R<sup>2</sup>
- ⑥ slope sig. test  $\rightarrow t$   
overall sig. test  $\rightarrow F$
- ⑦ Construction      Application

 ↓ 進入 Dummy 变量

Correlation test ~~X1X2~~  $X_1 X_2$



$$Y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon$$



Assumption:



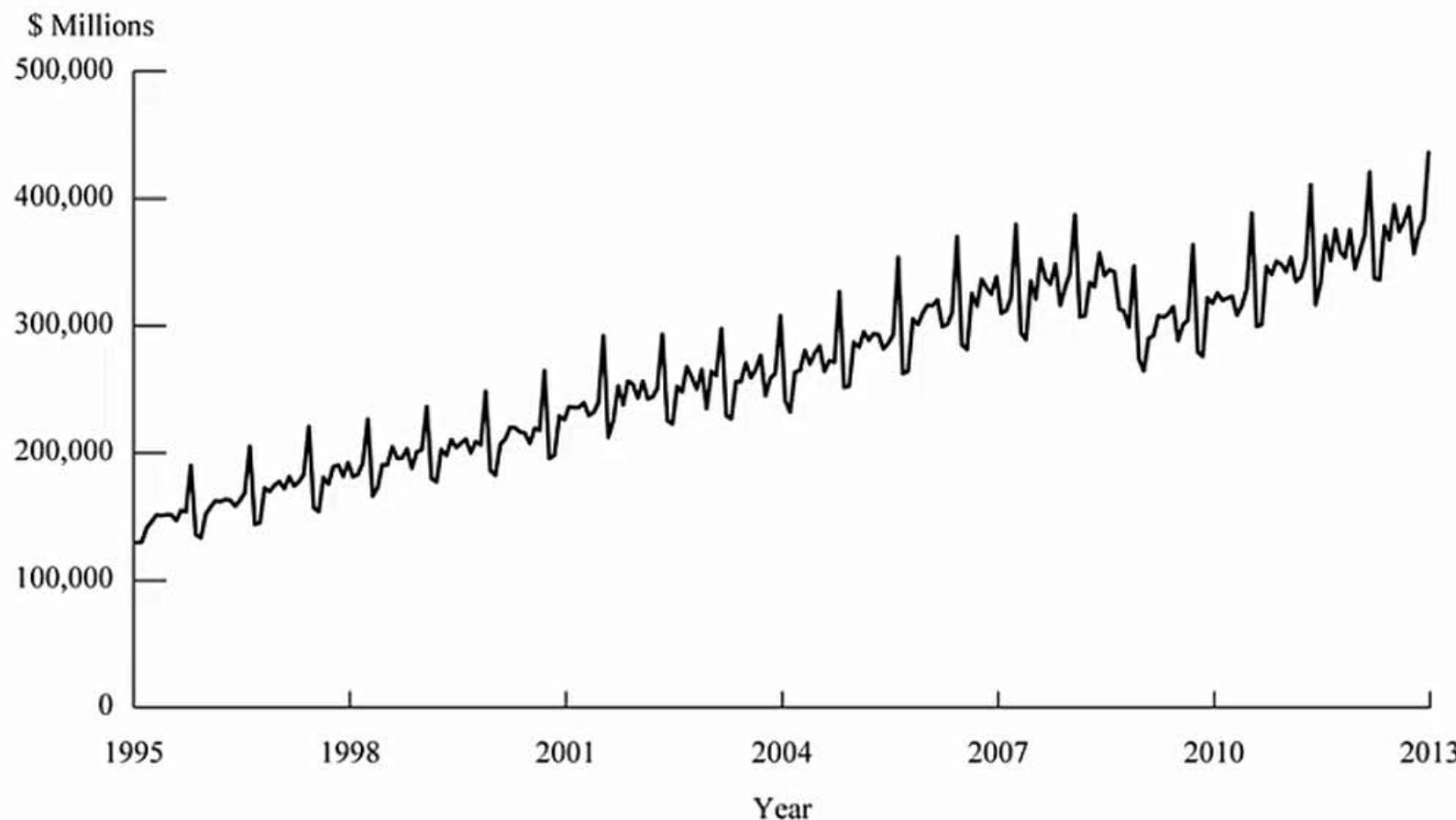
$$SE\epsilon, R^2 \rightarrow \bar{R}^2$$



(t) (F)  
Sig & test of slope / model

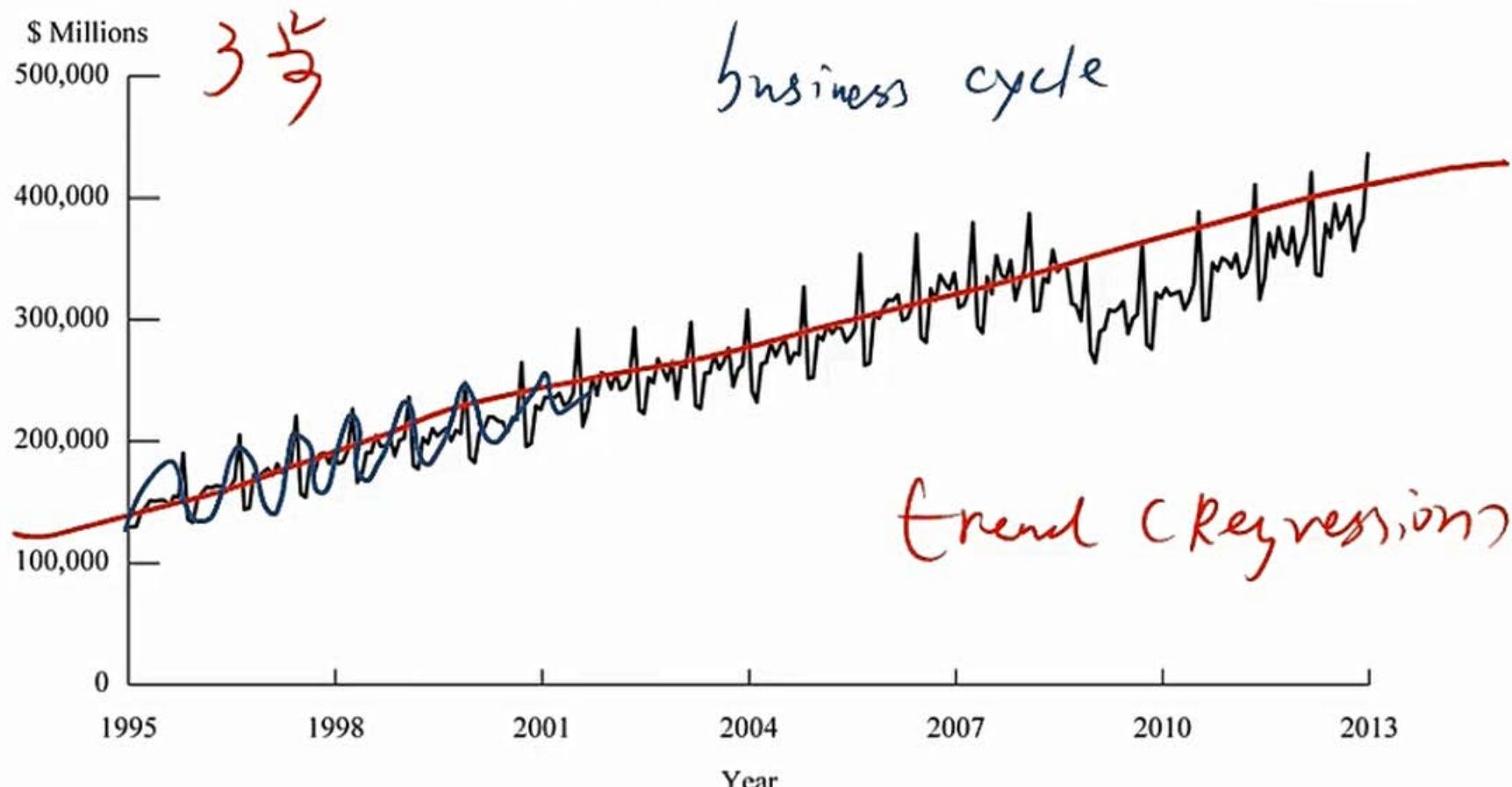
# ◆ Time Series Data and Analysis

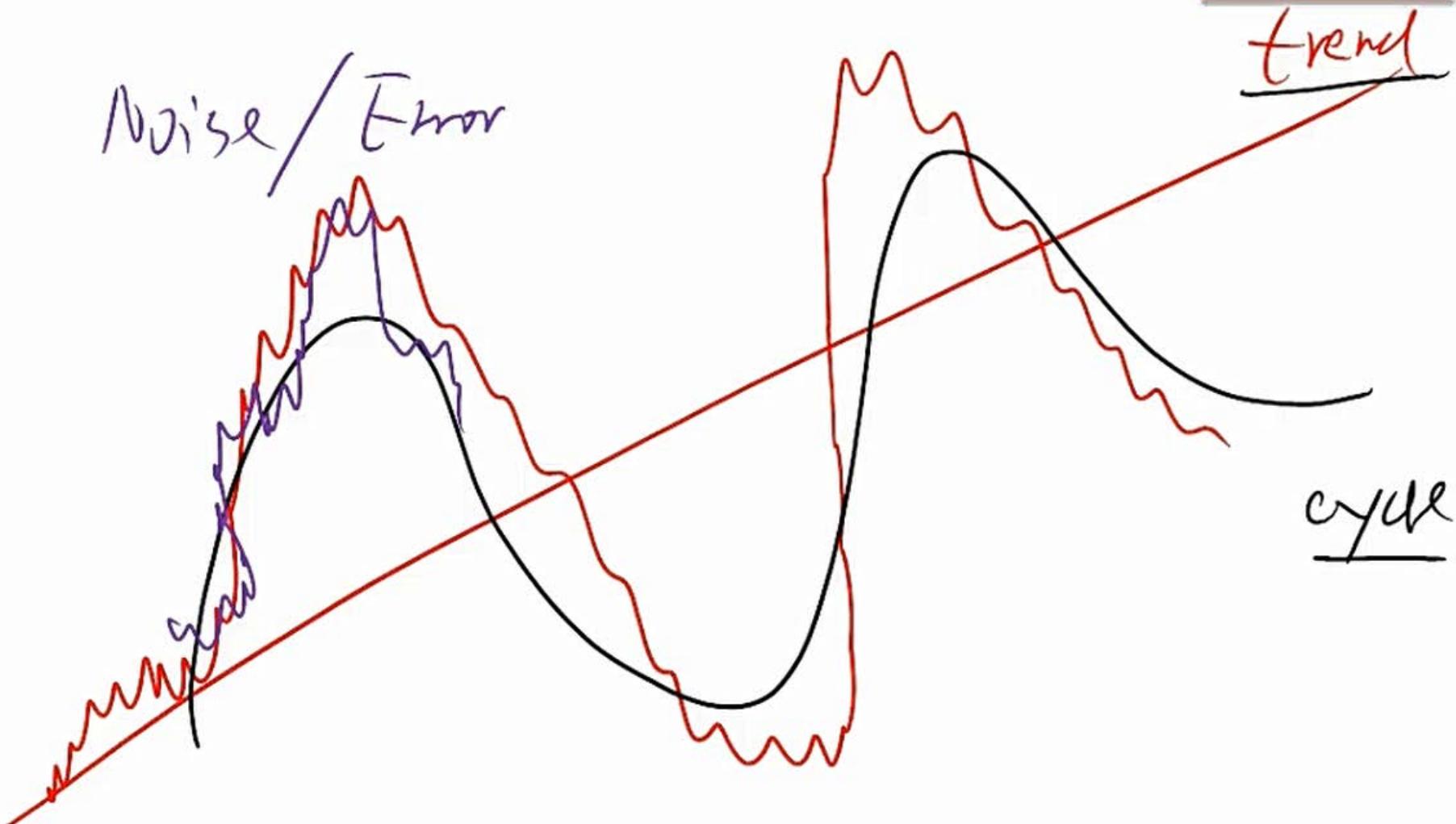
- A **time series** is a set of observations on a variable's outcomes in different time periods, taking the example of a set of **monthly U.S. retail sales data**.



# Time Series Data and Analysis

- A **time series** is a set of observations on a variable's outcomes in different time periods, taking the example of a set of **monthly U.S. retail sales data**.



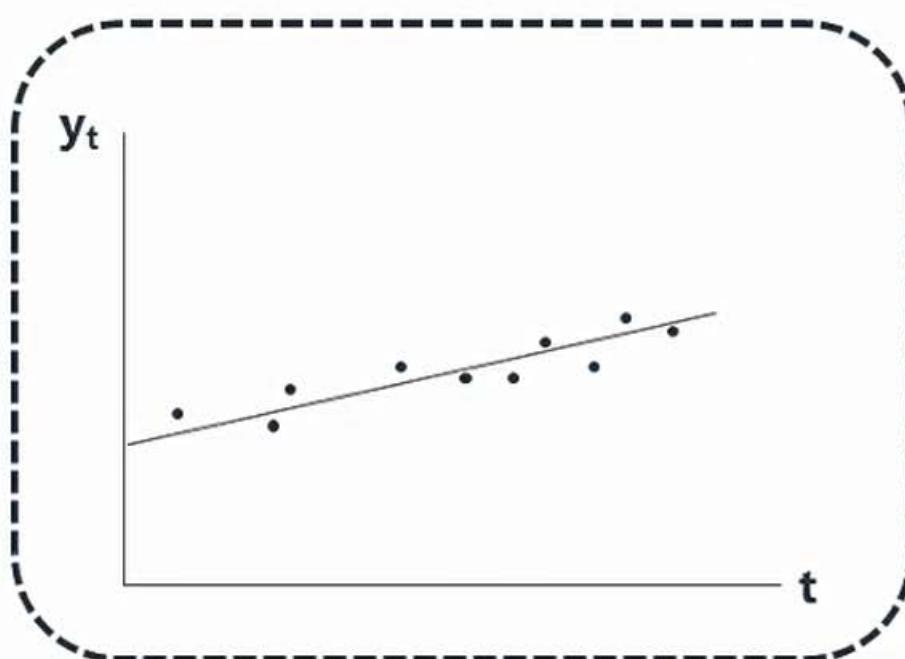


# Trend Models

## ➤ Linear trend model ✓

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

- Same as linear regression, except for that the independent variable is time  $t$  ( $t=1, 2, 3, \dots$ )



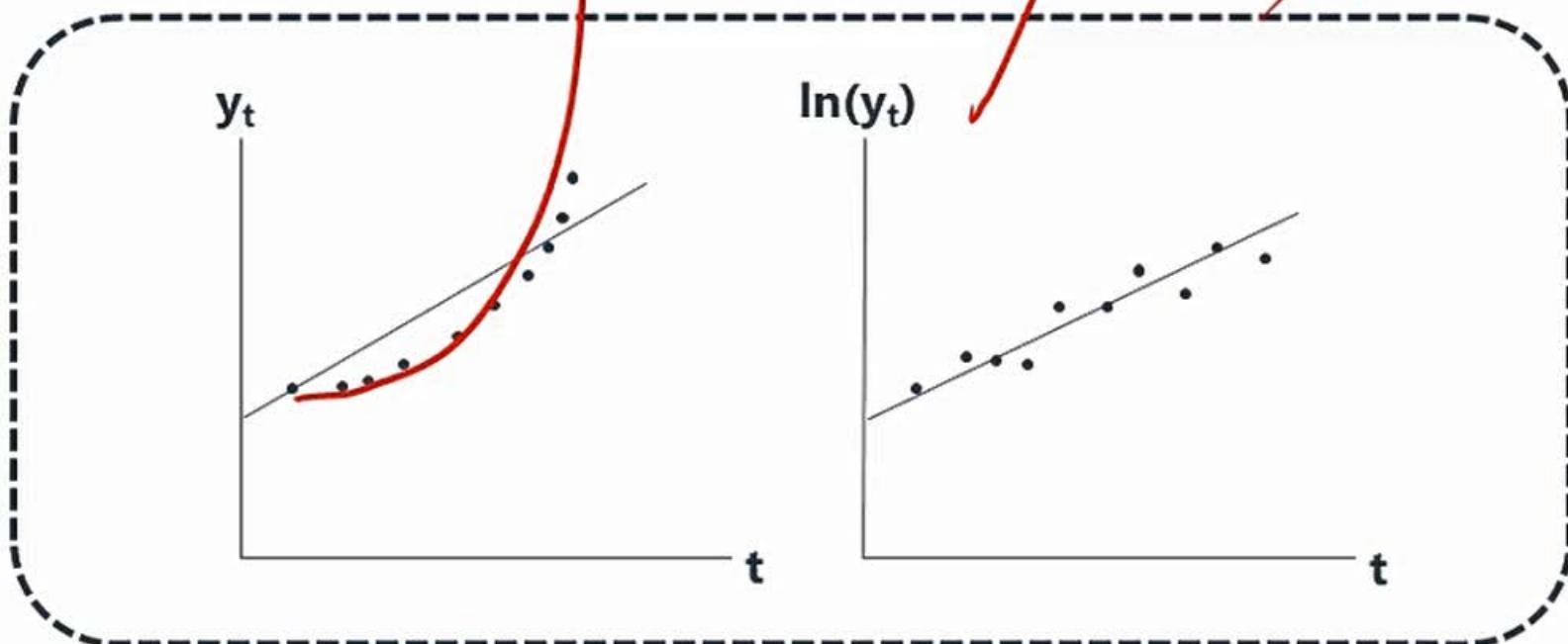
# Trend Models

## ➤ Log-linear trend model

- $\ln(y_t) = b_0 + b_1 t + \varepsilon_t$
- Model the natural log of the series using a linear trend
- Use the Durbin Watson statistic to detect autocorrelation

$$y = e^x$$

$$\ln x = x$$



# Trend Models

## ➤ How to select a trend model

- A linear trend model may be appropriate if fitting a linear trend to a time series leads to uncorrelated errors.
- A log-linear model may be more appropriate if a time series grows at an exponential rate, we can model the natural log of that series using a linear trend.

## ➤ Limitations of Trend Model

- Usually the time series data exhibit serial correlation, which means that the model is not appropriate for the time series
- **Existence of serial correlation** suggests we build better forecasting model than trend model



## Autoregressive Models (AR)

- An autoregressive model uses past values of dependent variables as independent variables

- AR(p) model

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

- AR (p): AR model of order p (p indicates the number of lagged values that the autoregressive model will include as independent variable).  
For example, a model with two lags is referred to as a second-order autoregressive model or an AR (2) model.

$$\underline{y_t} = b_0 + b_1 \underline{y_{t-1}} + \underline{\varepsilon}$$

# Autoregressive Models (AR)

- An autoregressive model uses past values of dependent variables as independent variables

- AR(p) model

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

- AR (p): AR model of order p (p indicates the number of lagged values that the autoregressive model will include as independent variable).
- For example, a model with two lags is referred to as a second-order autoregressive model or an AR (2) model.

$$x_t = b + b_1 x_{t-1} + \varepsilon$$

*AR(1)*



$$y_t = b_0 + b_1 x_1 + b_2 x_{t-3} + \zeta$$

$\rightarrow$  Sec 3)

# Autoregressive Models (AR)

## ➤ How to estimate the coefficients in AR Model?

OLS

- We can estimate an autoregressive model using ordinary least squares if
  - ✓ the time series is covariance stationary
  - ✓ the errors are uncorrelated
  - ✓ homoskedasticity of the error term variance from the independent variable, which is also called autoregressive conditional heteroskedasticity (ARCH) in AR Model.

# Autoregressive Models (AR)

## ➤ Covariance-stationary series

- The basic idea is that a time series is covariance stationary if its properties, such as mean and variance, do not change over time.

## ➤ Three conditions for covariance stationary

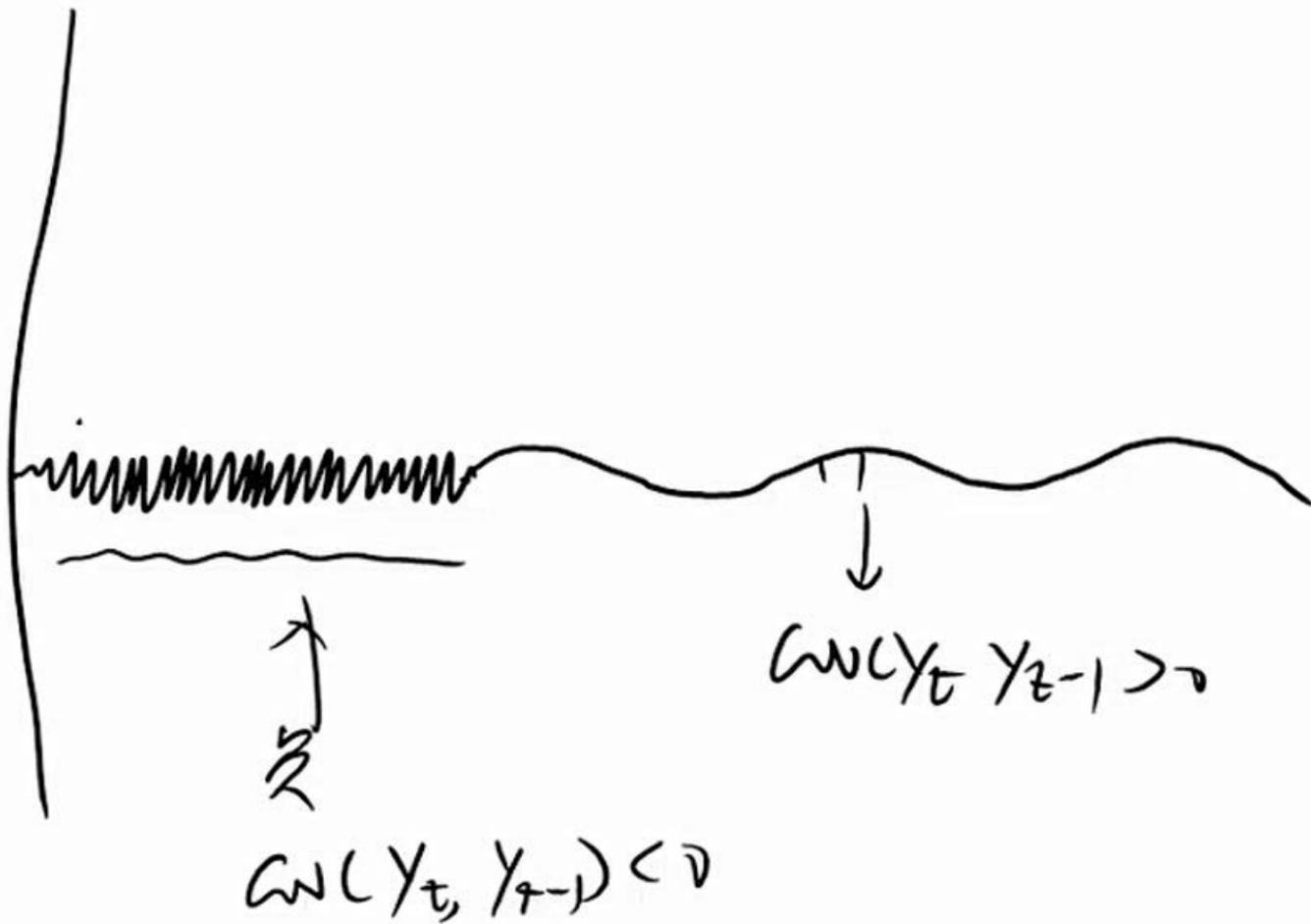
- The expected value of the time series must be constant and finite in all periods.

$$E(y_t) = \mu \text{ and } |\mu| < \infty, t = 1, 2, \dots, T$$

- The variance of the time series must be constant and finite in all periods.
- The covariance of the time series with itself for a fixed number of periods in the past or future must be constant and finite in all periods.

$\text{Cov}(y_t, y_{t-s}) = \lambda, |\lambda| < \infty, t = 1, 2, \dots, T; s = 0, \pm 1, \pm 2, \dots, \pm T$

- Stationary in the past does not guarantee stationary in the future.
- All covariance-stationary time series have a finite **mean-reverting level**.



# ◆ Autoregressive Models (AR)

## ➤ Mean reversion

- A time series shows mean reversion if it tends to fall when its level is above its mean and rise when its level is below its mean.
- For an AR(1) model, the mean-reverting level,  $x_t$ , is given by:  $x_i = \frac{b_0}{1-b_1}$ 
  - ✓ The time series will increase if  $x_i < \frac{b_0}{1-b_1}$
  - ✓ The time series will decrease if  $x_i > \frac{b_0}{1-b_1}$

$$Y_t = b_0 + b_1 Y_{t-1} + \varepsilon$$

$$\mu = b_0 + b_1 \mu + \sigma$$

$$\underbrace{EY_t}_{\mu} = b_0 + b_1 \underbrace{EY_{t-1}}_{\mu} + \cancel{\varepsilon} = \mu$$

$$\mu = b_0 + b_1 \cdot \mu$$

$$(1 - b_1) \mu = b_0 \quad \mu = \frac{b_0}{1 - b_1}$$

3

$$x_t = 2$$

$$\underline{x_{t+1}} \uparrow$$

$$x_{t+1} = 4 \quad x_{t+2} \downarrow$$

# Autoregressive Models (AR)

## ➤ Mean reversion

- A time series shows mean reversion if it tends to fall when its level is above its mean and rise when its level is below its mean.



- For an AR(1) model, the mean-reverting level,  $x_t$ , is given by:

$$x_t = \frac{b_0}{1-b_1}$$

✓ The time series will increase if  $x_i < \frac{b_0}{1-b_1}$

= 3

✓ The time series will decrease if  $x_i > \frac{b_0}{1-b_1}$

# ◆ Autoregressive Models (AR)

## ✓ Random walk

- $x_t = x_{t-1} + \varepsilon_t$  ( $b_0 = 0$  and  $b_1 = 1$ )
- The best forecast of  $x_t$  that can be made in period  $t-1$  is  $x_{t-1}$ .

## ➤ Random walk with a drift

- $x_t = b_0 + x_{t-1} + \varepsilon_t$  ( $b_0 \neq 0$ ,  $b_1 = 1$ )
- $X_t$  should increase or decrease by a constant amount in each period.

## ➤ Features

- A random walk is not covariance stationary
  - ✓ It has an undefined mean reverting level
  - ✓ It has an infinite variance

random walk  $\rightarrow b_1 = 1 \rightarrow$  mean reverting level / ~~not~~  $\rightarrow$  Not Covariance stationary

# ◆ Autoregressive Models (AR)

## ➤ The unit root test of nonstationarity

- The time series is said to have a unit root if the lag coefficient is equal to one.
- A common t-test of the hypothesis that  $b_1 = 1$  is invalid to test the unit root, however, it is not often the case.

## ➤ Dickey-Fuller test (DF test) to test the unit root

- Start with  $x_t - x_{t-1} = b_0 + (b_1 - 1)x_{t-1} + \varepsilon_t$ , or  $x_t - x_{t-1} = b_0 + g x_{t-1} + \varepsilon_t$
- Make hypothesis
  - $H_0: g=0$  (has a unit root and is nonstationary)
  - $H_a: g < 0$  (does not have a unit root and is stationary)
- Calculate conventional t-statistic and use revised t-table, which is computed by Dickey and Fuller.
- If we can reject the null, the time series **does not have a unit root and is stationary**.

$$b_1 = 1$$

~~$$t = \frac{b_1 - 1}{\text{SE}(b_1)}$$~~

➤ Dickey-Fuller test (DF test) to test the unit root

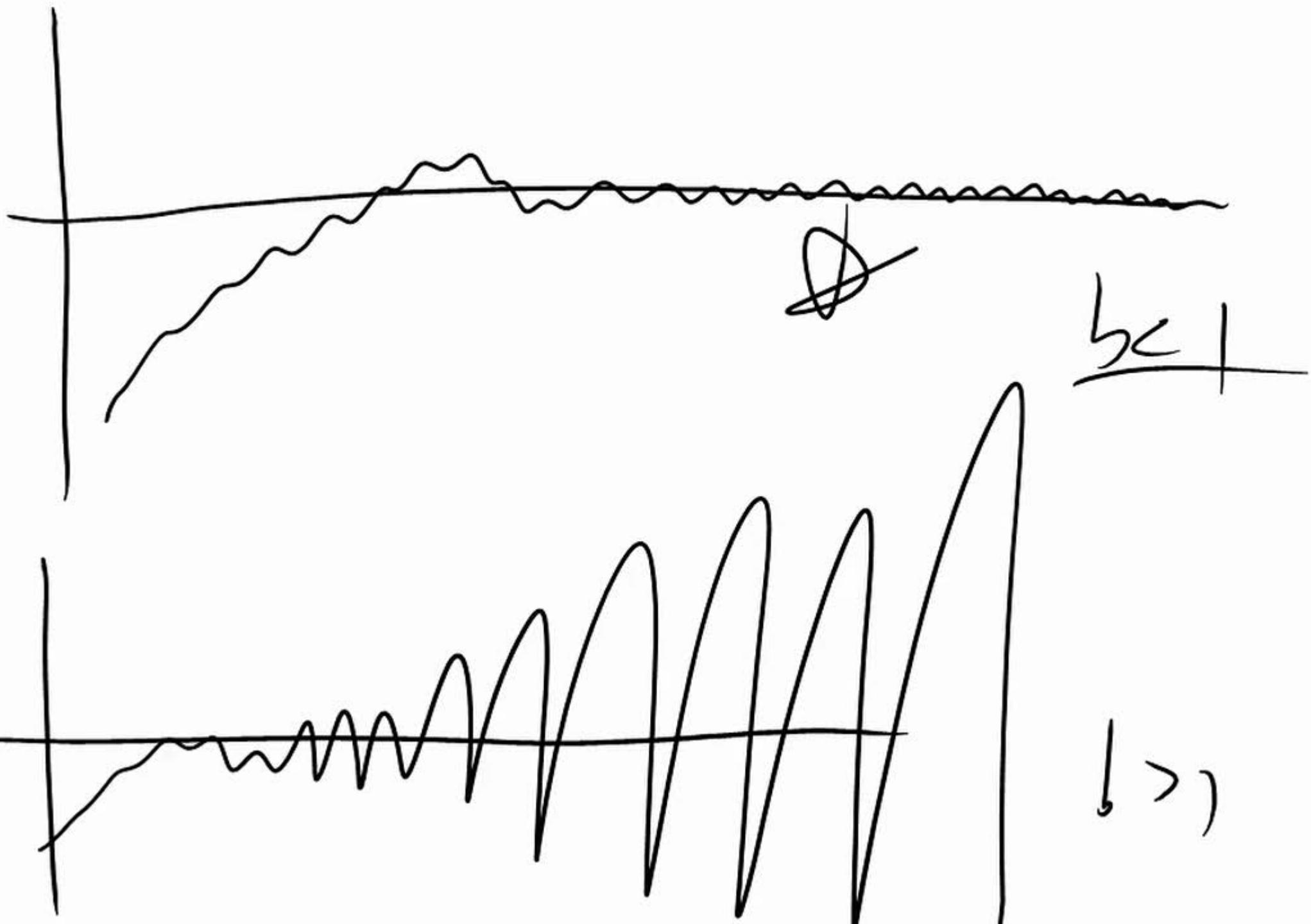
- Start with  $x_t - \underline{x_{t-1}} = b_0 + (b_1 - 1)x_{t-1} + \varepsilon_t$ , or  $\underline{x_t - x_{t-1}} = b_0 + g x_{t-1} + \varepsilon_t$
- Make hypothesis
  - ✓  $H_0: g=0$  (has a unit root and is nonstationary)
  - \* ✓  $H_a: \underline{g < 0}$  (does not have a unit root and is stationary)
- Calculate conventional t-statistic and use revised t-table, which is computed by Dickey and Fuller.
- If we can reject the null, the time series **does not have a unit root and is stationary**.

$s < 0$

$\Rightarrow b_1 - 1 < 0$

$\Rightarrow b_1 < 1 \rightarrow$  负数

$b_1 > 1 \rightarrow$  正数



PF  $\rightarrow$  can't reject  $H_0$

~~hazard~~

level

$b_1$ ,  $\text{retr} \bar{y}_0 = 1$



mean reverting level  $\bar{y}_{\text{PDT}}$

# Autoregressive Models (AR)

- If a time series appears to have a unit root  $\rightarrow b_1 = 1$ 
  - One method that is often successful is to first-difference the time series (as discussed previously) and try to model the first-differenced series as an autoregressive time series.
- First differencing  $- \text{原差分} \rightarrow \text{一阶差分}$ 
  - Define  $y_t$  as  $y_t = x_t - x_{t-1} = \varepsilon_t$
  - The first-differenced variable  $y_t$  is covariance stationary



# Autoregressive Models (AR)



## ➤ Autocorrelation of residuals in an AR model

- When the error terms are correlated, standard errors are unreliable.
- Durbin-Watson statistic is invalid when the independent variables include past values of the dependent variable
- Using **t-test** to test the significance of residual autocorrelation

## ➤ Detecting autocorrelation in an AR model

- Compute the autocorrelations of the residual
- t-tests to see whether the residual autocorrelations differ significantly from 0,

$$t - \text{statistics} = \frac{r_{\varepsilon_t, \varepsilon_{t-k}} - 0}{S_r} = \frac{r_{\varepsilon_t, \varepsilon_{t-k}}}{1/\sqrt{n}}$$

- If the residual autocorrelations differ significantly from 0, the model is not correctly specified, so we may need to modify it.

## Correlation Analysis

$$\underline{r(x,y)}$$

$$t = \frac{r - D}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

## Autocorrelation in AR

$$r(\varepsilon_i, \varepsilon_{i-1})$$

$$r(\varepsilon_i, \varepsilon_{i-2})$$

$$t = \frac{r - D}{\sqrt{\frac{1 - r^2}{n}}}$$

# Autoregressive Models (AR)

## ➤ Seasonality – a special question

- Time series shows regular patterns of movement within the year
- The seasonal autocorrelation of the residual will differ significantly from 0
- We should add a seasonal lag in an AR model
  - ✓ For example:  $x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-4} + \varepsilon_t$ , AR(1) model with a seasonal lag

$$\underline{Y_t = b_0 + b_1 Y_{t-1} + \varepsilon_t}$$

季节

$$Y_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-4} + \varepsilon_t$$

## Correlation Analysis

$$\underline{r(x,y)}$$

$$t = \frac{r - D}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Run a simple  
Regression

## Autocorrelation in AR

$$r(\varepsilon_i, \varepsilon_{i-1})$$

$$r(\varepsilon_i, \varepsilon_{i-2})$$

$$t = \frac{r_D}{\sqrt{\frac{1}{n}}}$$

Add back ~~sign~~  
sig. lags

# Example

- Suppose we decide to use an autoregressive model with a seasonal lag because of the seasonal autocorrelation in the previous problem. We are modeling quarterly data, so we estimate Equation:
  - $(\ln \text{Sales}_t - \ln \text{Sales}_{t-1}) = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + b_2(\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}) + \varepsilon_t.$
- **Q1:** Using the information in Table 1, determine if the model is correctly specified.
- **Q2:** If sales grew by 1 percent last quarter and by 2 percent four quarters ago, use the model to predict the sales growth for this quarter.

- Suppose we decide to use an autoregressive model with a seasonal lag because of the seasonal autocorrelation in the previous problem. We are modeling quarterly data, so we estimate Equation:
  - $(\ln \text{Sales}_t - \ln \text{Sales}_{t-1}) = b_0 + b_1(\ln \text{Sales}_{t-1} - \ln \text{Sales}_{t-2}) + b_2(\ln \text{Sales}_{t-4} - \ln \text{Sales}_{t-5}) + \varepsilon_t$
- **Q1:** Using the information in Table 1, determine if the model is correctly specified.
- **Q2:** If sales grew by 1 percent last quarter and by 2 percent four quarters ago, use the model to predict the sales growth for this quarter.

$$\ln S_t - \ln S_{t-1} = \ln \frac{S_t}{S_{t-1}} = \ln(1+r) = r^{\frac{1}{4}}$$

# Table 1. Log Differenced Sales

➤ Table 1

Regression Statistics	
R-squared	0.4220
Standard error	0.0318
Observations	68
Durbin-Watson	1.8784

	Coefficient	Standard Error	t-Statistic
Intercept	0.0121	0.0053	2.3055
Lag 1	-0.0839	0.0958	-0.8757
Lag 4	0.6292	0.0958	6.5693

## Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	t-Statistic
1	0.0572	0.1213	0.4720
2	-0.0700	0.1213	-0.5771
3	0.0065	0.1213	-0.0532
4	-0.0368	0.1213	-0.3033

$\delta(\varepsilon_t, \varepsilon_{t-4}) \neq 0$

$\downarrow$   
 $y_t + y_{t-4}$   $\neq$

# ◆ Table 1. Log Differenced Sales

➤ Table 1

Regression Statistics	
R-squared	0.4220
Standard error	0.0318
Observations	68
Durbin-Watson	1.8784

$$\sqrt{1/n}$$

$$\sqrt{\frac{1}{68}} = 0.11213$$

	Coefficient	Standard Error	t-Statistic
Intercept	0.0121	0.0053	2.3055
Lag 1	-0.0839	0.0958	-0.8757
Lag 4	0.6292	0.0958	6.5693

## Autocorrelations of the Residual

Lag	Autocorrelation	Standard Error	t-Statistic
1	0.0572 ✓	0.1213	0.4720
2	-0.0700	0.1213	-0.5771
3	0.0065	0.1213	-0.0532
4	-0.0368	0.1213	-0.3033 3.033

➤ **Answer to Q1**

- At the 0.05 significance level, with 68 observations and three parameters, this model has 65 degrees of freedom. **The critical value** of the t-statistic needed to reject the null hypothesis is thus about 2.0.
- The absolute value of the t-statistic for each autocorrelation is all below 2.0, so we cannot reject the null hypothesis that each autocorrelation is not significantly different from 0. **We have determined that the model is correctly specified.**

## ➤ Answer to Q2

- If sales grew by 1 percent last quarter and by 2 percent four quarters ago, then the model predicts that sales growth this quarter will be  $e^{0.0121 - 0.0839 \ln(1.01) + 0.6292 \ln(1.02)} - 1 = 2.40\%$ .



# Autoregressive Models (AR)

- **Heteroskedasticity** refers to the situation that the variance of the error term is not constant and dependence of the error term variance on the independent variable.
- **Test whether a time series is ARCH(1)**
  - $\varepsilon_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + u_t$
  - If the estimate of  $a_1$  is statistically significantly different from zero, we conclude that the time series is ARCH(1).
    - ✓ If a time-series model has ARCH(1) errors, then the variance of the errors in period  $t + 1$  can be predicted in period  $t$ .
- If ARCH exists,
  - the standard errors for the regression parameters will not be correct.  
Generalized least squares must be used to develop a predictive model.
  - ARCH model can be used to predict the variance of the residuals in future periods.

Regression

CH

BP test

$$BP = n \times R^2 (\text{residual})$$

$$\Sigma^2 = C_0 + C_1 X + \xi$$

AR

ARCH

t-test

$$t = \frac{\hat{b}_1 - \sigma}{SE(\hat{b}_1)}$$

$$\varepsilon_t^2 = b_0 + b_1 \tilde{\varepsilon}_{t-1}^2 + \xi$$

# Regression

CH

BP test

$$BP = n \times R^2_{\text{residual}}$$

$$\sum^2 = C_0 + C_1 X + \xi$$

Reject  $H_0 \rightarrow \underline{R^2_{\text{residual}}} \uparrow$

$\rightarrow$   $X$  与  $y$  之间有关系

$\rightarrow$  CH Hypo

AR

ARCH

t-test

$$t = \frac{\hat{b}_1 - D}{SE(\hat{b}_1)}$$

$$\xi_t^2 = b_0 + b_1 \xi_{t-1}^2 + \xi$$

Reject  $H_0 \rightarrow b_1 \neq D$

$\rightarrow \xi_t \pm \xi_{t-1}$  有关联

$\rightarrow$  ARCH? 假设

# ◆ Autoregressive Models (AR)

- **Heteroskedasticity** refers to the situation that the variance of the error term is not constant and dependence of the error term variance on the independent variable. *conditional*
- **Test whether a time series is ARCH(1)**
  - $\varepsilon_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + u_t$
  - If the estimate of  $a_1$  is statistically significantly different from zero, we conclude that the time series is ARCH(1).
    - ✓ If a time-series model has ARCH(1) errors, then the variance of the errors in period  $t + 1$  can be predicted in period  $t$ .
- If ARCH exists,
  - the standard errors for the regression parameters will not be correct.  
Generalized least squares must be used to develop a predictive model.
  - ARCH model can be used to predict the variance of the residuals in future periods.



# Autoregressive Models (AR)

## ➤ Multiperiod forecasts

- Chain rule of forecasting

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

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

- ✓ If we want to forecast  $x_{t+2}$  using an AR(1) model, our forecast will be based on

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

2013

$$Y_t = 2 + 3 Y_{t-1}$$

$$\underline{Y_3 = 5} \rightarrow Y_4 = 2 + 3 \times 5 = 17$$

$$\text{or } Y_6 = ?$$



$$Y_5 = 2 + 3 \times 17 = 53$$

$$2 + 3 \times 53 = 161$$

$$Y_t - Y_{t-1} = 2 + 3(Y_{t-1} - Y_{t-2})$$

$$Y_3 = 3 \quad Y_4 = 4 \quad \text{so } Y_5 = ?$$

$$\checkmark Y_5 - Y_4 = 2 + 3(Y_4 - Y_3)$$

$$\checkmark Y_6 - Y_5 = 2 + 3(Y_5 - Y_4)$$

$$\ln Y_t - \ln Y_{t-1} = 2+3 (\ln Y_{t-1} - \ln Y_{t-2})$$

$$Y_3 = 3 \quad Y_4 = 4 \quad Y_6 = 2$$

~~$$\ln Y_5 - \ln Y_4$$~~
$$\checkmark \qquad \checkmark$$
$$\ln \underline{Y_5} - \ln Y_4 = 2+3 (\ln Y_4 - \ln Y_3)$$

# ◆ Autoregressive Models (AR)

## ➤ Comparing forecasting model performance

- In-sample forecast errors are the residuals from a fitted time-series model.
- Out-of-sample forecast errors are the differences between actual and predicted inflation, if we use this model to make a prediction outside this period.

- ✓ Root mean squared error (RMSE) the model with the smallest RMSE is most accurate for out-of-sample
- ✓ RMSE is the square root of the average squared error

$$\sqrt{\frac{\sum_i e_i^2}{df}} \rightarrow \text{SEE}$$

$$R^2$$

$$\checkmark \frac{RSS}{}$$

$$\checkmark \frac{SSE}{}$$

~~$$\checkmark \frac{SEE}{}$$~~

$$\checkmark \frac{SST}{}$$

$$\left\{ \begin{array}{l} MSE \\ MSR \end{array} \right.$$

$$RMSE$$

$$RSS + SSE = SST$$

~~$$\sqrt{\frac{SSE}{df}} = SEE$$~~

~~$$MSE = \frac{SSE}{n-k-1}$$~~

~~$$MSR = \frac{RSS}{k}$$~~

$$RMSE = \sqrt{\frac{\sum \varepsilon_i^2}{df}} \leftarrow \textcircled{SSE}$$

# Autoregressive Models (AR)

## ➤ Instability of regression coefficients

- The regression coefficient estimates of a time-series model estimated using an earlier sample period can be quite different from those of a model estimated using a later sample period.
  - ✓ e.g. Exchange rate with fixed regime and floating regime
- The estimates can be different between models estimated using relatively **shorter** and **longer** sample periods.
  - ✓ Unfortunately, there is usually no clear-cut basis in economic or financial theory for determining whether to use data from a longer or shorter sample period to estimate a time-series model.

# ◆ Regression with More Than One Time Series

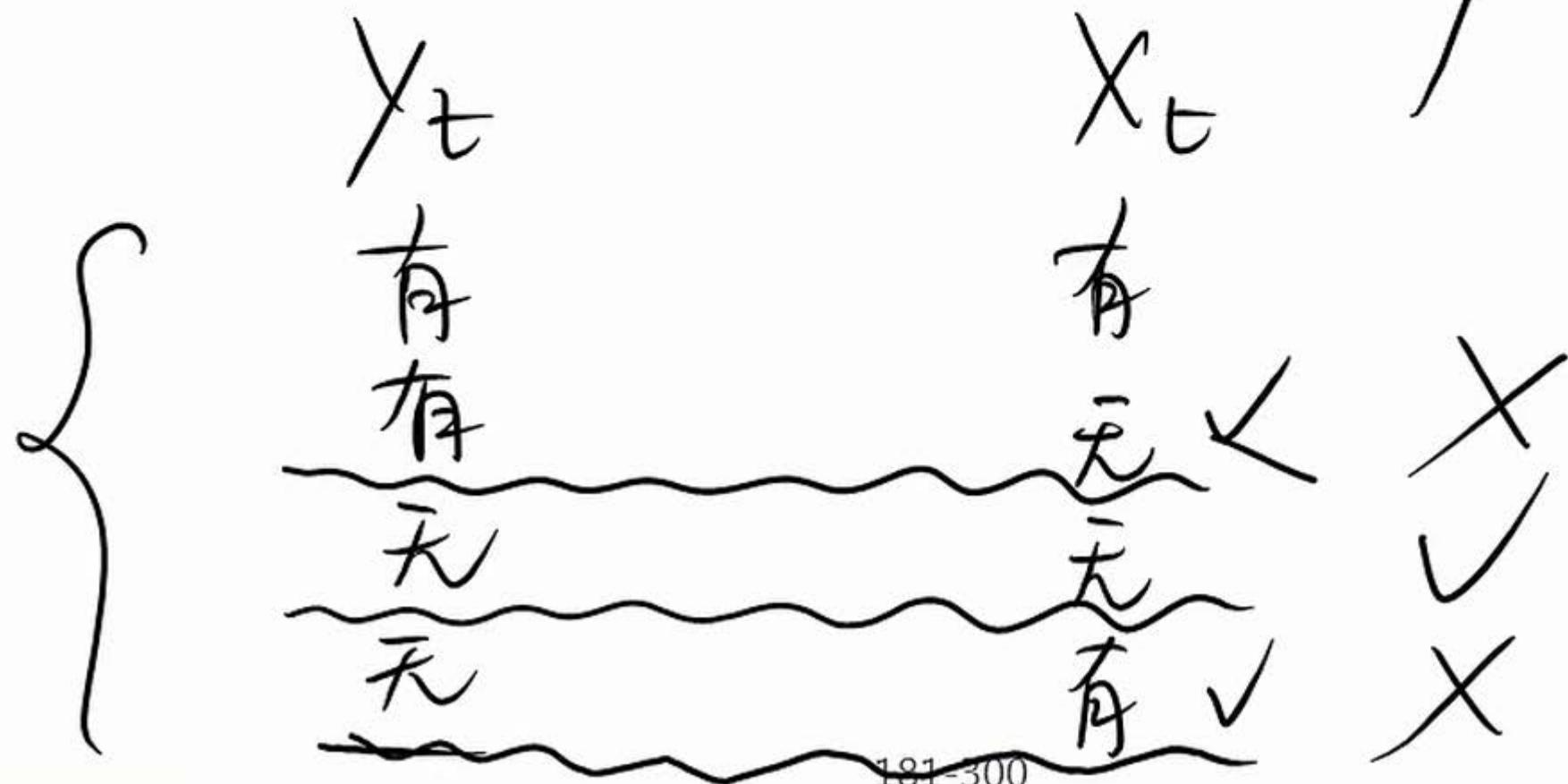
- In linear regression, if any time series contains a unit root, OLS may be invalid
- Use DF tests for each of the time series to detect unit root, we will have 3 possible scenarios
  - None of the time series has a unit root: we can use multiple regression
  - At least one time series has a unit root while at least one time series does not: we cannot use multiple regression
  - Each time series has a unit root: we need to establish whether the time series are cointegrated.
    - ✓ If conintegrated, can estimate the long-term relation between the two series (but may not be the best model of the short-term relationship between the two series).

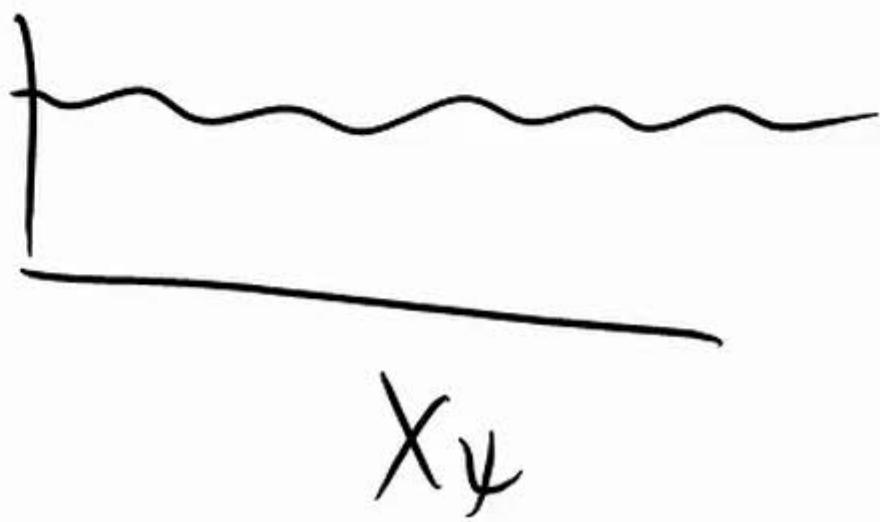
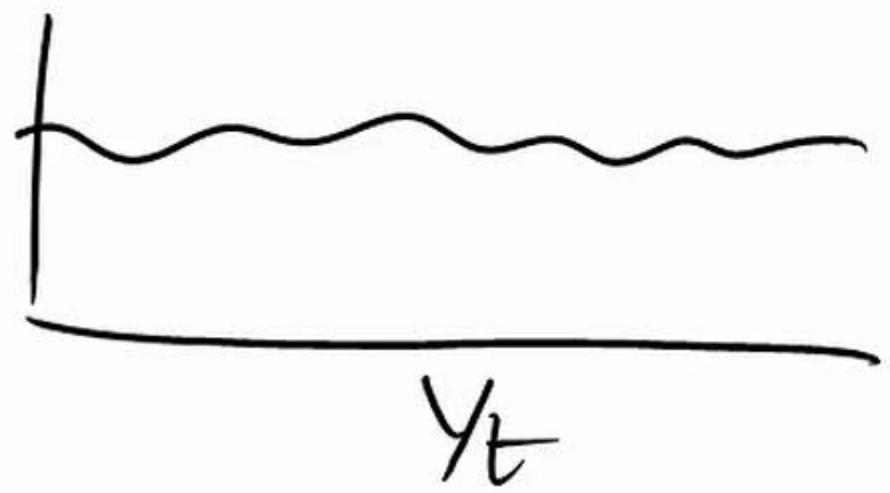
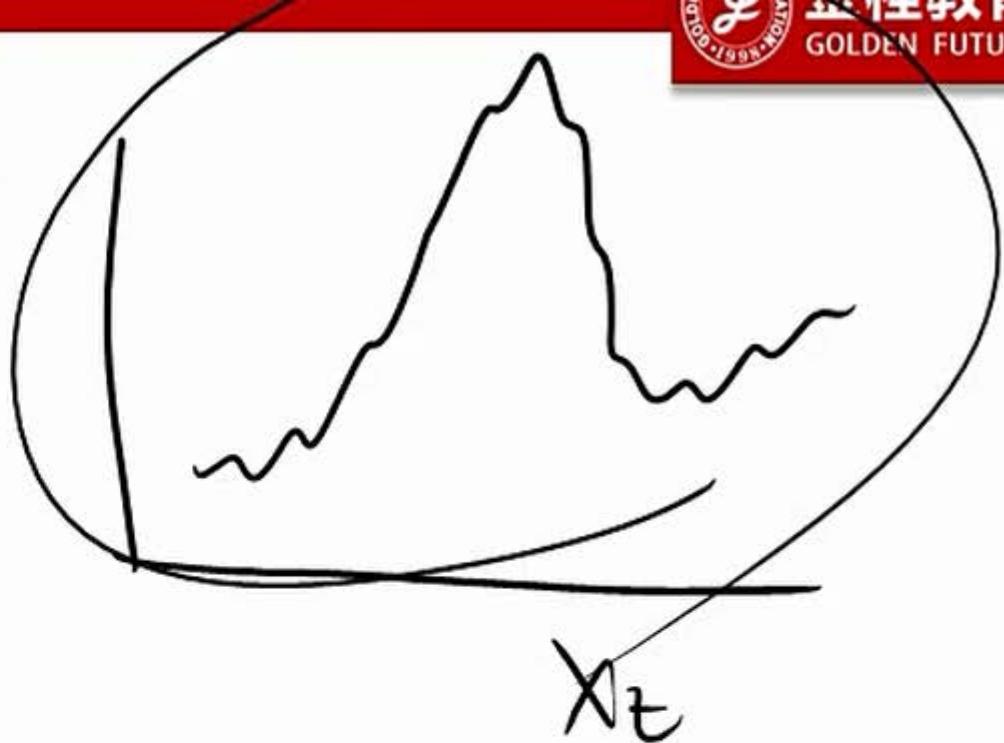
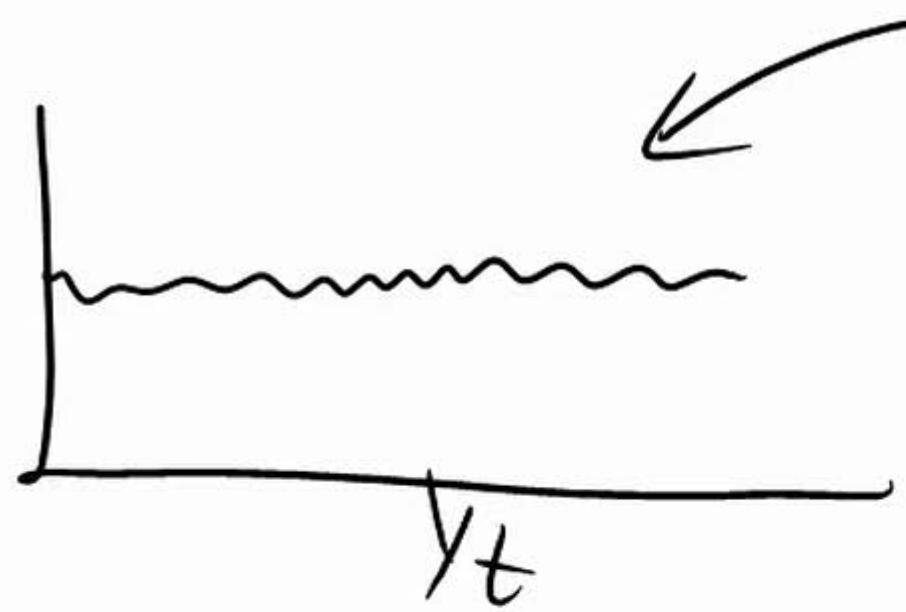
$$\checkmark_t = b_0 + b_1 \checkmark_{t-1} + b_2 x_t + \zeta$$

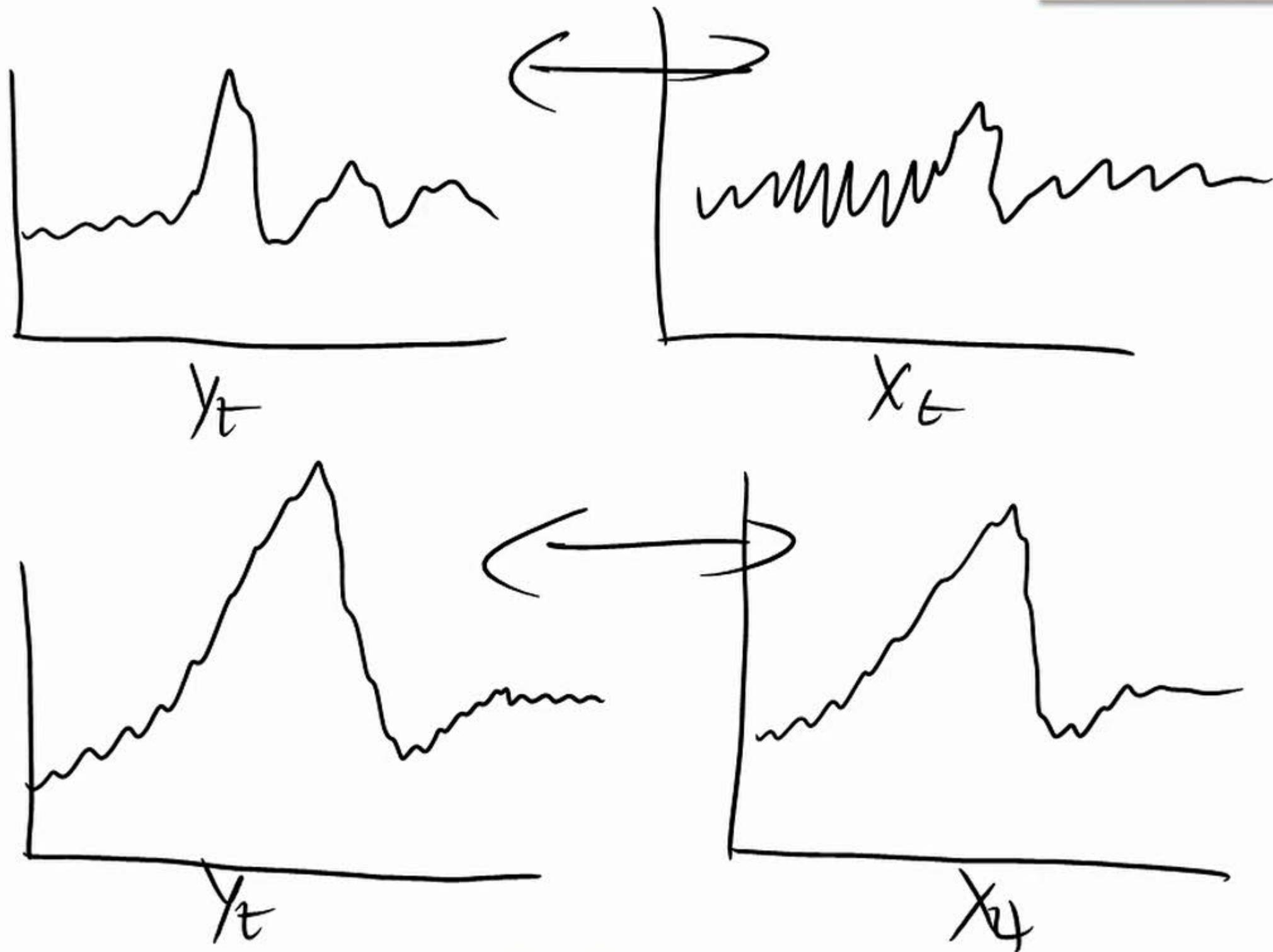
消  
吸收

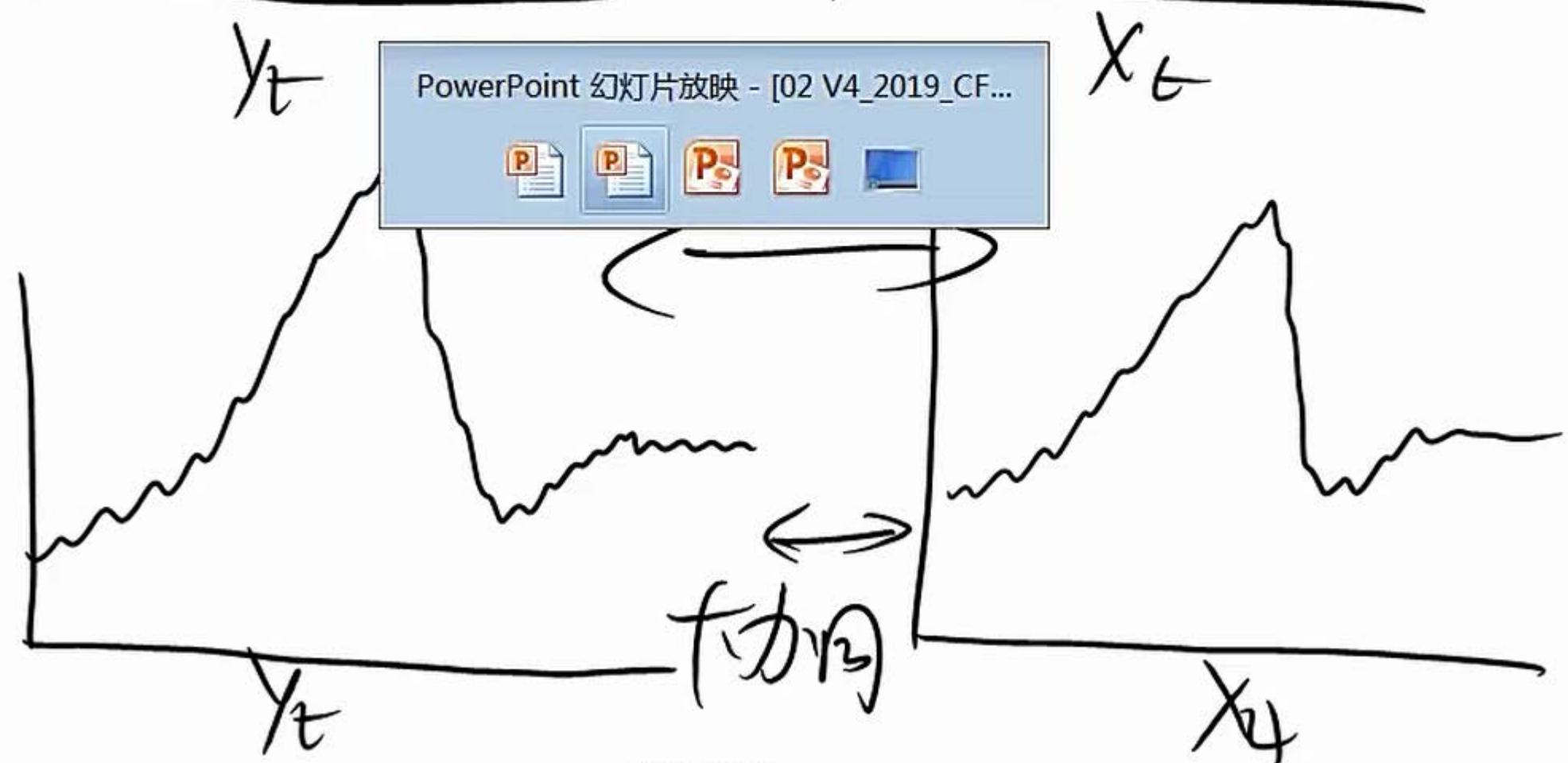
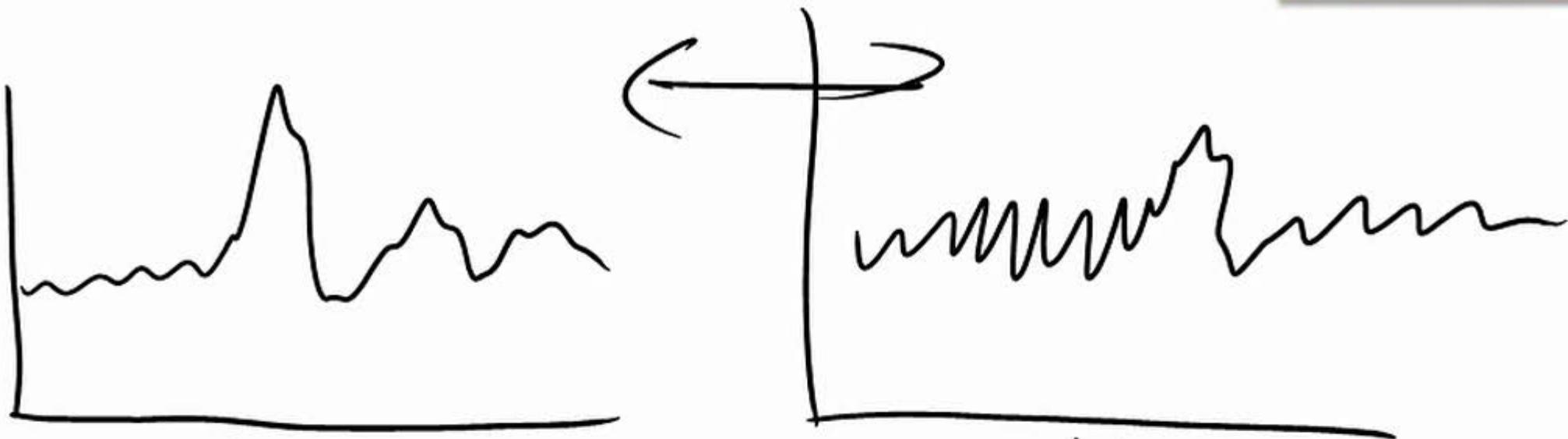
$\checkmark_t$

$x_t$









## Regression with More Than One Time Series

- In linear regression, if any time series contains a unit root, OLS may be invalid
- Use DF tests for each of the time series to detect unit root, we will have 3 possible scenarios
  - None of the time series has a unit root: we can use multiple regression
  - At least one time series has a unit root while at least one time series does not: we cannot use multiple regression
  - Each time series has a unit root: we need to establish whether the time series are cointegrated.
    - ✓ If cointegrated, can estimate the long-term relation between the two series (but may not be the best model of the short-term relationship between the two series).

# ◆ Regression with More Than One Time Series

- Use the Dickey-Fuller Engle-Granger test (DF-EG test) to test the cointegration

- $H_0$ : no cointegration       $H_a$ : cointegration
- If we cannot reject the null, we cannot use multiple regression
- If we can reject the null, we can use multiple regression
- Critical value calculated by Engle and Granger

↑  
禁用多变量

↑  
禁用多变量

e calculated by Engle and Granger

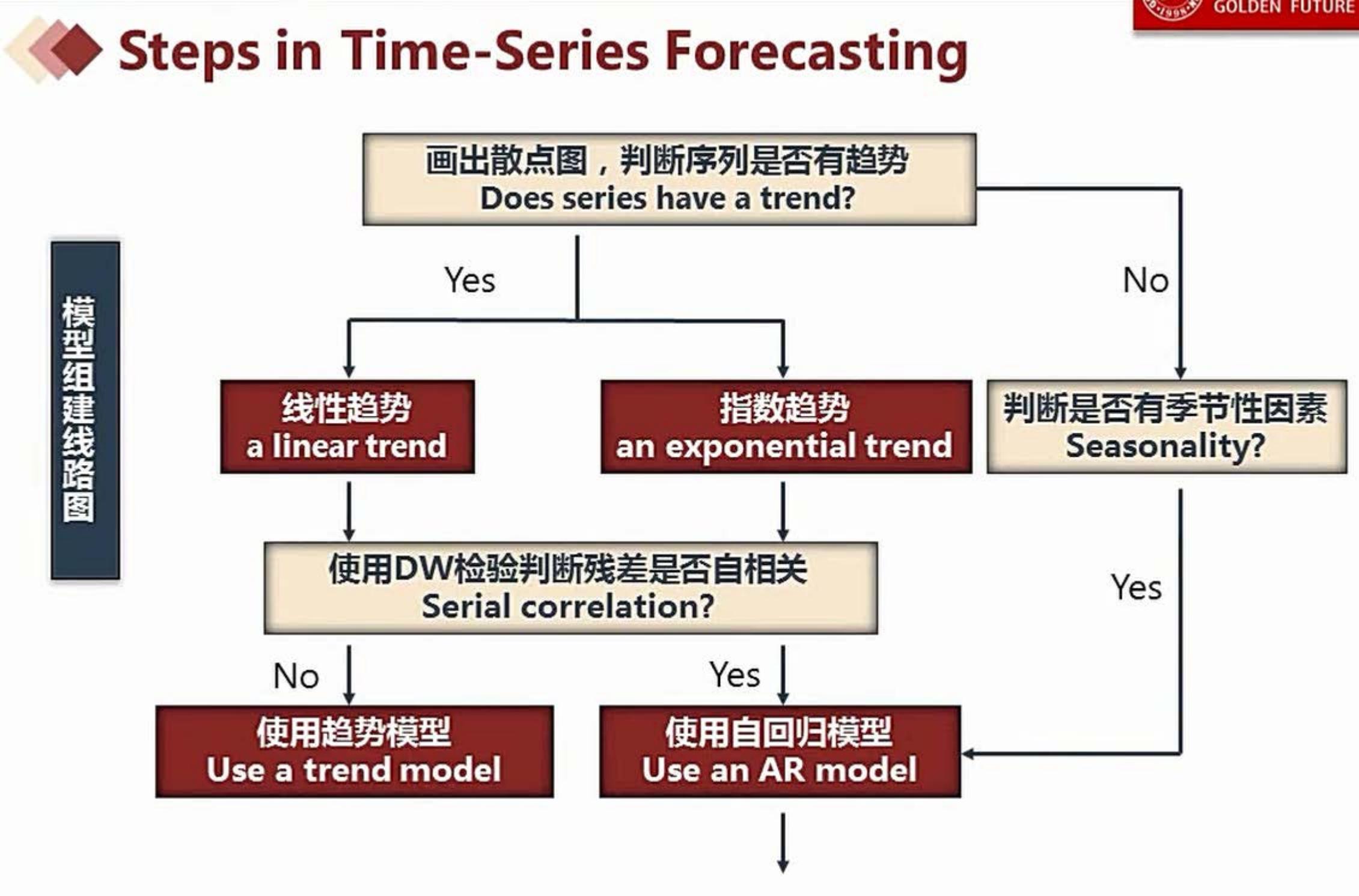
$X_t$  与  $Y_t$  之相关，下图  $Y_t$

$\left. \begin{array}{c} DP \\ DP \end{array} \right\}$   
DEG

calculated by Engle and Granger

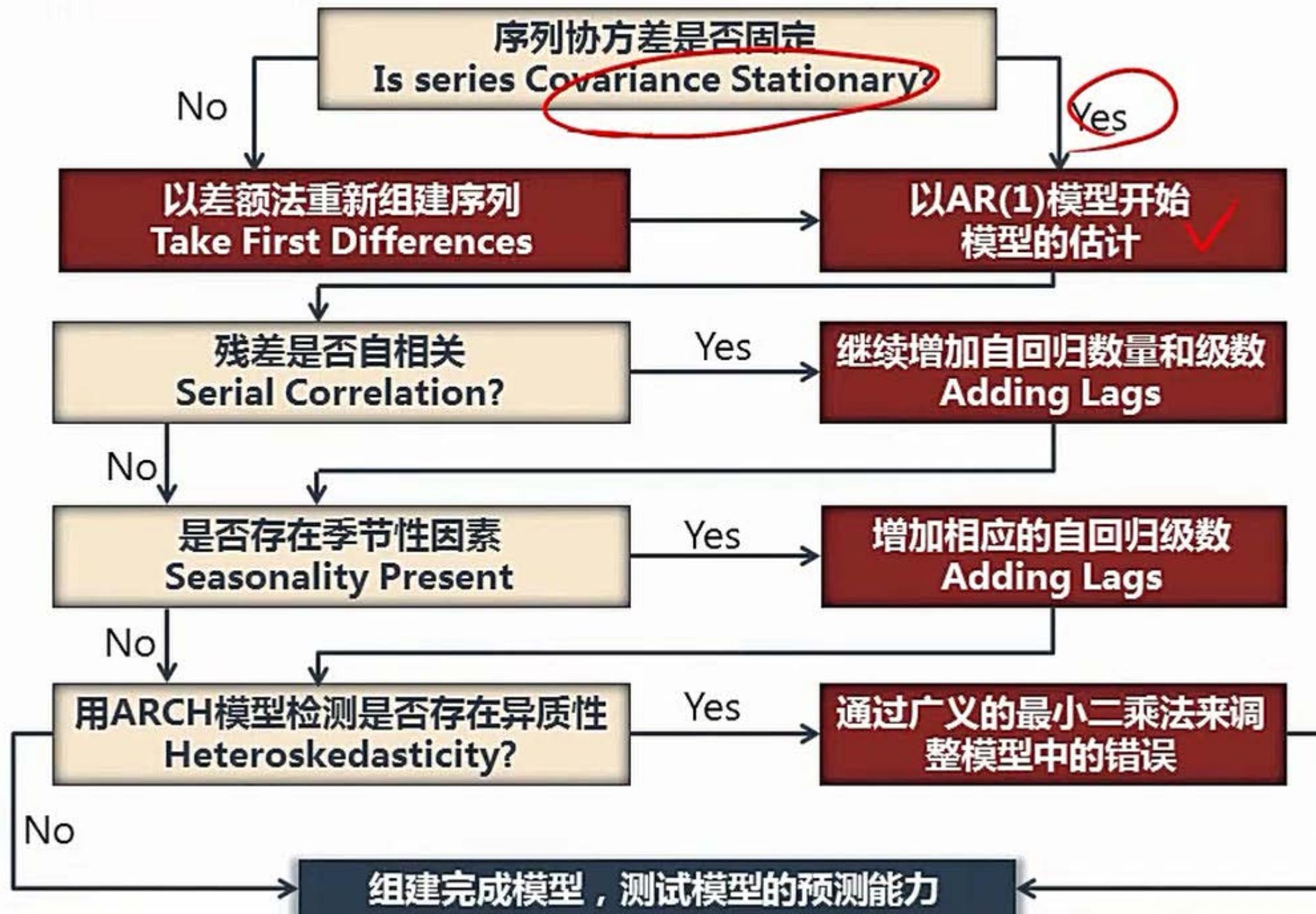
$X_t$  布单径报，下  $y_t$  {  
DP  
DPEG ✓}

$X_t$  元单径报。 -  $y_t$  {  
PF ✓  
NPZG}



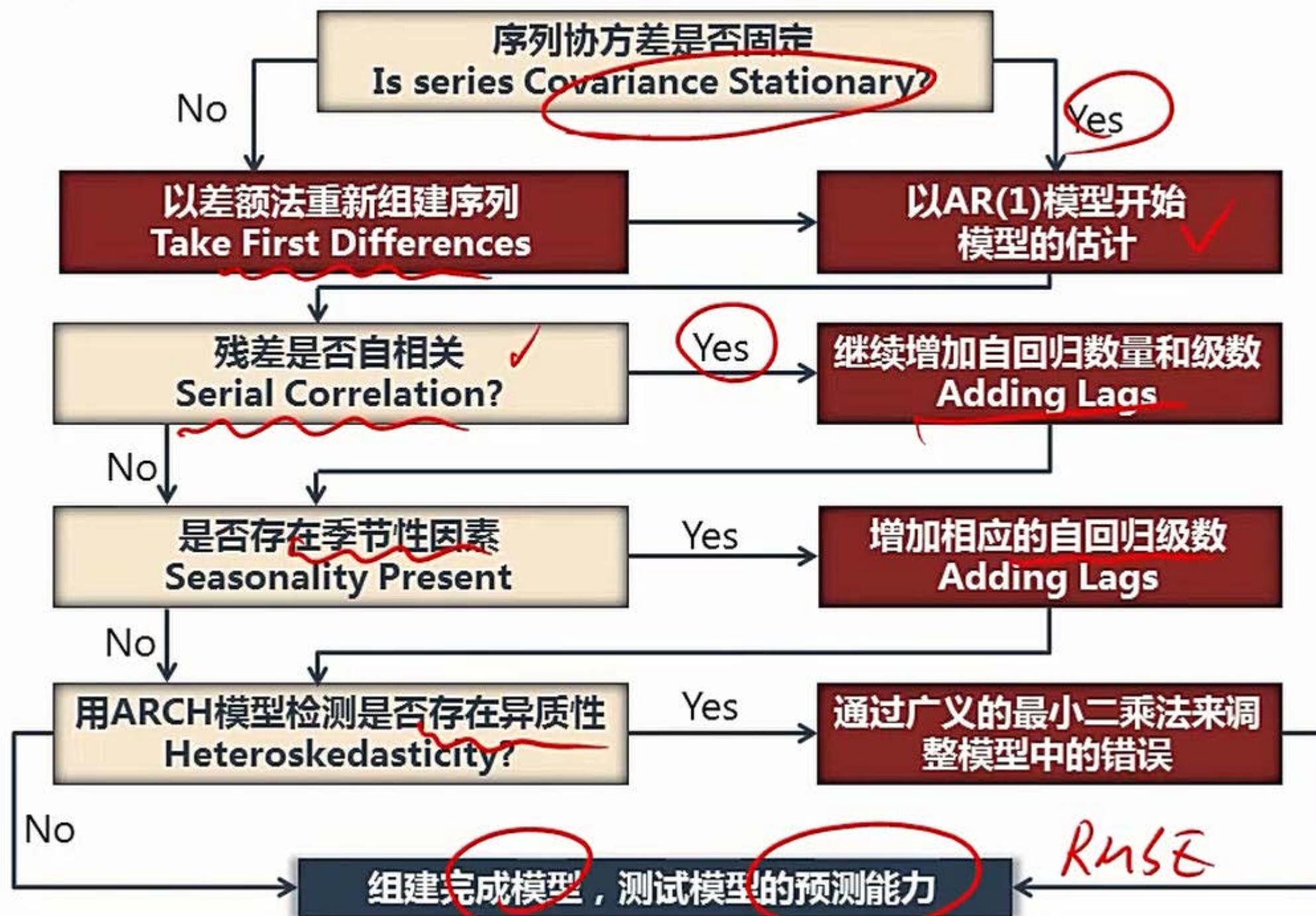
# ◆ Steps in Time-Series Forecasting

自回归模型组建路线图



# ◆ Steps in Time-Series Forecasting

自回归模型组建路线图





## Example



- Angela Martinez, an energy sector analyst at an investment bank, is considering whether to recommend a hedge for the bank portfolio's exposure to changes in oil prices. Martinez examines West Texas Intermediate (WTI) monthly crude oil price data, expressed in US dollars per barrel, for the 181-month period from August 2000 through August 2015. The end-of-month WTI oil price was \$51.16 in July 2015 and \$42.86 in August 2015 (Month 181).
- After reviewing the time-series data, Martinez determines that the mean and variance of the time series of oil prices are not constant over time. She then runs the following four regressions using the WTI time-series data.
  - Linear trend model: Oil price<sub>t</sub> = b<sub>0</sub> + b<sub>1</sub>t + e<sub>t</sub>
  - Log-linear trend model: ln Oil price<sub>t</sub> = b<sub>0</sub> + b<sub>1</sub>t + e<sub>t</sub>
  - AR(1) model: Oil price<sub>t</sub> = b<sub>0</sub> + b<sub>1</sub>Oil price<sub>t-1</sub> + e<sub>t</sub>
  - AR(2) model: Oil price<sub>t</sub> = b<sub>0</sub> + b<sub>1</sub>Oil price<sub>t-1</sub> + b<sub>2</sub>Oil price<sub>t-2</sub> + e<sub>t</sub>

# Exhibit 1

## Regression Statistics (t-statistics for coefficients are reported in parentheses)

	Linear	Log-Linear	AR(1)	AR(2)
R <sup>2</sup>	0.5703	0.6255	0.9583	0.9656
Standard error	18.6327	0.3034	5.7977	5.2799
Observations	181	181	180	179
Durbin-Watson	0.10	0.08	1.16	2.08
RMSE			2.0787	2.0530

## Coefficients:

Intercept	28.3278 (10.1846)	3.3929 (74.9091)	1.5948 (1.4610)	2.0017 (1.9957)
t (Trend)	0.4086 (15.4148)	0.0075 (17.2898)		
Oil Price <sub>t-1</sub>			0.9767 (63.9535)	1.3946 (20.2999)
Oil Price <sub>t-2</sub>				-0.4249 (6.2064)



## Example



In Exhibit 1, at the 5% significance level, the lower critical value for the Durbin-Watson test statistic is 1.75 for both the linear and log-linear regressions. Exhibit 2 presents selected autocorrelation data from the AR(1) models.

### Exhibit 2. Autocorrelations of the Residual from AR(1) Model

Lag	Autocorrelation	t-Statistic
1	0.4157	5.5768
2	0.2388	3.2045
3	0.0336	0.4512
4	-0.0426	-0.5712

Note: At the 5% significance level, the critical value for at-statistic is 1.97.

After reviewing the data and regression results, Martinez draws the following conclusions.

- Conclusion 1: The time series for WTI oil prices is covariance stationary.
- Conclusion 2: Out-of-sample forecasting using the AR(1) model appears to be more accurate than that of the AR(2) model.

1. Based on Exhibit 1, the predicted WTI oil price for October 2015 using the linear trend model is *closest* to:
  - A. \$29.15.
  - B. \$74.77.
  - C. \$103.10.
2. Based on Exhibit 1, the predicted WTI oil price for September 2015 using the log-linear trend model is *closest* to:
  - A. \$29.75.
  - B. \$29.98.
  - C. \$116.50.
3. Based on the regression output in Exhibit 1, there is evidence of positive serial correlation in the errors in:
  - A. the linear trend model but not the log-linear trend model.
  - B. both the linear trend model and the log-linear trend model.
  - C. neither the linear trend model nor the log-linear trend model.

# Exhibit 1

**Regression Statistics** (t-statistics for coefficients are reported in parentheses)

	Linear	Log-Linear	AR(1)	AR(2)
R <sup>2</sup>	0.5703	0.6255	0.9583	0.9656
Standard error	18.6327	0.3034	5.7977	5.2799
Observations	181	181	180	179
Durbin-Watson	0.10	0.08	1.16	2.08
RMSE			2.0787	2.0530
<b>Coefficients:</b>				
Intercept	28.3278 (10.1846)	3.3929 (74.9091)	1.5948 (1.4610)	2.0017 (1.9957)
t (Trend)	0.4086 (15.4148)	0.0075 (17.2898)		
Oil Price <sub>t-1</sub>			0.9767 (63.9535)	1.3946 (20.2999)
Oil Price <sub>t-2</sub>	<i>t-statistics</i>			
			-0.4249 (6.2064)	

## Example

1. Based on Exhibit 1, the predicted WTI oil price for October 2015 using the linear trend model is closest to:
- A. \$29.15.
  - B. \$74.77.
  - C. \$103.10.
- $t = 183$
2. Based on Exhibit 1, the predicted WTI oil price for September 2015 using the log- linear trend model is closest to:
- A. \$29.75.
  - B. \$29.98.
  - C. \$116.50.
- $t = 182$
3. Based on the regression output in Exhibit 1, there is evidence of positive serial correlation in the errors in:
- A. the linear trend model but not the log-linear trend model.
  - B. both the linear trend model and the log-linear trend model.
  - C. neither the linear trend model nor the log-linear trend model.

**Regression Statistics** (t-statistics for coefficients are reported in parentheses)

	<b>Linear</b>	<b>Log-Linear</b>	<b>AR(1)</b>	<b>AR(2)</b>
R <sup>2</sup>	0.5703	0.6255	0.9583	0.9656
Standard error	18.6327	0.3034	5.7977	5.2799
Observations	181	181	180	179
Durbin-Watson	<u>0.10</u>	<u>0.08</u>	1.16	2.08
RMSE			2.0787	2.0530

## Example

$d_L \quad d_u$

In Exhibit 1, at the 5% significance level, the lower critical value for the Durbin-Watson test statistic is 1.75 for both the linear and log-linear regressions. Exhibit 2 presents selected autocorrelation data from the AR(1) models.

### Exhibit 2. Autocorrelations of the Residual from AR(1) Model

Lag	Autocorrelation	t-Statistic
1	0.4157	5.5768
2	0.2388	3.2045
3	0.0336	0.4512
4	-0.0426	-0.5712

Note: At the 5% significance level, the critical value for t-statistic is 1.97.

After reviewing the data and regression results, Martinez draws the following conclusions.

- Conclusion 1: The time series for WTI oil prices is covariance stationary.
- Conclusion 2: Out-of-sample forecasting using the AR(1) model appears to be more accurate than that of the AR(2) model.

## Example

*d<sub>L</sub>* *d<sub>U</sub>*

In Exhibit 1, at the 5% significance level, the lower critical value for the Durbin-Watson test statistic is 1.75 for both the linear and log-linear regressions.  
 Exhibit 2 presents selected autocorrelation data from the AR(1) models.

### Exhibit 2. Autocorrelations of the Residual from AR(1) Model

Lag	Autocorrelation	t-Statistic
1	0.4157	5.5768
2	0.2388	3.2045
3	0.0336	0.4512
4	-0.0426	-0.5712

AR(2)

Note: At the 5% significance level, the critical value for at-statistic is 1.97.

After reviewing the data and regression results, Martinez draws the following conclusions.

- Conclusion 1: The time series for WTI oil prices is covariance stationary.
- Conclusion 2: Out-of-sample forecasting using the AR(1) model appears to be more accurate than that of the AR(2) model.

# Exhibit 1

Regression Statistics (t-statistics for coefficients are reported in parentheses)				
	Linear	Log-Linear	AR(1)	AR(2)
R <sup>2</sup>	0.5703	0.6255	0.9583	0.9656
Standard error	18.6327	0.3034	5.7977	5.2799
Observations	181	181	180	179
Durbin-Watson	0.10	0.08	1.16	2.08
RMSE			2.0787	2.0530
Coefficients:				
Intercept	28.3278 (10.1846)	3.3929 (74.9091)	1.5948 (1.4610)	2.0017 (1.9957)
t (Trend)	0.4086 (15.4148)	0.0075 (17.2898)		0.9767-1
Oil Price <sub>t-1</sub>			0.9767 (63.9535)	1.3946 (20.2999)
Oil Price <sub>t-2</sub>	<i>t-statistics</i>		0.9767-0.63.9535	-0.4249 (6.2064)

➤ After reviewing the time-series data, Martinez determines that the mean and variance of the time series of oil prices are not constant over time. She then runs the following four regressions using the WTI time-series data.

## Example

4. Martinez's Conclusion 1 is:
- A. correct.
  - B. incorrect because the mean and variance of WTI oil prices are not constant over time. 
  - C. incorrect because the Durbin-Watson statistic of the AR(2) model is greater than 1.75.
5. Based on Exhibit 1, the forecasted oil price in September 2015 based on the AR(2) model is closest to:
- A. \$38.03.
  - B. \$40.04.
  - C. \$61.77.

5. Based on Exhibit 1, the forecasted oil price in September 2015 based on the AR(2) model is closest to:
- A. \$38.03.
  - B. \$40.04.
  - C. \$61.77.
- 7,8 → ↗

6. Based on the data for the AR(1) model in Exhibits 1 and 2, Martinez can conclude that the:
- A. residuals are not serially correlated.
  - B. autocorrelations do not differ significantly from zero.
  - C. standard error for each of the autocorrelations is 0.0745.
7. Based on the mean-reverting level implied by the AR(1) model regression output in Exhibit 1, the forecasted oil price for September 2015 is most likely to be:
- A. less than \$42.86.
  - B. equal to \$42.86.
  - C. greater than \$42.86.

6. Based on the data for the AR(1) model in Exhibits 1 and 2, Martinez can conclude that the:

- A. residuals are not serially correlated.  $\times$
- B. autocorrelations do not differ significantly from zero.  $\times$
- C. standard error for each of the autocorrelations is 0.0745.

$$\sqrt{\frac{T}{n}}$$

**Regression Statistics** (t-statistics for coefficients are reported)

	<b>Linear</b>	<b>Log-Linear</b>	<b>AR(1)</b>
R <sup>2</sup>	0.5703	0.6255	0.9583
Standard error	18.6327	0.3034	5.7977
Observations	181	181	180
Durbin-Watson	<u>0.10</u>	<u>0.08</u>	1.16
RMSE			2.0787

5

 $x_5 \longrightarrow x_1$ 

$$x_5 = 2 + 3 \times 4$$

$$x_4 = 2 + 3 \times 3$$

$$x_3 = 2 + 3 \times 2$$

$$x_2 = 2 + 3 \times 1$$

 $\checkmark \quad \checkmark$

**AR(1)**

0.9583

5.7977

180

1.16

2.0787

**AR(2)**

0.9656

5.2799

179

2.08

2.0530

7. Based on the mean-reverting level implied by the AR(1) model regression output in Exhibit 1, the forecasted oil price for September 2015 is most likely to be:

- A. less than \$42.86.
- B. equal to \$42.86.
- C. greater than \$42.86.

7. Based on the mean-reverting level implied by the AR(1) model regression output in Exhibit 1, the forecasted oil price for September 2015 is most likely to be:

- A. less than \$42.86.
- B. equal to \$42.86.
- C. greater than \$42.86.

$$\frac{b_0}{1-b_1} = 60$$

42.86

7. Based on the mean-reverting level implied by the AR(1) model regression output in Exhibit 1, the forecasted oil price for September 2015 is most likely to be:

- A. less than \$42.86.
- B. equal to \$42.86.
- C. greater than \$42.86.

$$\frac{b_0}{1-b_1} = \cancel{60} \quad 40$$

42.86

# ◆ Simulation

## ➤ Steps in simulation

- Determine "probabilistic" variables
- Define probability distributions for these variables
  - ✓ Historical data
  - ✓ Cross sectional data
  - ✓ Statistical distribution and parameters
- Check for correlation across variables
  - ✓ When there is strong correlation across inputs,
    - ◆ One solution is to pick only one of the two inputs to vary;
    - ◆ The other is to build the correlation explicitly into the simulation.
- Run the simulation



# Simulation

➤ **Advantage of using simulation in decision making**

- Better input estimation
- It yields a distribution for expected value rather than a point estimate

*big picture*

➤ **Simulations with constraints**

- Book value constraints
  - ✓ Regulatory capital restrictions
    - Financial service firms
  - ✓ Negative book value for equity
- Earnings and cash flow constraints
  - ✓ Either internally or externally imposed
- Market value constraints
  - ✓ Explicitly model the effect of distress on expected cash flows and discount rates.



# Simulation

## ➤ Issues in using simulation

- GIGO
- Real data may not fit distributions
- Non-stationary distributions
- Changing correlation across inputs

*Garbage in / out*

# Comparing The Approaches

## ➤ Choose scenario analysis, decision trees, or simulations

- Selective versus full risk analysis
- Type of risk
  - ✓ Discrete risk vs. Continuous risk
  - ✓ Concurrent risk vs. Sequential risk
- Correlation across risk
  - ✓ Correlated risks are difficult to model in decision trees

Risk type and Probabilistic Approaches			
Discrete/ Continuous	Correlated/ Independent	Sequential/ Concurrent	Risk approach
Discrete	Correlated	Sequential	Decision trees
Discrete	Independent	Concurrent	Scenario analysis
Continuous	Either	Either	simulations

## Selective versus full risk analysis

- Type of risk

- ✓ Discrete risk vs. Continuous risk

- ✓ Concurrent risk vs. Sequential risk

$\mu_2/\beta_2$

- Correlation across risk

- ✓ Correlated risks are difficult to model in decision trees

Risk type and Probabilistic Approaches			
Discrete/ Continuous	Correlated/ Independent	Sequential/ Concurrent	Risk approach
Discrete	Correlated	Sequential	Decision trees
Discrete	Independent	Concurrent	Scenario analysis
Continuous	Either	Either	simulations

# ◆ Define Algorithmic Trading

- **Algorithm** is “a sequence of steps to achieve a goal”, and algorithmic trading is “using a computer to automate a trading strategy.”
- **Execution algorithms** are about automating ‘**how to trade**’—that is, how to place orders in the market.
  - Execution algorithms are about minimizing market impact and trying to ensure a fair price.
- **High-frequency trading algorithms** add “**when to trade**” and even sometimes “**what to trade**” .
  - HFT algorithms are about profit.

# The ‘Money Machine’ Application Areas

- Liquidity aggregation and smart order routing
  - Markets have become increasingly fragmented as the number of venues trading the same instruments has proliferated. This phenomenon is known as **market fragmentation** and creates the potential for price and liquidity disparities across venues.
- Real-time pricing of instruments
  - High-frequency pricing can thus influence prices and spreads based on the **up-to-millisecond** view of the market and the tier of the customer.
- Trading on news
  - Firms can trade **automatically on news**, such as announcement of a war, or unexpected weather events, before a human trader can react.
- Genetic tuning
  - The algorithms that have the most profitable theoretical P&L profile can be put into the market to **trade live**. Over time, live algorithms may become less profitable and can be **deactivated**.

# Execution Algorithms Take Various Approaches

- **VWAP** uses the historical trading volume distribution for a particular security over the course of a day and divides the order into slices, proportioned to this distribution. *Volume weighted average*
- **Implementation shortfall** dynamically adjusts the schedule of the trade price in response to market conditions to minimize the difference between the price at which the buy or sell decision was made and the final execution price.
- **Market participation** slices the order into segments intended to participate on a pro-rata basis with volume throughout the course of the execution period.

四  
四  
四

600

$$\frac{590+600}{2}$$

< 600

590 → 601

→ 610

$$\begin{array}{r} \overline{10\bar{1}} \\ \text{buy} \end{array}$$

~~20\bar{1}~~ X

595

V

$$\frac{595+610}{2} > 600$$

< 600

➤ Implementation shortfall dynamically adjusts the schedule of the trade ~~in~~<sup>Pr</sup> ~~to~~  
response to market conditions to minimize the difference between the price  
at which the buy or sell decision was made and the final execution price.

600



610



speed



wst

- **Implementation shortfall** dynamically adjusts the schedule of the trade in response to market conditions to minimize the difference between the price at which the buy or sell decision was made and the final execution price.
- **Market participation** slices the order into segments intended to participate on a pro-rata basis with volume throughout the course of the execution period.

*Strategy*

- **Market participation** slices the order into segments intended to participate on a pro-rata basis with volume throughout the course of the execution period.

strategy

市场参与

逐笔执行



# Data for High-Frequency Trading Algorithms

- **Market data** feeds stream directly from trading venues. (交易场所的数据)
- **Quote events**, is a new bid or offer in the market for a certain instrument at a certain price level and with a certain available quantity (volume)
- **Trade events**, shows a new trade that has taken place at a certain price and a certain volume.
- **News events**, contains **news** related to particular instruments or economic indicators. Although all news offers value, some news is more relevant than other news. If the news contained in a news event merely confirms pre-existing expectations, that event is likely to have a lesser impact than news "surprises."

# Data for High-Frequency Trading Algorithms

- **Market data** feeds stream directly from trading venues. (交易场所的数据)
- **Quote events**, is a new bid or offer in the market for a certain instrument at a certain price level and with a certain available quantity (volume)
- **Trade events**, shows a new trade that has taken place at a certain price and a certain volume.
- **News events**, contains **news** related to particular instruments or economic indicators. Although all news offers value, some news is more relevant than other news. If the news contained in a news event merely confirms pre-existing expectations, that event is likely to have a lesser impact than news "surprises."

All the info.

# HFT-Statistical Arbitrage

- Stat. arb. algorithms monitor instruments that are known to be statistically correlated with the goal of detecting breaks in the correlation that indicate trading opportunities.

- Pairs trading
- Index arbitrage
- Basket trading
- Spread trading

+ in - re  
ETF

- ✓ Crack spread(crude oil, petroleum products)
- ✓ Spark spread(market price of electricity and its cost of production)
- ✓ Crush spread(soybean futures, soybean oil futures, soybean meal)

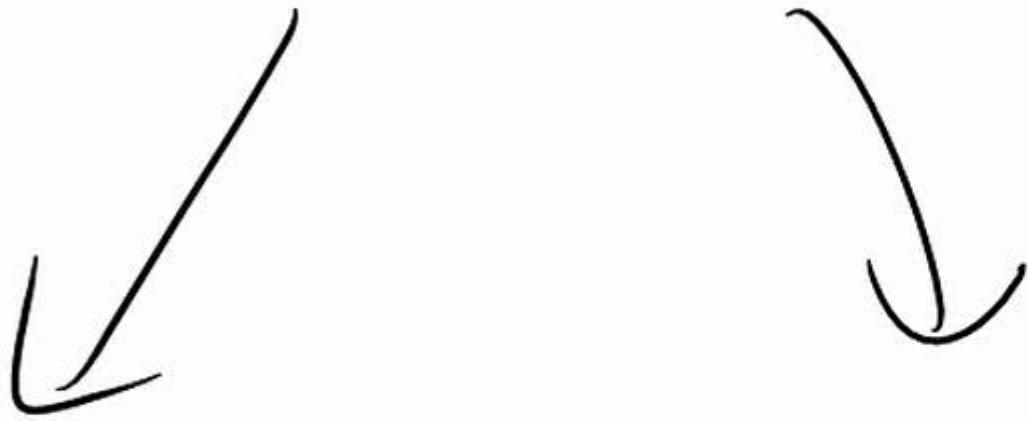
- Mean reversion
- Delta neutral strategies

100

2

98

100 ~~Y~~<sup>f</sup> 币



0.5 中

99.5 法

# The ‘Money Machine’ Application Areas

- Liquidity aggregation and smart order routing
  - Markets have become increasingly fragmented as the number of venues trading the same instruments has proliferated. This phenomenon is known as **market fragmentation** and creates the potential for price and liquidity disparities across venues.
- Real-time pricing of instruments
  - High-frequency pricing can thus influence prices and spreads based on the up-to-millisecond view of the market and the tier of the customer.
- Trading on news
  - Firms can trade **automatically on news**, such as announcement of a war, or unexpected weather events, before a human trader can react.
- ~~Genetic tuning~~
  - The algorithms that have the most profitable theoretical P&L profile can be put into the market to trade live. Over time, live algorithms may become less profitable and can be **deactivated**.

# Latency

- **Latency** is the time **difference** between stimulus and response. Low latency is important to get first mover advantage and to act on an opportunity before a competitor does.
- Low latency decision making is particularly relevant when placing multiple trades as part of a stat arb strategy. This process is called a **multi-legged** trade, in which each trade is a leg.
  - First, it is important to act quickly on the liquidity opportunity seen in the market.
  - Second, it is important not to get "legged out," with one leg of the strategy executing but another leg being confronted with a market that has moved, which means the opportunity is lost.

多单  
空单

# Risk Management and Regulatory Oversight

- HFT can scale the capabilities of a trader hundreds or thousands of times. However, it can also increase trading risk. To complement high-frequency trading, high-frequency pre-trade risk capabilities are needed.
- Many firms embraced this concept some time ago. However, some groups turned off their pre-trade risk management because it increased latency.
- Two approaches being successfully used to mitigate trading risk
  - Real-time pre-trade risk firewall
  - Back testing and market simulation.

➤ The kinds of patterns that can be detected include the following:

- Insider trading
- Front running orders
- Painting the tape(manipulating pricing)

十  
萬 10 万 手

支  
支 2  
1

600  
599 10 手  
10 万 手

支 1

支 2

支 4 支 10 万 手

580

吸筹

- 
- Fictitious orders(quote stuffing, layering and spoofing from Flash Crash of American stock at May 6<sup>th</sup> ,2010)

- Wash trading
- Trade collusion

美

手

美

手

美

手

美

33

1分

5↑

支 1手

支 1手

支 1手

一  
次 3手

foot

5J

➤ The kinds of patterns that can be detected include the following:

- Insider trading
- Front running orders
- Painting the tape(manipulating pricing)
- Fictitious orders(**quote stuffing, layering and spoofing** from Flash Crash of American stock at May 6<sup>th</sup>, 2010)
- Wash trading
- Trade collusion

同买同卖 → 洗盘  
勾结

# ◆ Impact of Algorithmic on Securities Markets

## ➤ Positive ✓

- Minimized market impact of large trades
- Lower cost of execution
- Improved efficiency in certain markets
- More open and competitive trading markets
- Improved and more efficient trading venues

## ➤ Negative ✓

- Fear of an unfair advantage
- Acceleration and accentuation of market movements(no emotion)
- Gaming the market(fictitious orders)
- Increased risk profile
- Algorithms gone wild

妄想

失控