Quantitative Methods

CFA二级培训项目

讲师：王慧琳

101% Contribution Breeds Professionalism

Irene

- **工作职称：**金程教育资深培训师
- **教育背景：**上海财经大学经济学学士，美国约翰霍普金斯大学金融学硕士，CFA持证人。
- **工作背景：**金程教育资深培训师，一次性以高分通过CFA考试，对于考试重点和应试技巧有自己的心得。学术背景深厚，在硕士学习期间，主要研究投资组合管理方向，曾独立撰写NWL公司财务状况分析报告，参与校园基金投资政策声明项目。本科期间曾赴德参加学术交流项目，获得上海市理财规划大赛一等奖。
- **服务客户：**中国银行、中国平安、中国建设银行、工商银行、杭州联合银行、杭州银行、国泰君安证券、苏州元禾控股等

Topic Weightings in CFA Level II

Content	Weightings
Quantitative Methods	**5-10**
Economics	5-10
Financial Statement Analysis	10-15
Corporate Issuers	5-10
Equity	10-15
Fixed Income	10-15
Derivatives	5-10
Alternative Investments	5-10
Portfolio Management	10-15
Ethical and Professional Standards	10-15

Framework

Quantitative Methods

- LM5 Time-Series Analysis
- LM6 Machine Learning
- LM7 Big Data Projects

Quantitative Methods

- LM1 Basics of Multiple Regression and Underlying Assumptions
- LM2 Evaluating Regression Model Fit and Interpreting Model Results
- LM3 Model Misspecification
- LM4 Extensions of Multiple Regression

Learning Module 1

Basics of Multiple Regression and Underlying Assumptions

Framework

1. Linear regression
   - Multiple linear regression
2. Assumptions
3. Detection of violations: diagnostic plots
   - Scatter plots
   - Residual plots
4. Regression process

1. Multiple Linear Regression

- The Multiple Linear Regression model

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + \varepsilon$$

  - k = number of independent variables (k=1 simple linear regression)
  - n = number of observations (n must > k)
  - b_0 = intercept: estimated value of Y when X_1, X_2,..., X_k are all equal to zero.
  - b_1,..., b_k = partial slope coefficients: the expected change in the dependent variable for a 1-unit increase in an independent variable, holding all the other independent variables constant.
  - ε=error/residual term: the stochastic or random part of the model.
- Predicted value of the dependent variable

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

2. Multiple Regression Assumptions

- **1. Linearity:** The relationship between the dependent variable and the independent variables is linear.
- **2. Homoskedasticity:** The variance of the regression residuals is the same for all observations.
- **3. Independence of errors:** The observations are independent of one another. This implies the regression residuals are uncorrelated across observations.
- **4. Normality:** The regression residuals are normally distributed.
- **5. Independence of independent variables:**
  - 5a. Independent variables are not random.
  - 5b. There is no exact linear relation between two or more of the independent variables or combinations of the independent variables.

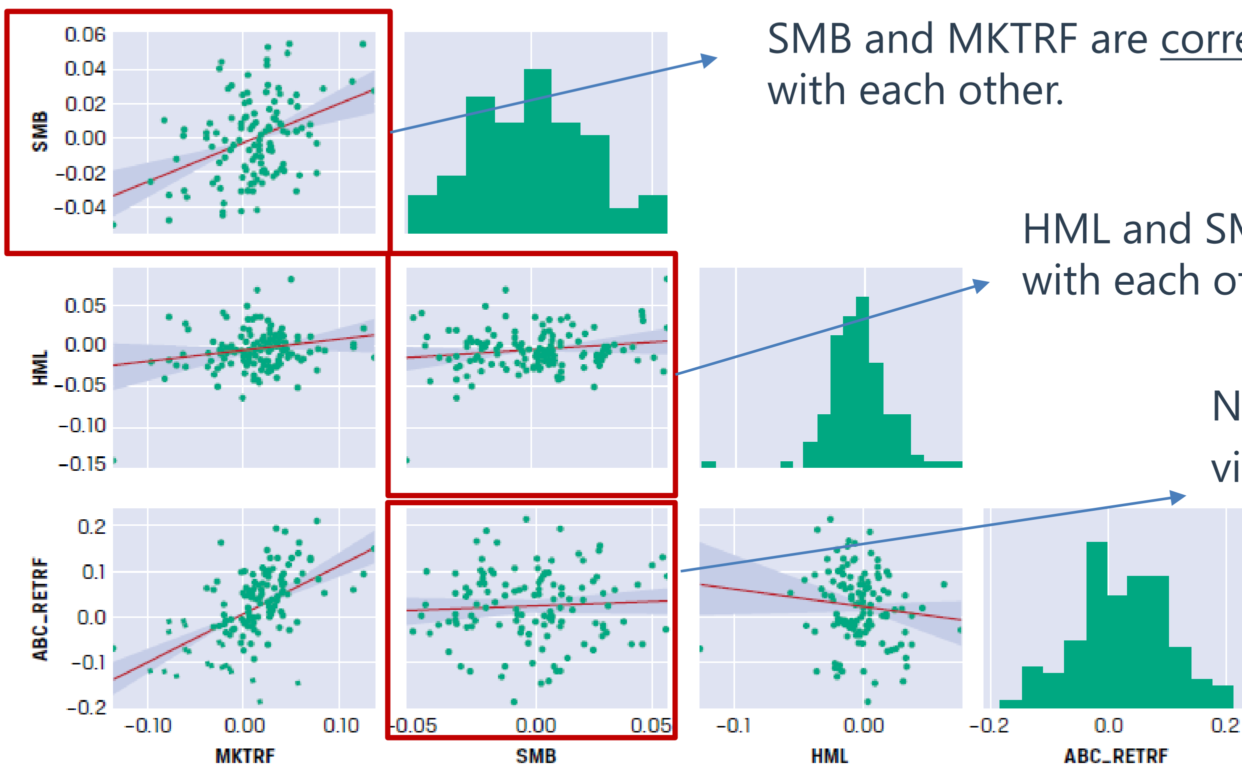
3. Detection of Violations

- **Diagnostic plots** can help detect whether these assumptions are satisfied.
  - **Scatterplots of dependent and independent variables**
    - ✓ Useful for detecting non-linear relationships.
  - **Scatterplot of residuals**
    - ✓ Residuals vs. Predicted value of dependent variable or independent variables
    - ✓ Useful for detecting violations of homoskedasticity, independence of errors and independence of independent variables.
  - **Scatterplot matrix**
    - ✓ Can be used to identify extreme values and outliers.

3.1 Violation of Independence of Independent Variables

➢ **Example:**

$$ABC\_RETRF_t = b_0 + b_1MKTRF_t + b_2SMB_t + b_3HML_t + \varepsilon_t$$



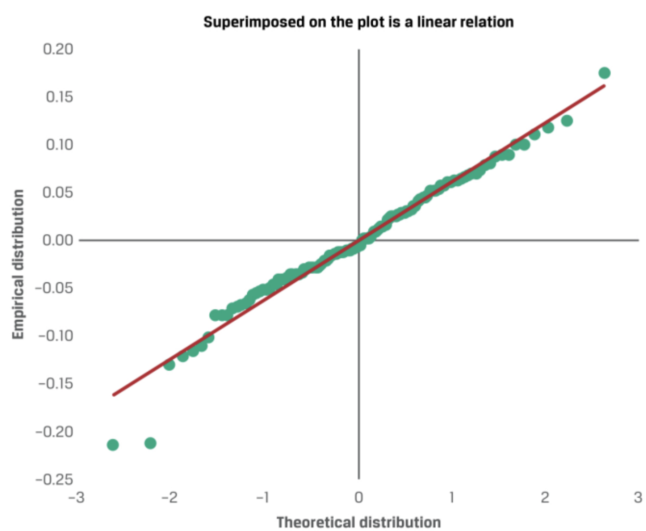
SMB and MKTRF are <u>correlated</u> with each other.

HML and SMB are <u>uncorrelated</u> with each other.

Near-zero slope indicates violation of linearity.

3.2 Violation of Normality

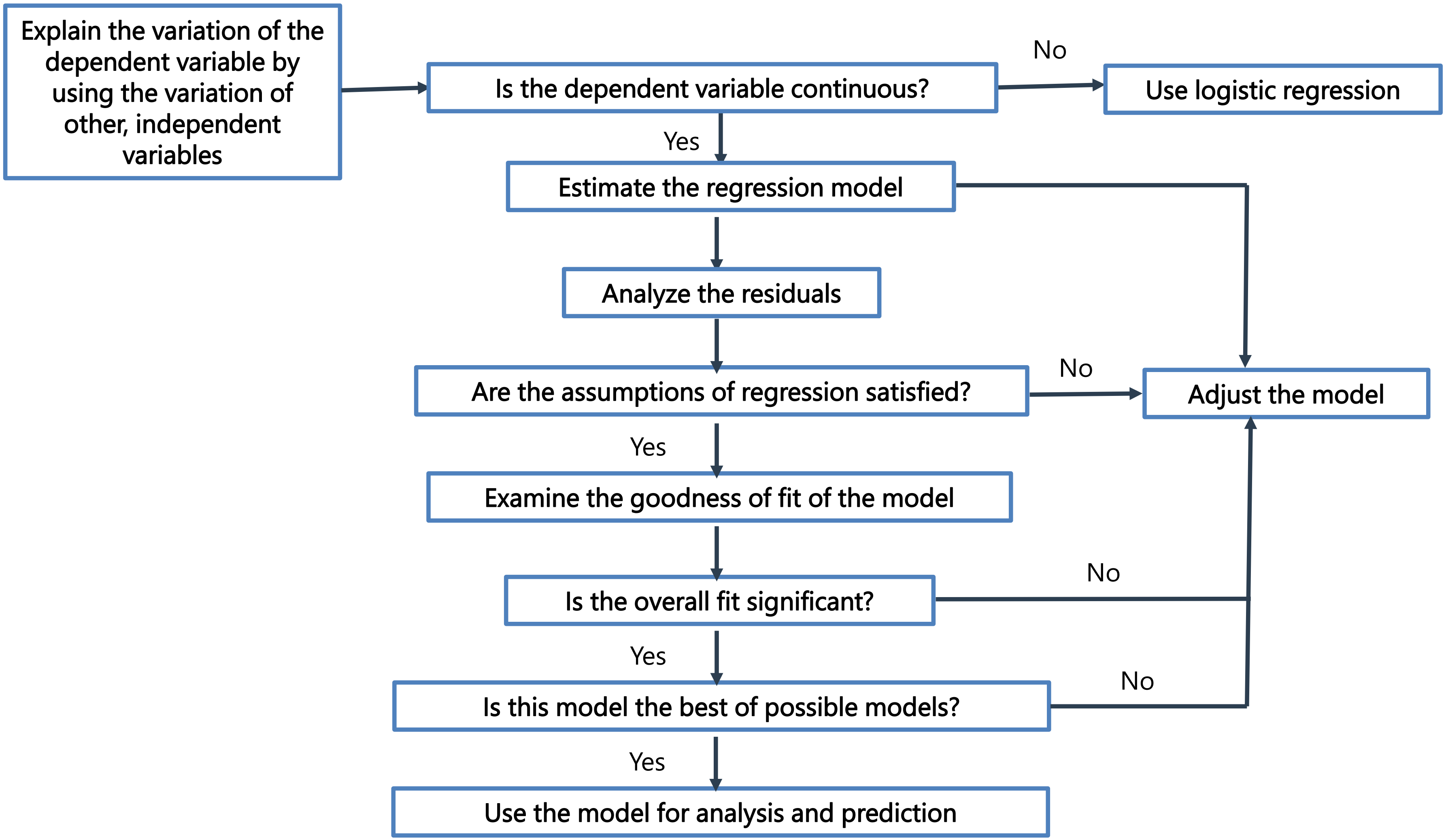
- **(Normal) Q-Q Plot**
  - Q-Q plot is used to visualize the distribution of a variable by comparing it to a normal distribution.
    - ✓ E.g. Use a Q-Q plot to compare the model's standardized residuals to a theoretical standard normal distribution.



Superimposed on the plot is a linear relation

If the residuals are normally distributed, they should align along the diagonal.

4. Regression Process



Learning Module 2

Evaluating Regression Model Fit and Interpreting Model Results

Framework

1. Measures of goodness of fit
   - ANOVA table
   - R^2
   - Adjusted R^2
   - AIC & BIC
2. Significance test for regression coefficient
   - Joint hypothesis tests
   - General linear F-test
3. Forecasting using multiple regression

1.1 Measures of goodness of fit-ANOVA table

- **Analysis of variance (ANOVA) table**

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

- **Coefficient of determination (R^2)**

- $$R^2 = \frac{RSS}{SST} = 1 - \frac{SSE}{SST}$$
$$= \frac{explained\ variation}{total\ variation} = 1 - \frac{unexplained\ variation}{total\ variation}$$

✓ $0 \le R^2 \le 1$

✓ The higher R^2 , the better fitness.

1.2 Adjusted R^2

- **Adjusted R^2 ($\bar{R}^2$):**
  - $\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$
    - ✓ adjusted for degrees of freedom
    - ✓ if k>=1, R^2 is strictly greater than adjusted R^2
    - ✓ adjusted R^2 may be less than zero
    - ✓ a high adjusted R^2 does not necessary mean the correct choice of variables
  - The following are two key observations about $\bar{R}^2$ **when adding a new variable** to a regression:
    - ✓ If the coefficient's $|t\text{-statistic}| > 1.0$, then $\bar{R}^2$ increases.
    - ✓ If the coefficient's $|t\text{-statistic}| < 1.0$, then $\bar{R}^2$ decreases.

1.3 R^2 and adjusted R^2

- **Both of R^2 and adjusted R^2 cannot:**
  - Provide information on whether the coefficients are **statistically significant**.
  - Provide information on whether there are **biases in the estimated coefficients and predictions**.
  - Tell whether the **model fit is good**.
- Therefore, we explore the ANOVA further,
  - Calculating the **F-statistic**
  - Other goodness-of-fit metrics **(AIC & BIC)**.

1.4 AIC & BIC

➢ Akaike's information criterion **(AIC)** and Schwarz's Bayesian information criteria **(BIC)** are used to **evaluate model fit** and **select the "best" model** among a group with the same dependent variable.

- $\mathbf{AIC} = \boldsymbol{n}\ln\left(\frac{\textbf{Sum of squares error}}{\boldsymbol{n}}\right) + \mathbf{2}(\boldsymbol{k}+\mathbf{1})$
- $\mathbf{BIC} = \boldsymbol{n}\ln\left(\frac{\textbf{Sum of squares error}}{\boldsymbol{n}}\right) + \ln(\boldsymbol{n})\,(\boldsymbol{k}+\mathbf{1})$
  - ✓ $2(k+1)$ or $\ln(n)\,(k+1)$ is the **penalty** assessed for adding independent variables to the model.
  - ✓ Since ln(n) >2 (for n≥ 8), BIC assesses a <u>greater penalty</u> for having more parameters in a model.
  - ✓ **Lower AIC & BIC** indicates a **better-fitting model.**

1.4 AIC & BIC

- Practically speaking**:**
  - AIC is preferred if the model is used for prediction purposes.
  - BIC is preferred when the best goodness of fit is desired.
- Importantly,
  - the value of these measures considered alone is **meaningless**;
  - the **relative values** of AIC or BIC among a set of models is what really matters.

2. Testing Joint Hypotheses for Coefficients

2.1 Tests of a single coefficient

- H_0: b_1= hypothesized value of b_1
- Test Statistic:

$$t = \frac{\hat{b}_1 - hypothesized\ value\ of\ b_1}{s_{\hat{b}_1}}\text{, df=n-k-1}$$

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

2. Testing Joint Hypotheses for Coefficients

- **2.2 Joint F-test**
  - H_0: $b_m = b_{m+1} = \ldots = b_{m+q-1} = 0$
  - H_a: At least one slope of the q slopes $\neq 0$.
    - ✓ m is the first restricted slope,
    - ✓ m + 1 is the second restricted slope, and so on, up to the *q*th restricted slope.
- **F-statistic**$= \dfrac{(\textbf{Sum of squares error restricted model} - \textbf{Sum of squares error unrestricted})/q}{\textbf{Sum of squares error unrestricted model}/(n-k-1)}$
  - q is the number of restrictions
- **Critical value (查表):** F_α (q, n-k-1) "**one-tailed**".
- **Decision rule:**
  - Reject H_0 : if F-statistic > F_α (q, n-k-1)

1606 (1) 1706 (1) 1806 (1) 1906(1) 2012(1) 2108(1) 2202(1) 2305(1) 2311(1)

2. Testing Joint Hypotheses for Coefficients

- **2.3 General linear F-test**
  - Tests the null hypothesis that slope coefficients on all variables are equal to zero:
    - ✓ Assesses the effectiveness of the model as a whole in explaining the dependent variable.
- **Define hypothesis:**
  - H_0: $b_1 = b_2 = b_3 = \ldots = b_k = 0$
  - H_a: at least one $b_j \neq 0$ (for j = 1, 2, …, k)
- **F-statistic**$= F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/(n-k-1)}$
- **Critical value (查表)**: F_{α} (k, n-k-1) "**one-tailed**" F-test; alpha=5%
- **Decision rule**
  - Reject H_0 : if F-statistic > F_{α} (k, n-k-1)

3. Forecasting Using Multiple Regression

- Predicting the value of the dependent variable
  - $\hat{Y} = \hat{b}_0 + \hat{b}_1\hat{X}_1 + \hat{b}_2\hat{X}_2 + \cdots + \hat{b}_k\hat{X}_k$
- **Two sources of uncertainty** when using the regression model to predict the dependent variable.
  - Model error: The error term itself contains uncertainty.
  - Sampling error: Uncertainty in the independent variable forecasts.
- The **confidence interval** around the forecasted value of the dependent variable reflects both *model error* and *sampling*.
  - the **larger** the sampling error, the **larger** is the standard error of the forecast of Y and the **wider** is the confidence interval.

Learning Module 3

Model Misspecification

Framework

1. Model Specification
   - Omitted variables
   - Inappropriate form of variables
   - Inappropriate variable scaling
   - Inappropriate data pooling
2. Multiple Regression Assumption Violations
   - Heteroskedasticity
   - Serial correlation
   - Multicollinearity

1.1 Model Specification

- **Principles for Proper Regression Model Specification**
  - Model should be grounded in **economic reasoning**.
  - Model should be **parsimonious**.
  - Model should **perform well out of sample**.
  - Model **functional form should be appropriate**.
  - Model should **satisfy regression assumptions**.

1.2 Misspecified Functional Form

Failures in Regression Functional Form	Explanation	Consequence
Omitted variables	One or more important variables are omitted from the regression.	May lead to heteroskedasticity or serial correlation
Inappropriate form of variables	Ignoring a nonlinear relationship between the dependent and independent variable	May lead to heteroskedasticity
Inappropriate variable scaling	One or more regression variables may need to be transformed before estimating the regression	May lead to heteroskedasticity or Multicollinearity
Inappropriate data pooling	Regression model pools data from different samples that should not be pooled	May lead to heteroskedasticity or serial correlation

1.2 Misspecified Functional Form

- **Omitted Variable**
  - If the omitted variable is uncorrelated with existing independent variable.
    - ✓ The estimate of the intercept will be **biased,**
    - ✓ The coefficient of existing independent variable will still be **estimated correctly**.
  - If the omitted variable is correlated with the existing variable
    - ✓ The estimated coefficient on X_j, the intercept, and the residuals **will be incorrect**.
    - ✓ The estimates of the coefficients' standard errors will also be **inconsistent**, so these cannot be used for conducting statistical tests.

2. Multiple Regression Assumption Violations

- **Three Multiple Regression Assumption Violations**
  - Heteroskedasticity（异方差）
  - Serial correlation (autocorrelation) （序列相关，自相关）
  - Multicollinearity（多重共线性）

2.1 Heteroskedasticity

- Heteroskedasticity may **arise from** model misspecification, including:
  - omitted variables,
  - incorrect functional form,
  - incorrect data transformations,
  - extreme values of independent variables.
- **Effect of heteroskedasticity on regression analysis**
  - **Not** affect:
    - ✓ **Consistency** of regression parameter estimators ($\hat{b}_j$).
  - Heteroskedasticity introduces **bias** into estimators of the **standard error** of regression coefficients.
    - ✓ **t-tests** for the significance of individual regression coefficients are **unreliable.**
      - In regressions with financial data, the most likely impacts of conditional heteroskedasticity are that standard errors will be underestimated, so t-statistics will be inflated. **(Type I error)**.
    - ✓ The **F-test** for overall significance of the regression is **unreliable**.

2.1 Heteroskedasticity

- **Testing for Conditional Heteroskedasticity**
  - **Residual scatter plots** (residual vs. independent variable)
  - The **Breusch-Pagan χ^2 test**
    - ✓ H_0: No heteroskedasticity, one-tail, right-side test
    - ✓ Chi-square test: $\chi^2_{BP,k} = n \times R_{residual}^2$, df=k
      - ◆Tips: Regress squared residuals with independent variable, X, and $R_{residual}^2$ is the coefficient of determination.
    - ✓ Decision rule: if BP test statistic > critical value or p-value<alpha, reject H_0, indicating conditional heteroskedastic residuals.
- **Correcting heteroskedasticity**
  - Computing **robust standard errors**, to correct the standard error of estimated coefficients, (aka. White-corrected standard error, Heteroskedasticity-consistent standard error).

2.2 Serial Correlation

- Serial correlation (or Autocorrelation) is often found **in time series data** and **panel data**.
  - **Positive** serial correlation is much more **common** in economic and financial data.
- **Effect of serial correlation**

Independent Variable Is Lagged Value of Dependent Variable	Invalid Coefficient Estimates	Invalid Standard Error Estimates
No	No	Yes
Yes	Yes	Yes

- Positive serial correlation → **Type I error & F-test, t-test unreliable**.
  - ✓ Standard errors for the regression coefficients are underestimated, so t-statistics are inflated →the prob. of type I error increased.
  - ✓ F-statistic may be inflated because the mean squared error will tend to underestimate the population error variance.

2.2 Serial Correlation

- **Testing Serial Correlation**
  - **Durbin-Watson (DW) test**
    - ✓ Compares the squared differences of successive residuals with the sum of the squared residuals.
    - ✓ **Limitation**: applies only to testing for **first-order** serial correlation.
      - ◆ H_0: No serial correlation

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

      - ◆ Decision rule

Reject H_0, conclude positive serial correlation	Inconclusive	Do not reject H_0	Inconclusive	Reject H_0, conclude negative serial correlation
0 – d_L	d_L – d_U	d_U – 4-d_U	4-d_U – 4-d_L	4-d_L – 4

2.2 Serial Correlation

- **Testing Serial Correlation**
  - **Breusch–Godfrey (BG) test**
    - ✓ **More robust** because it can detect autocorrelation up to a pre-designated order p.
    - ✓ Step 1: run the initial regression
      - ◆ $Y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + u_t$
    - ✓ Step 2: run a new model with the fitted residuals from Step 1:
      - ◆ e.g.: $\hat{u}_t = a_0 + a_1 X_{1t} + a_2 X_{2t} + p_1 u_{t-1} + e_t$, order p = 1.
    - ✓ Step 3: Test hypotheses
      - ◆ H_0: $p_1 = 0$, no serial correlation in the model's residuals up to lag p.
      - ◆ H_a: $p_1 \neq 0$.
      - ◆ Test statistic is approximately ***F*-distributed** with <u>$n - p - k - 1$ and p degrees of freedom</u>, where p is the number of lags.
    - ✓ Step 4: Make decision.

2.2 Serial Correlation

- **Correcting for Serial Correlation**
  - **Adjust the coefficient standard errors** to account for the serial correlation.
    - ✓ The corrections are known by various names, including **serial-correlation consistent standard errors**, serial correlation and heteroskedasticity adjusted standard errors (HAC), Newey–West standard errors, and robust standard errors.
    - ✓ An **advantage** of these methods is that they also correct for conditional heteroskedasticity.

2.3 Multicollinearity

- In practice, **multicollinearity is often a matter of degree**.
  - Multicollinearity may occur:
    - ✓ when two or more independent variables are highly correlated;
    - ✓ or when there is an approximate linear relationship among independent variables.
- **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 →a little power to reject the null hypothesis. (Type II error)

2.3 Multicollinearity

- **Methods to Detect Multicollinearity**
  - **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.
    - ✓ A high R^2 would reflect the overall significance of the regression (significant F-Statistic).
  - **Check pairwise correlations**: Using the magnitude of **pairwise correlations** among the independent variables.
    - ✓ High pairwise correlations among the independent variables can usually indicate multicollinearity **(|r|>0.7)**.

2.3 Multicollinearity

- **Methods to Detect Multicollinearity (cont.)**
  - **Variance inflation factor (VIF):**
    - ✓ $VIF_j = \frac{1}{1-R_j^2}$
    - ✓ The minimum VIF_j is 1.
    - ✓ VIF increases as the correlation increases.
      - ◆ $VIF_j > 5$ warrants further investigation of the given independent variable.
      - ◆ $VIF_j > 10$ indicates serious multicollinearity requiring correction.
- **Methods to Correct Multicollinearity**
  - Excluding one or more of the regression variables;
  - Using a different proxy for one of the variables;
  - Increasing the sample size.

Summary of Assumption Violations

Assumption violation	Impact	Detection	Solution
Conditional heteroskedasticity	Biased estimates of coefficients' standard errors	✓ Visual inspection of residuals; ✓ Breusch-Pagen test	✓ Revise model; ✓ Use robust standard errors
Serial correlation	Inconsistent estimates of coefficients and biased standard errors	✓ Durbin-Watson test; ✓ Breusch– Godfrey test	✓ Revise model; ✓ Use serial correlation consistent standard errors
Multicollinearity	Inflated standard errors	✓ Classic method; ✓ Check pairwise correlations; ✓ Variance inflation factor	✓ Revise model; ✓ Increase sample size

Learning Module 4

Extensions of Multiple Regression

Framework

1. Influence Analysis
   - Leverage
   - Studentized residual
   - Cook's distance
2. Qualitative Independent Variable
   - Dummy Variables
3. Qualitative Dependent Variable
   - Logistic Regression

1. Influence Analysis

- **Influential Data Points:** Two kinds of observations may potentially influence regression results.
  - A **high-leverage point**, a data point having an extreme value of an independent variable.
  - An **outlier**, a data point having an extreme value of the dependent variable.
- **Detecting Influential Points**
  - Single linear regression: **Scatterplot**
  - Multiple linear regression: Quantitative way
    - ✓ **Leverage;**
    - ✓ **Studentized residual;**
    - ✓ **Cook’s distance.**

1.1 Detecting Influential Points

- **Leverage (h_{ii}):** identifying high-leverage points.
  - measures the distance between the value of the *i*th observation of that independent variable and the mean value of that variable across all *n* observations.
  - h_{ii} has a value between 0 and 1.
  - higher the leverage, the more influence the *i*th observation can potentially exert on the estimated regression.
  - $h_{ii} > \mathbf{3}\left(\frac{\boldsymbol{k+1}}{\boldsymbol{n}}\right)$, then it is a <u>potentially influential observation</u>.
    - ✓ k=number of independent variables;
    - ✓ n=number of observations.

1.2 Detecting Influential Points

- **Studentized residual (t_{i^*}):** identifying outliers

  - $t_{i^*} = \frac{e_{i^*}}{S_{e^*}} = e_i \times \sqrt{\frac{n-k-1}{SSE(1-h_{ii})-e_i^2}}$
    - ✓ e_{i^*} is the residual with the ith observation deleted.
    - ✓ S_{e^*} is the standard deviation of the residuals.
    - ✓ h_{ii} is the leverage value for the ith observation.
  - Rule of thumb:
    - ✓ $|\boldsymbol{t_{i^*}}| > \mathbf{3}$ → Flag observation as being an **outlier**.
    - ✓ $|\boldsymbol{t_{i^*}}| >$ **critical value** of t−statistic with n – k – 2 degrees of freedom at selected significance level → Flag outlier observation as being **potentially influential**.

1.3 Detecting Influential Points

➢ **Cook's distance, or Cook's *D*** **(**D_i **)**

- $D_i = \frac{e_i^2}{k \times MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$
- It depends on both residuals and leverages, so it is a composite measure for detecting extreme values of both types of variables.
- It summarizes how much all of the regression's estimated values change when the ith observation is deleted from the sample.
- A **large** $\boldsymbol{D_i}$ indicates that the ith observation **strongly influences** the regression's estimated values.
  - ✓ D_i >0.5 → The ith observation may be influential and merits further investigation.
  - ✓ $D_i > 1.0$ → The ith observation is highly likely to be an influential data point.
  - ✓ $\boldsymbol{D_i > 2 \times \sqrt{k/n}}$ → **The ith observation is highly likely to be an influential data point.**

Summary of Influence Analysis

Measure	Y	X	Process	Is observation influential?
Leverage		√	h_{ii} ranges from 0 to 1	If $h_{ii} > 3(\frac{k+1}{n})$, then potentially influential
Studentized residual	√		Compare calculated |*t*-statistic| with critical *t*-value	If calculated |*t*-statistic| > critical *t*-value, then potentially influential
Cook's distance	√	√	Compare calculated Cook's D against $2 \times \sqrt{k/n}$	If calculated Cook's *D* $> 2 \times \sqrt{k/n}$, then highly likely influential

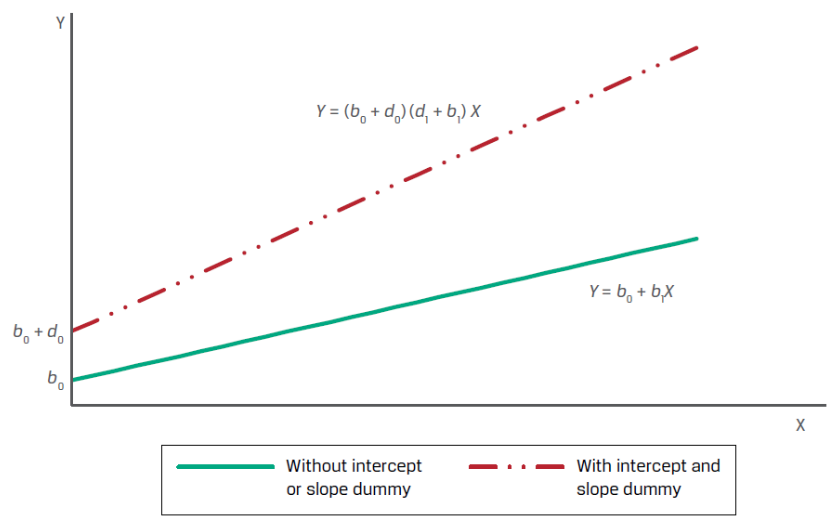
2.1 Defining Dummy Variables

- One type of **qualitative independent variable**, called a **(simple) dummy variable,** or **indicator variable.**
  - $X=\begin{cases}1 & true \\ 0 & false\end{cases}$
    - ✓ 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.
- Types of dummy variables:
  - Intercept Dummy;
  - Slope Dummy.

2.2 Different Types of Dummy Variables

- **Dummies in both slope and intercept**
  - $Y = b_0 + d_0 D_i + b_1 X_i + d_1\, D_i X_i + \varepsilon_i$
    - ✓ If D = 0, then the equation becomes $Y = b_0 + b_1 X + \varepsilon$ (*base category*).
    - ✓ If D=1, then $Y = (b_0{+}d_0) + (b_1{+}d_1)X + \varepsilon$ (*category to which both changed intercept and changed slope apply*).
    - ✓ The slope dummy variable creates an **interaction term** between the X variable and the condition represented by D = 1.

Panel C: Model with intercept and slope dummy variable



专业 · 创新 · 增值

3.1 Logistic Regression

- The **logistic transformation** tends to linearize the relation between the dependent and independent variables.
  - $\ln(\frac{P}{1-P})$, *P* refers to a condition is fulfilled or an event happens.
  - The natural logarithm (ln) of the odds of an event happening is the **log odds.**
- **Logistic Regression**
  - ✓ $\ln(\frac{P}{1-P}) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \varepsilon$
  - ✓ $P = \frac{1}{1+\exp[-(b_0+b_1X_1+b_2X_2+b_3X_3)\ ]}$
  - Logistic regression coefficients are typically estimated using the **maximum likelihood estimation (MLE)** method.
    - ✓ **Slope:** change in the log odds that the event happens per unit change in the independent variable, *holding all other independent variables constant.*
    - ✓ **Intercept**: log odds of Y=1 if all independent variables are zero.

3.2 Hypothesis test

- **Test for single coefficient**
  - the same process as the test in ordinary least squares regression.
- **Test For model fitness**
  - **Likelihood ratio (LR) test**
    - ✓ Unrestricted Model A: $\ln(\frac{P}{1-P}) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \varepsilon$
    - ✓ Restrictions model B, Model B: $\ln(\frac{P}{1-P}) = b_0 + b_1X_1 + \varepsilon$
  - H_0: $b_2 = b_3 = 0$; H_a: at least one of the coefficients is different from zero.
  - LR = −2 (Log likelihood restricted model − Log likelihood unrestricted model)
    - ✓ log-likelihood <0, so higher values (|log-likelihood |↓) → better fitting model.
  - Rejecting the null hypothesis is a rejection of the **smaller, restricted model** in favor of the larger, unrestricted model.

Learning Module 5

Time-Series Analysis

专业 · 创新 · 增值

Framework

1. Trend models
2. Autoregressive models (AR)
   - Chain rule of forecasting
   - Assumption
     - No autocorrelation
     - Covariance-stationary series
     - No conditional heteroskedasticity
   - RMSE
3. Regression with more than one time series

1906(1) 2012(1) 2108(1) 2105(1) 2311(1)

Trend Models

- **Linear trend model**
  - $y_t = b_0 + b_1 t + \varepsilon_t$
- **Log-linear trend model**
  - $y_t = e^{(b0+b1t)}$
  - $Ln(y_t) = b_0 + b_1 t + \varepsilon_t$
- **Factors that Determine Which Model is Best**
  - A linear trend model may be appropriate if the data points appear to be equally distributed above and below the regression line (inflation rate data).
    - ✓ **Growth at a constant amount**
  - A log-linear model may be more appropriate if the data plots with a non-linear (curved) shape.
    - ✓ **Growth at a constant rate**
    - ✓ **Persistently positive or negative**
- **Limitations of Trend Model**
  - Usually the time series data exhibit serial correlation.
    - ✓ Use the Durbin Watson statistic to detect autocorrelation

Autoregressive Models (AR)

- **An autoregressive model uses past values of dependent variables as independent variables**
  - AR(p) model:

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

- **Chain rule of forecasting**
  - A one-period-ahead forecast for an AR (1) model is determined in the following manner:

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

  - Likewise, a two-step-ahead forecast for an AR (1) model is calculated as:

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

Autoregressive Models (AR)

- **Assumption violations**
  - No autocorrelation
  - covariance-stationary series
  - No conditional heteroskedasticity

Autoregressive Models (AR)

- **Detecting autocorrelation in an AR model**
  - Compute the autocorrelations of the residual
  - t-tests to see whether the residual autocorrelations differ significantly from 0,

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

n is the number of observations in the time series.

  - If the residual autocorrelations differ significantly from 0, the model is not correctly specified, so we may need to modify it (eg. seasonality)
  - Correction: add lagged values

1706(1) 2202(1) 2308(1)

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 **adds** a seasonal lag in an AR model
- For example: $x_t=b_0+b_1\ x_{t-1}+\ b_2\ x_{t-4}+\varepsilon_t$

Autoregressive Models (AR)

- **Covariance-stationary series**
  - Statistical inference based on OLS estimates for a lagged time series model assumes that the time series is covariance stationary
  - Three conditions for covariance stationary
    - ✓ **Constant** and finite expected value of the time series
    - ✓ **Constant** and finite variance of the time series
    - ✓ **Constant** and finite covariance with leading or lagged values
  - 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 exhibits mean reversion if it has a tendency to move towards its mean
  - For an AR(1) model, the mean reverting level is: $x_i = \dfrac{b_0}{1 - b_1}$
  - If $x_i > \dfrac{b_0}{1 - b_1}$ the model predicts that x_{t+1} will be lower than x_t,
  - if $x_i < \dfrac{b_0}{1 - b_1}$ the model predicts that x_{t+1} will be higher than x_t

> 均值复归的反面：
> Autoregressive model 如果没有mean reverting level说明follow random walk.

Random Walks

- **Random walk (unit root)**
  - A special AR(1) model with $b_0=0$ and $b_1=1$
  - Simple random walk: $x_t = x_{t-1} + \varepsilon_t$
  - The best forecast of x_t is x_{t-1}
- **Random walk with a drift**
  - $x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$
  - $b_0 \neq 0$, $b_1 = 1$
  - The time series is expected to increase/decrease by a constant amount

Random Walks

- **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 b1=1 is invalid to test the unit root
- **Dickey-Fuller test (DF test) to test the unit root**
  - Start with an AR(1) model $x_t=b_0+b_1 x_{t-1}+\varepsilon_t$

    Subtract x_{t-1} from both sides $x_t-x_{t-1}=b_0+(b_1-1)x_{t-1}+\varepsilon_t$

    $$\mathbf{x_t-x_{t-1}=b_0+g\,x_{t-1}+\varepsilon_t}$$
  - 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
  - If we can **reject the null**, the time series **does not have a unit root** and is stationary

Random Walks – if a time series appears to have a unit root

- If a time series appears to have a unit root, how should we model it ???
- 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**
  - Define y_t as $y_t = x_t - x_{t-1}$
  - This is an AR(1) model $y_t = b_0 + b_1\, y_{t-1} + \varepsilon_t$
  - The first-differenced variable y_t is covariance stationary

Autoregressive Conditional Heteroskedasticity (ARCH)

- **Heteroskedasticity** refers to the situation that the variance of the error term is not constant

多元回归中用BP test

- **Test whether a time series is ARCH(1)**
  - $\varepsilon_t^2 = a_0 + a_1\varepsilon_{t-1}^2 + u_t$
  - If the coefficient a1 is significantly different from 0, the time series is ARCH(1)
- If ARCH exists,
  - Generalized least squares must be used to develop a predictive model
  - Use the ARCH model to predict the variance of the residuals in following periods

Compare forecasting power with RMSE

- **Comparing forecasting model performance**
  - **In-sample forecasts** are within the range of data (i.e., time period) used to estimate the model, which for a time series is known as the sample or test period.
  - **Root mean squared error (RMSE)**: the model with the **smallest RMSE** is most accurate for **out-of-sample**
    - ✓ **Out-of-sample** forecasts are made outside. In other words, we compare how accurate a model is in forecasting the y variable value for a time period outside the period used to develop the 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**
  - 1. None of the time series has a unit root: we can use multiple regression
  - 2. At least one time series has a unit root while at least one time series does not: we cannot use multiple regression
  - 3. 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

Learning Module 6

Machine Learning

Framework

1. Overview of machine learning
2. Supervised Learning Algorithms
3. Unsupervised Learning Algorithms
4. Neutral networks

1.1 Defining Machine Learning

- **Machine learning** seeks to extract knowledge from large amounts of data with no such restrictions. The **goal** of machine learning algorithms is to automate decision-making processes **by generalizing (i.e., "learning")** from known examples to determine an underlying structure in the data.
- **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**
    - ✓ X variables are called **features**
  - **Hyperparameter**: model input specified by the researcher.

1.2 Types of Machine Learning

➢ **Supervised learning** uses **labeled training data** to guide the ML program toward superior forecasting accuracy.

- Applying the ML algorithm to this data set to infer the pattern between the inputs and output is called "**training**" the algorithm.
- Once the algorithm has been trained, the inferred pattern can be used to **predict output** values based on new inputs (i.e., ones not in the training data set).

✓ Two types of supervised learning

- ☐ **Regression model**: making prediction of **continuous** target variables
  - ✓ Multiple regression is an example of supervised learning.
- ☐ **Classification model**: sorting observations into distinct categories
  - ✓ Binary classification – e.g. default or not likely default
  - ✓ Multicategory classification – e.g. bond rating

Types of Machine Learning

Unsupervised learning

- In unsupervised learning, the ML program **is not given labeled training** data; instead, inputs (i.e., features) are provided without any conclusions about those inputs.
  - ✓ The algorithm seeks to **discover structure within** the data themselves.
  - ✓ Two types of unsupervised learning
    - **Dimension reduction** focuses on **reducing the number of features** while retaining variation across observations to preserve the information contained in that variation.
    - **Clustering** focuses on **sorting observations into groups** (clusters) such that observations in the same cluster are more similar to each other than they are to observations in other clusters.

Types of Machine Learning

- **Neural networks** (NNs, also called artificial neural networks, or ANNs) include highly flexible ML algorithms that have been successfully applied to a variety of tasks characterized by **non-linearities** and interactions among features.
  - Deep learning and reinforcement learning are themselves **based on neural networks.**
  - In deep learning, sophisticated algorithms address highly complex tasks, such as image classification, face recognition, speech recognition, and natural language processing.
  - In reinforcement learning, a computer learns from interacting with itself (or data generated by the same algorithm).

1.3 Data Sets

➢ To **measure** how well a model generalizes, data analysts create three **nonoverlapping** data sets:

- **Training sample** (used to develop the model) In-sample
  - ✓ In-sample prediction errors occur with the training sample
- **Validation sample** (used for tuning the model) Out-of-sample
- **Test sample** (used for evaluating the model using new data)

1.4 Overfitting

- **Challenges of Machine Learning**
  - Underfitted: make too little use of the data
  - Overfitting: make too much use of the data
- **Overfitting** is an issue with **supervised ML** that results when a **large number of features** are included in the data sample, resulting that the fitted algorithm <u>does fit well to training data but not generalize well to new data</u>.
  - It results in inaccuracy forecasts on out of sample data, randomness is misperceived to be a pattern
    - ✓ When a model **generalizes well**, it means that the model <u>retains its explanatory power</u> when it is applied to new (i.e., out-of-sample) data.

Overfitting

- Data scientists then decompose these errors into the following:
  - **Bias error.** This is the in-sample error resulting from models with a poor fit.
  - **Variance error.** This is the out-of-sample error resulting from overfitted models that do not generalize well.
  - **Base error.** These are residual errors due to random noise.
- Variance error increases with model complexity, while bias error decreases with complexity. Data scientists often express this as a trade-off between cost and complexity.
  - An optimal level of complexity **minimizes the total error** and is a key part of successful model generalization.
  - **Linear functions** are more susceptible to bias error and underfitting;
  - **Non-linear functions** are more prone to variance error and overfitting.

Overfitting – Addressing methods

- **Two common guiding principles and two methods are used to reduce overfitting:**
  - 1) preventing the algorithm from getting too complex during selection and training, which requires estimating an **overfitting penalty**;
    - ✓ it means **limiting the number of features** and penalizing algorithms that are too complex or too flexible by constraining them to include only parameters that reduce out-of-sample error.
  - 2) proper data sampling achieved by using **cross-validation**, a technique for estimating out-of-sample error directly by determining the error in validation samples.

2. Supervised Learning Algorithms

- **Supervised machine learning** models are trained using labeled data and depending on the nature of the target (Y) variable, they can be divided into two types: regression for a continuous target variable and classification for a categorical or ordinal target variable.
  - **Penalized regression**
  - **Support vector machine (SVM)**
  - **K-nearest neighbor (KNN)**
  - **Classification and regression tree (CART) algorithms**
  - **Ensemble and Random forest**

2.1 Penalized Regression

- **Penalized regressions.**
  - Reduce the problem of **overfitting** by imposing a **penalty term**.
    - ✓ The number of features increases, penalty term increases.
  - E.G. Least absolute shrinkage and selection operator (LASSO). LASSO can be used to build parsimonious models.

penalty term, $\lambda > 0$

$$\sum_{i=1}^{n}(Yi - \widehat{Y}_i)^2 + \lambda \sum_{k=1}^{K} |\hat{b}_k|$$

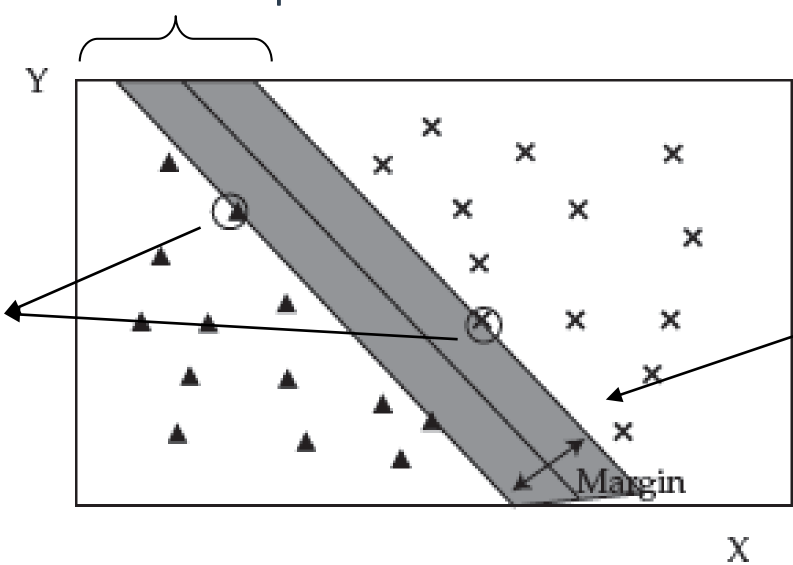
    - ✓ Lambda (λ) is a hyperparameter
      - ◆ balance between fitting the model versus keeping the model parsimonious.
      - ◆ Note: λ = 0, LASSO penalized regression = OLS regression.
- **Regularization describes methods that reduce statistical variability in high dimensional data estimation problems.**
  - Regularization can be applied to non-linear models.

2.2 SVM

- **Support vector machine (SVM)** is a **linear classifier** that determines the **hyperplane** that optimally separates the observations into two possible classifiers (e.g., sell vs. buy, default and non-default).
- SVM **maximizes the probability of making a correct prediction** by determining the boundary that is farthest away from all the observations.
  - SVM separates the data by the **maximum margin.**

shaded strip

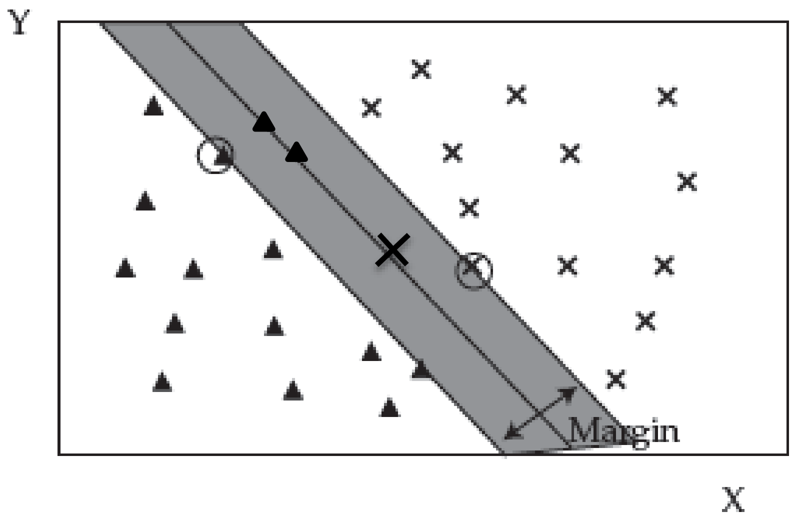
support vectors, observations that are closest to the boundary



discriminant boundary

SVM

- Many real-world data sets, however, are **not perfectly linearly separable**, in that case**, soft margin classification** is applied.

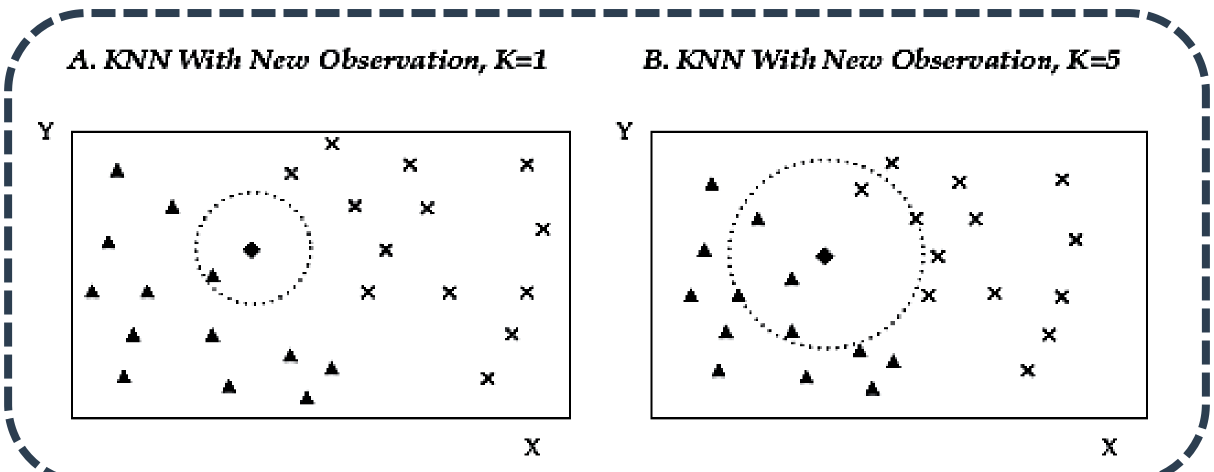


This adaptation **adds a penalty** to the objective function for observations in the training set that are misclassified, it ***optimizes the tradeoff between a wider margin and classification error.***

- As an alternative to soft margin classification, a **non-linear SVM** algorithm can be run by introducing more advanced, non-linear separation boundaries.

2.3 K-Nearest Neighbor

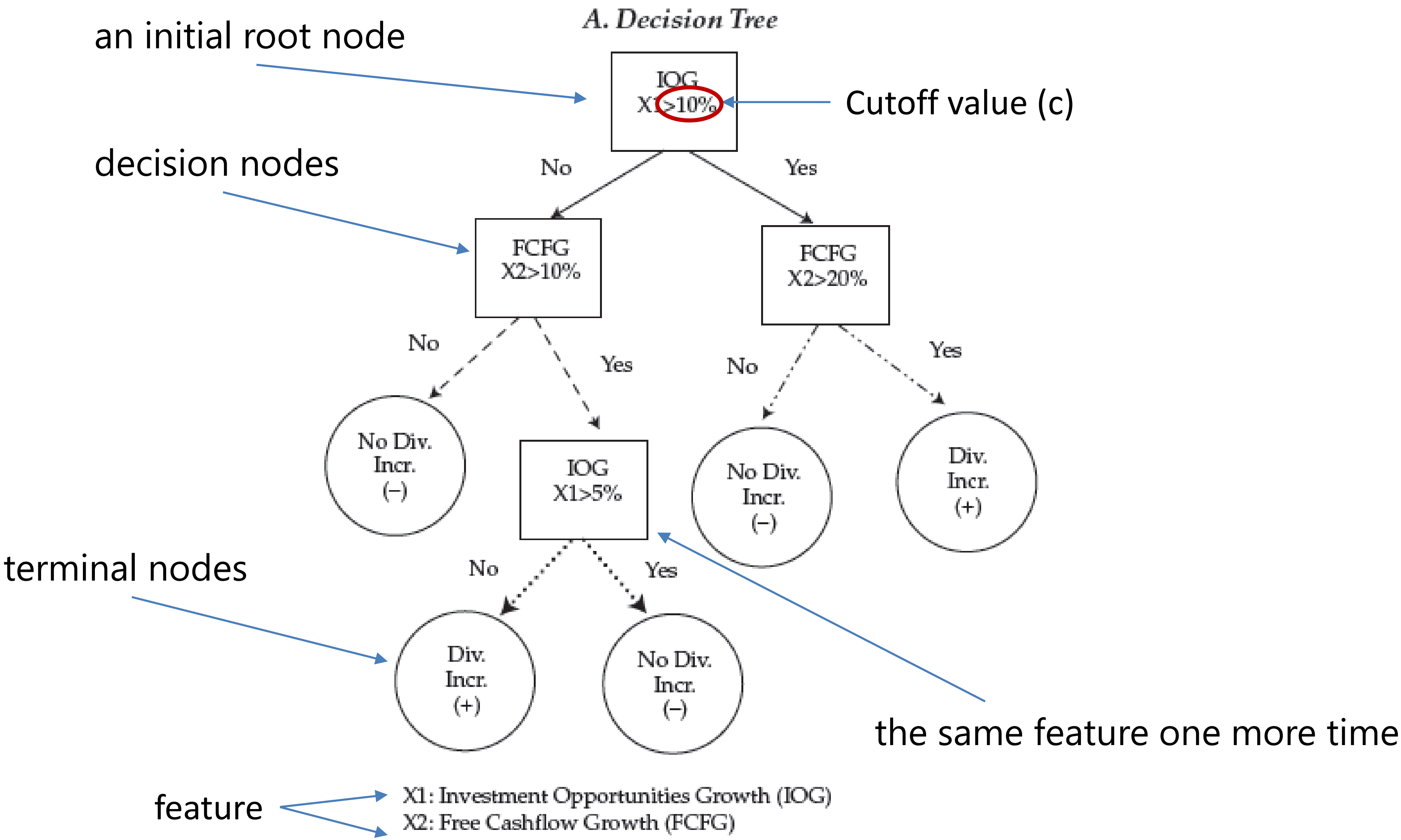
- **K-nearest neighbor (KNN).** More commonly used in **classification** (but sometimes in regression), this technique is used to classify a new observation by **finding similarities ("nearness")** between this new observation and the training sample.
- **Two vital concerns**
  - A critical challenge of KNN, however, is defining what it means to be **"similar" (or near)**.
  - The researcher specifies the **value of k**, the hyperparameter, triggering the algorithm to look for the k observations in the sample that are closest to the new observation that is being classified.



2.4 Classification and Regression Tree

- **Classification and regression trees (CART).**
  - **Classification trees** are appropriate when the target variable is **categorical**.
    - ✓ typically used when the target is binary. (e.g., an IPO will be successful vs. not successful.)
  - **Regression trees** are appropriate when the target is **continuous**.
    - ✓ If the goal is regression, then the prediction at each terminal node is the mean of the labeled values.
- CART can be used when there are significant **nonlinear relationships** among variables.
- **To avoid overfitting**,
  - **regularization criteria** such as maximum tree depth, maximum number of decision nodes, and so on are specified by the researcher.
  - Alternatively, sections of tree with minimal explanatory power are **pruned**.

Classification and Regression Tree



2.5 Ensemble learning

- **Ensemble learning:** combining predictions from multiple models.
  - The ensemble method results in a lower average error rate because the different models cancel out noise.
  - **Two kinds of ensemble methods**
    - ✓ Under **aggregation of heterogeneous learners**, **different algorithms** are combined together via a voting classifier.
    - ✓ Under **aggregation of homogenous learners**, the **same algorithm** is used, but on **different training data** sourcing from:
      - **Bootstrap aggregating (or bagging)**. The process relies on generating random samples (bags) with replacement from the initial training sample.

Random Forest

- **Random forest** is a variant of classification trees whereby a large number of classification trees are trained using data bagged from the same data set.
  - A randomly selected **subset** of **features** is used in creating each tree, and each tree is slightly different from the others. Because each tree only uses a subset of features, random forests can mitigate the problem of overfitting.
  - The process of using multiple classification trees to determine the final classification is akin to the practice of "wisdom of crowd".
- **Investment applications** of random forest include factor-based asset allocation, and prediction models for the success of an IPO.

Voting Classifiers

- Suppose you have been working on a machine learning project for some time and have trained and compared the results of several algorithms, such as SVM, KNN, and CART. A **majority-vote classifier** will assign to a new data point the predicted label with the most votes.
  - For example, if the SVM and KNN models are both predicting the category "stock outperformance" and the CART model is predicting the category "stock underperformance," then the majority-vote classifier will choose '"stock outperformance."
  - The **more individual models** you have trained, the **higher the accuracy** of the aggregated prediction up to a point.

3. Unsupervised Learning Algorithms

- **Unsupervised learning** is machine learning that does not use labeled data (i.e., no target variable); thus, the algorithms are tasked with finding patterns within the data themselves.
- The two main types are
  - **Dimension reduction**
    - ✓ principal components analysis.
  - **Clustering**
    - ✓ k-means clustering;
    - ✓ hierarchical clustering.

3.1 Principal Component Analysis

- **Dimension reduction.** Problems associated with too much noise often arise when the number of features in a data set (i.e., its dimension) is excessive.
  - **Principal components analysis (PCA).**
    - ✓ PCA is used to summarize or reduce highly correlated features of data into a few main, uncorrelated **composite variables**.
    - ✓ Eigenvectors (composite variable): define new, mutually uncorrelated composite variables that are linear combinations of the original features.
    - ✓ Eigenvalue (类似R^2): the proportion of total variance in the initial data explained by each eigenvector.
  - In practice, the smallest number of principal components that collectively capture 85%—95% of the total variance are retained.

Principal Component Analysis

- The **main drawback** of PCA is that since the principal components are combinations of the data set's initial features, they typically cannot be easily labeled or directly interpreted by the analyst. Compared to modelling data with variables that represent well-defined concepts, the end user of PCA may perceive PCA as something of a "**black box**."

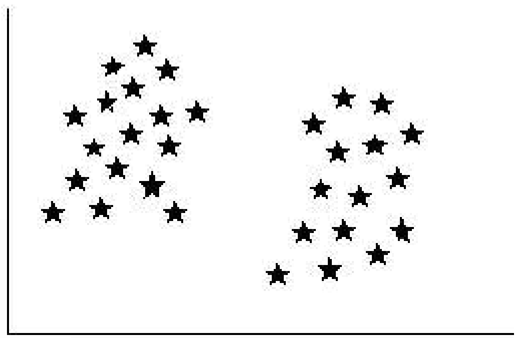
3.2 Clustering

- **Clustering.** Given a data set, clustering is the process of grouping observations into categories based on **similarities** in their attributes.
  - In practice, human judgment plays a role in defining what is similar.
    - **Euclidian distance**, the straight line distance between two observations, is one common metric that is used.
    - The **smaller the distance**, the **more similar the observations;** the larger the distance, the more dissimilar the observations.
- **Common types of clustering:**
  - k-means clustering
  - hierarchical clustering.

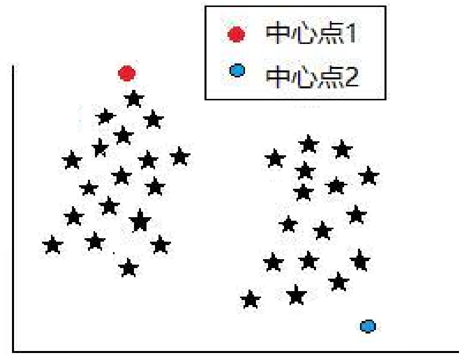
K-Means Clustering

- **K-means clustering** partitions observations into k non-overlapping clusters, where **k** is a **hyperparameter**.
  - Each cluster has a **centroid** (the center of the cluster), and each new observation is assigned to a cluster based on its proximity to the centroid.
  - One limitation of this type of algorithm is that the hyperparameter k is chosen before clustering starts, meaning that one has to have some idea about the nature of the data set.
- K-means clustering is used in investment management to classify thousands of securities based on patterns in high **dimensional data**.

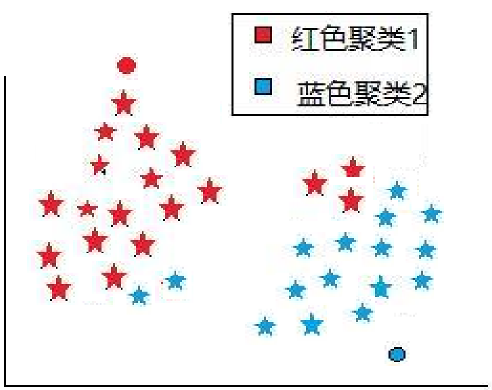
K-Means Clustering



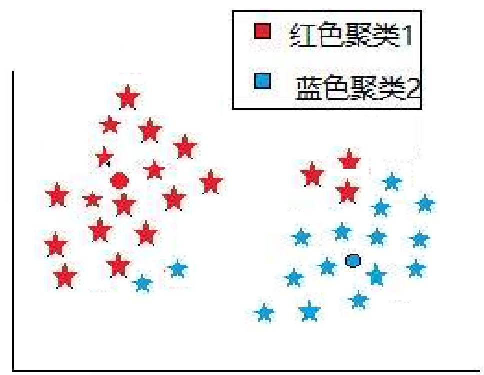
1）样本分布



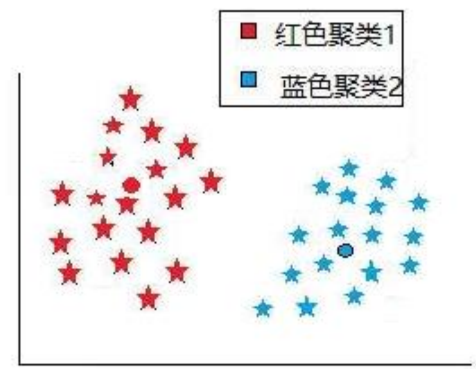
2）随机选取K个中心点（K=2）



3）计算每个样本和中心点的距离



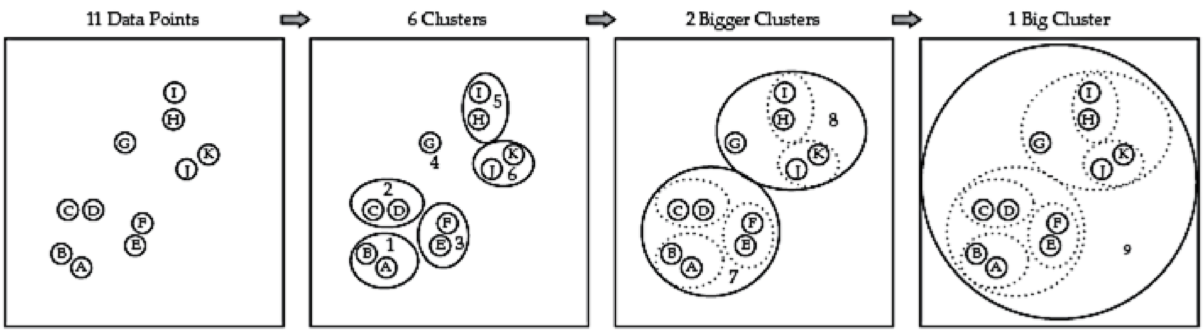
4）中心点移动到类中心



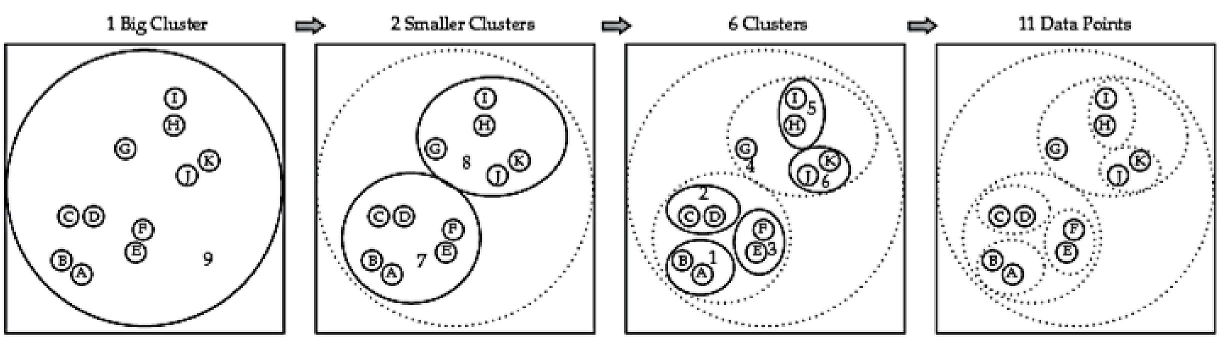
5）最终结果

Hierarchical clustering

- **Hierarchical clustering** builds a hierarchy of clusters **without any predefined number of clusters.**
  - ✓ **An agglomerative** (or bottom-up) **clustering**

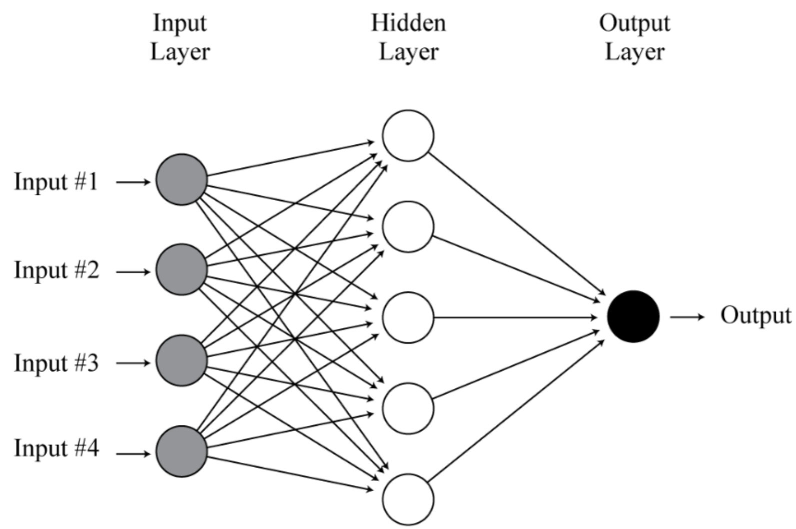


  - ✓ A **divisive** (or top-down) **clustering**



4. Neural Networks

- Neural networks have three types of layers
  - an input layer
  - hidden layers
  - an output layer



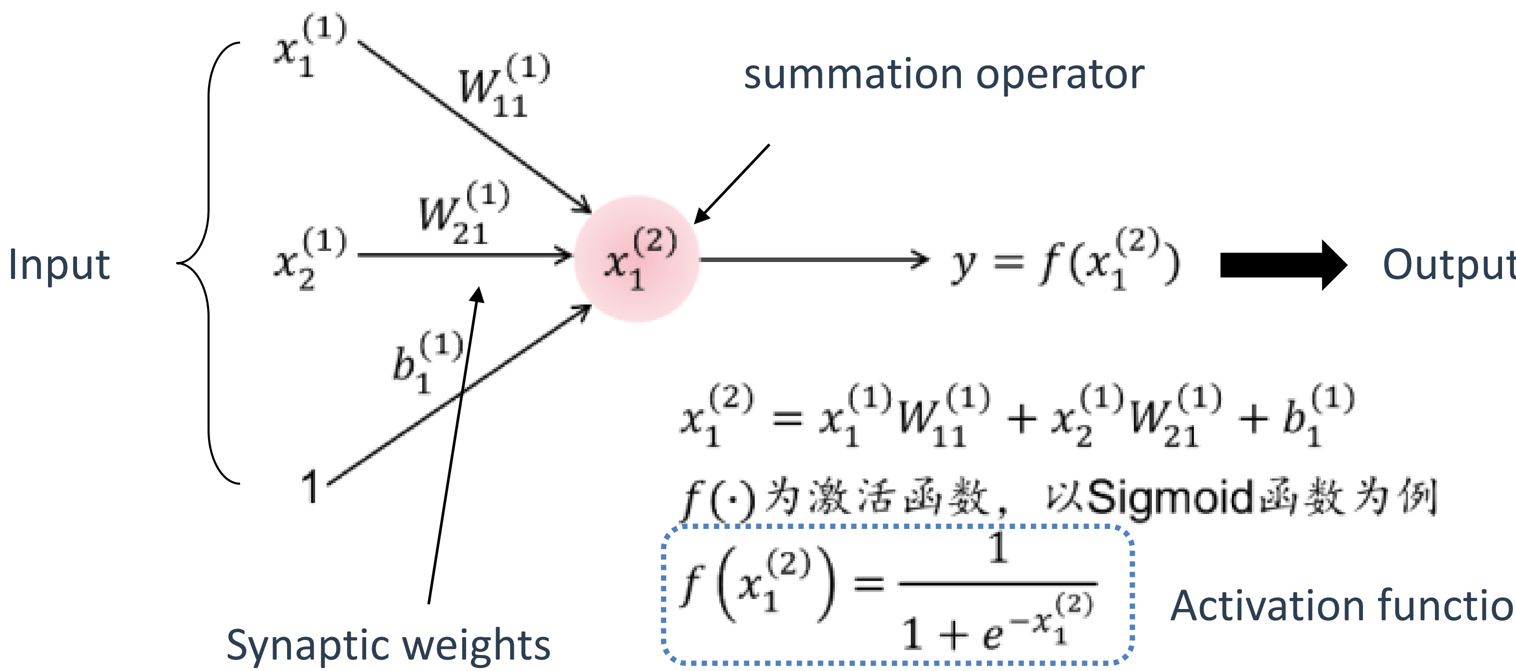
- For example, a neutral network with **hyperparameters of 4-5-1**
  - **four** nodes in input layer(four features)
  - **five** nodes in the single hidden layer(five ways of transmitting data)
  - **one** node in output layer(one predict result)

Neural Networks

- Each node has, conceptually, two functional parts: a **summation operator** and an **activation function**.
  - Once the node receives the four input values, the **summation operator** multiplies each value by a weight and sums the weighted values to form the total net input.
  - The total net input is then passed to the **activation function**, which transforms this input into the final output of the node.
    - ✓ Informally, the **activation function** operates like a **light dimmer switch** that decreases or increases the strength of the input.
    - ✓ The activation function is characteristically non-linear, such as an S-shaped (sigmoidal) function (with output range of 0 to 1) or the rectified linear unit function.

Neural Networks

- **Neurons modeling（forward propagation)**



$$x_1^{(2)} = x_1^{(1)} W_{11}^{(1)} + x_2^{(1)} W_{21}^{(1)} + b_1^{(1)}$$

$f(\cdot)$为激活函数，以Sigmoid函数为例

$$\boxed{f\left(x_1^{(2)}\right) = \frac{1}{1 + e^{-x_1^{(2)}}}}$$

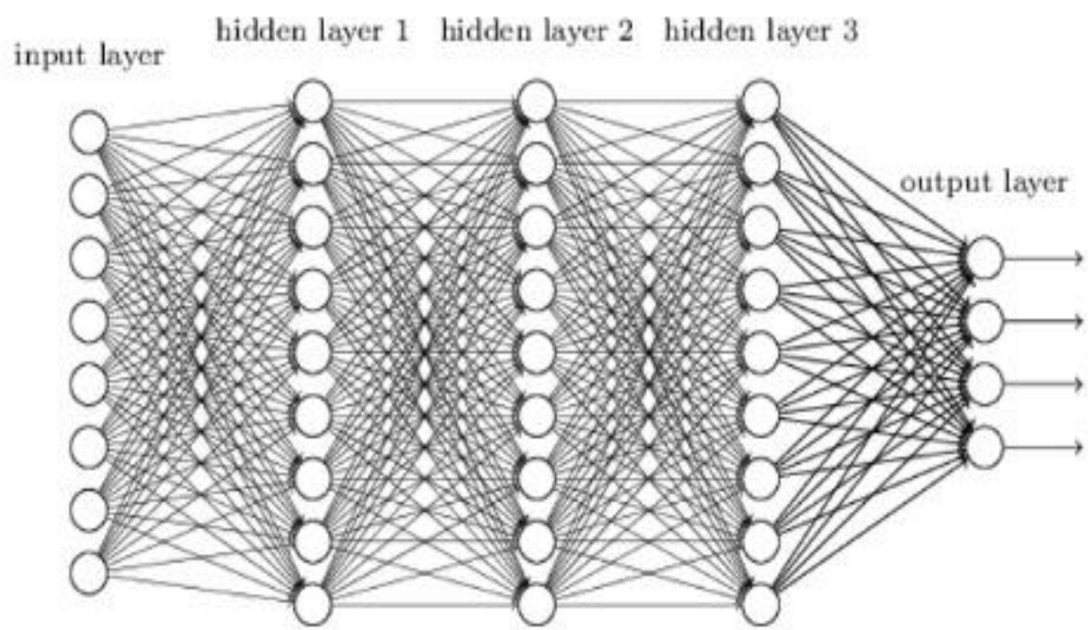
Activation function

Neural Networks

- **Backward propagation**
  - A related process, backward propagation, is employed to revise the weights used in the summation operator as the network learns from its errors.
- **Revision of hyperparameters**
  - Hyperparameters may be revised based on the out-of-sample performance of the model.

Deep Learning Nets

- Neural networks with many hidden layers (at least 3 but often more than 20) – known as **deep learning nets (DLNs)** – are the backbone of the intelligence revolution.
  - ✓ DLNs have been shown to be useful in general for image, pattern and speech recognition problems.



Reinforcement Learning

- **Reinforcement learning (RL) algorithms** have an agent that seeks to **maximize a defined reward** given defined constraints.
  - The RL agent does not rely on labeled training data, but rather learns based on immediate feedback from **(millions of) trials and errors**.
  - When applied to the ancient game of Go, DeepMind's AlphaGo algorithm was able to beat the reigning world champion.

Learning Module 7

Big Data Projects

Framework

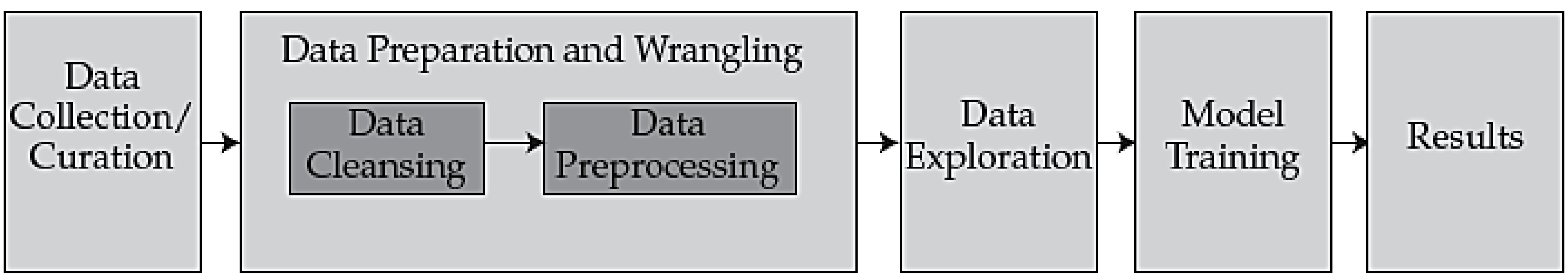
1. Structured Data Analysis
   - Conceptualization of the modeling task
   - Data collection
   - Data preparation and wrangling
   - Data exploration
   - Model training
2. Unstructured Data Analysis
   - Text problem formulation
   - Data (text) curation
   - Text preparation and wrangling
   - Text exploration
   - Model training

1. Structured Data Analysis

- To illustrate the steps involved in analyzing data for financial forecasting (traditional ML model), we will use an example of a consumer credit scoring model in the following **five** steps:
  1. **Conceptualization of the modeling task**
  2. **Data collection**
  3. **Data preparation and wrangling**
  4. **Data exploration**
  5. **Model training**

1.1 Data Preparation and Wrangling

3. **Data preparation and wrangling** involves **cleansing** the data set and **organizing** raw data into a consolidated format.
   - **Data preparation (Cleansing)** is the process of examining, identifying, and mitigating errors in raw data, includes addressing any missing values or verification of any out-of-range values.
   - **Data Wrangling (Preprocessing)** data may performs transformations and critical processing steps on the cleansed data to make the data ready for ML model training, involving aggregating, filtering, or extracting relevant variables.



Data Preparation (Cleansing)

- The possible errors in a raw dataset include the following:
  - **Incompleteness error** is where the data are not present, resulting in missing data.
  - **Invalidity error** is where the data are outside of a meaningful range, resulting in invalid data.
  - **Inaccuracy error** is where the data are not a measure of true value.
  - **Inconsistency error** is where the data conflict with the corresponding data points or reality.
  - **Non-uniformity error** is where the data are not present in an identical format.
  - **Duplication error** is where duplicate observations are present.

Data Preparation (Cleansing)

Inaccuracy error

Invalidity error

Non-uniformity error

Inconsistency error

1	ID	Name	Gender	Date of Birth	Salary	Income	State	Credit Card
2	1	Mr. ABC	M	12/5/1970	$50,200	$5,000	VA	Y
3	2	Ms. XYZ	M	15 Jan, 1975	$60,500	$0	NY	Yes
4	3	EFG		1/13/1979	$65,000	$1,000	CA	No
5	4	Ms. MNO	F	1/1/1900	—	—	FL	Don't Know
6	5	Ms. XYZ	F	15/1/1975	$60,500	$0		Y
7	6	Mr. GHI	M	9/10/1942	NA	$55,000	TX	N
8	7	Mr. TUV	M	2/27/1956	$300,000	$50,000	CT	Y
9	8	Ms. DEF	F	4/4/1980	$55,000	$0	British Columbia	N

Incompleteness error

Duplication error

Data Wrangling (Preprocessing)

- **Before data wrangling**, as **outliers** may be present in the data, and domain knowledge is needed to deal with them.
  - ✓ Any outliers that are present **must first be identified**.
  - ✓ The outliers then should be examined and a decision made to either remove or replace them with values imputed using statistical techniques.

1	ID	Name	Gender	Date of Birth	Salary	Other Income	State	Credit Card
2	1	Mr. ABC	M	12/5/1970	USD 50200	USD 5000	VA	Y
3	2	Ms. XYZ	F	1/15/1975	USD 60500	USD 0	NY	Y
4	3	Mr. EFG	M	1/13/1979	USD 65000	USD 1000	CA	N
5	6	Mr. GHI	M	9/10/1942	$\boxed{\text{USD 0}}$	USD 55000	TX	N
6	7	Mr. TUV	M	2/27/1956	$\boxed{\text{USD 300000}}$	USD 50000	CT	Y
7	8	Ms. DEF	F	4/4/1980	CAD 55000	CAD 0	British Columbia	N

outliers

Data Wrangling (Preprocessing)

- There are several practical methods for **handling outliers**.
  - When extreme values and outliers are simply removed from the dataset, it is known as **trimming** (also called **truncation**).
    - ✓ E.G. A 5% trimmed dataset is one for which the 5% highest and the 5% lowest values have been removed.
    - ✓ E.G. The truncated average score of a diving competition.
  - When extreme values and outliers are **replaced** with the maximum (for large value outliers) and minimum (for small value outliers) values of data points that are not outliers, the process is known as **winsorization**.

Data Wrangling: Transformation

- Data preprocessing primarily includes **transformations** and **scaling** of the data.
  - **Data transformations**
    - ✓ **Extraction**: A new variable can be created from the current variable for ease of analyzing and using for training the ML model.
    - ✓ **Aggregation**: Two or more variables can be aggregated into one variable to consolidate similar variables.
    - ✓ **Filtration**: The data rows that are not needed for the project must be identified and filtered.
    - ✓ **Selection**: The data columns that are intuitively not needed for the project can be removed.
    - ✓ **Conversion**: The variables can be of different types: nominal, ordinal, continuous, and categorical.

Data Wrangling: Transformation

- Data before transformation

(4)

1	ID	Name	Gender	Date of Birth	Salary	Other Income	State	Credit Card
2	1	Mr. ABC	M	12/5/1970	USD 50200	USD 5000	VA	Y
3	2	Ms. XYZ	F	1/15/1975	USD 60500	USD 0	NY	Y
4	3	Mr. EFG	M	1/13/1979	USD 65000	USD 1000	CA	N
5	6	Mr. GHI	M	9/10/1942	USD 0	USD 55000	TX	N
6	7	Mr. TUV	M	2/27/1956	USD 300000	USD 50000	CT	Y
7	8	Ms. DEF	F	4/4/1980	CAD 55000	CAD 0	British Columbia	N

(3)

(5)

- Data after transformation

1	ID	Gender	Age	Total Income	State	Credit Card
2	1	M	48	55200	VA	Y
3	2	F	43	60500	NY	Y
4	3	M	39	66000	CA	N
5	6	M	76	55000	TX	N

(1)

(2)

Data Wrangling: Scaling

- **Scaling** is a process of **adjusting the range of a feature** by shifting and changing the scale of data.
- Here are two of the most common ways of scaling:
  - **Normalization** is the process of rescaling numeric variables in the range of [0, 1].

$$X_{i(\mathrm{normalized})} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

    - ✓ **sensitive to outliers**, so treatment of outliers is necessary before normalization is performed.
    - ✓ used when the **distribution** of the data is **not known**.
  - **Standardization** is the process of both centering and scaling the variables.

$$X_{i(\mathrm{standardized})} = \frac{X_i - \mu}{\sigma}$$

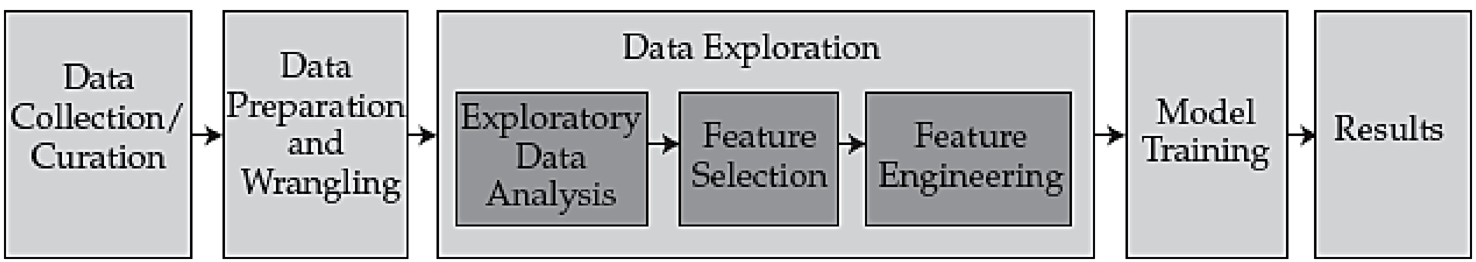
    - ✓ **less sensitive to outliers** as it depends on the mean and standard deviation of the data.
    - ✓ The data must be **normally distributed** to use standardization.

1.2 Data Exploration

4. **Data exploration**. Prepared data are explored to investigate data distribution and relationships.

    - This step involves **initial exploratory data analysis**, **feature selection** and **feature engineering**.

        ✓ In a credit scoring model, several variables may be combined to form an ability-to-pay score.



Data Exploration-EDA

- **Exploratory data analysis (EDA)** is the preliminary step in data exploration.
  - An important objective of EDA is to **help understand data** prosperities and characteristics, to **serve as a communication medium** among project stakeholders, including business users, domain experts, and analysts.
- **Tools of EDA as follow:**
  - **Summary statistics**
    - ✓ mean, variance, skewness, kurtosis, correlation matrix
  - **Visualizations**
    - ✓ histogram, box plot, heat maps, scatterplot

Data Exploration-FS

- ***After using EDA to discover relevant patterns in the data***, it is essential to **identify and remove unneeded, irrelevant, and redundant features**.
  - **Feature selection** is a process whereby **only pertinent features** from the dataset are selected for ML model training such as PCA.
    - ✓ Objective: selecting fewer features decreases ML model complexity and training time.

Data Exploration-FS

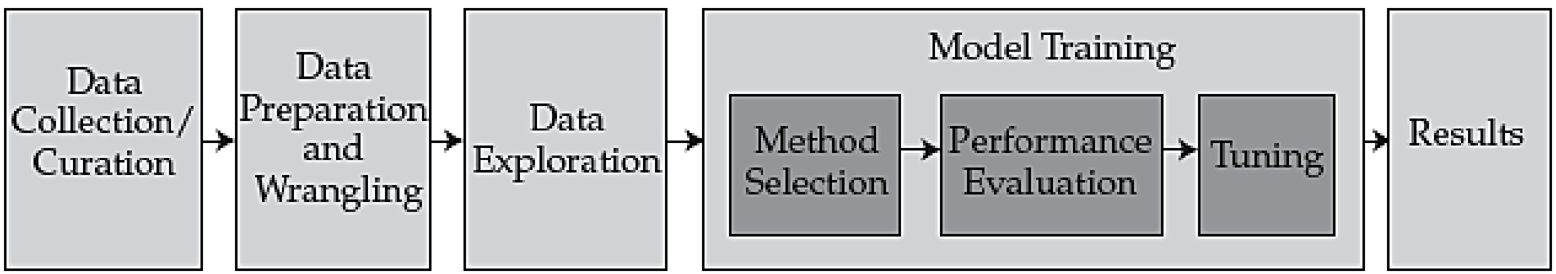
- Feature selection vs. Selection in data preprocessing steps
  - Good feature selection requires an understanding of the data and statistics, and comprehensive EDA must be performed to assist with this step.
  - Data preprocessing needs clarification only from data administrators and basic intuition.
- Feature selection vs. Dimensionality reduction
  - Both seek to reduce number of features.
  - Dimensionality reduction method creates new composite variables, which are combined by highly correlated features.
  - Feature selection excludes features that are unneeded, irrelevant and redundant without altering them.

Data Exploration-FE

- **Feature Engineering(FE)** helps further optimize and improve the features.
  - Feature engineering is a process of **creating new, more meaningful features** by changing or transforming existing features.
    - ✓ Objective: describe the structures inherent in the dataset.
    - ✓ New features can be created, made by combination of two features, or decompose one feature into many.
- Model performance heavily depends on feature selection and engineering.

1.3 Model Training

5. **Model training**. This step involves determining the appropriate ML algorithm to use, evaluating the algorithm using a training data set, and tuning the model. The choice of the model depends on the nature of the relationship between the features and the target variable.



Performance Evaluation

- It is important to measure the model training performance or goodness of fit for validation of the model.
- Several techniques to measure model performance that are well suited specifically for binary classification models:
  - Error analysis
  - Receiver Operating Characteristic
  - Root mean square error (RMSE)

Performance Evaluation

- **1) Error analysis**. For classification problems, error analysis involves computing four basic evaluation metrics: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) metrics.
  - FP is also called a Type I error, and FN is also called a Type II error.
- We can use **logistic regression** to get the predicted probability (p).
  - When target value p from a logistic regression model for a given observation is greater than the cutoff point (or threshold), here, for example is 0.5, then
    - ✓ the observation is classified as class = 1.
    - ✓ otherwise, the observation will be classified as class = 0.

Performance Evaluation

➢ **Confusion matrix:**

- Assume that Class “0” is “not defective” and,
- Class “1” is “defective.”

		Actual Training Labels	
		Class “1”	Class “0”
Predicted Results	Class “1”	True Positives (TP)	False Positives (FP) Type I Error
	Class “0”	False Negatives (FN) Type II Error	True Negatives (TN)

Performance Evaluation

- **Elements in error analysis.**
  - **Precision** is the ratio of correctly predicted positive classes to all predicted positive classes.

    $$\text{Precision (P)} = TP/(TP + FP)$$

  - **Recall** (sensitivity) is the ratio of correctly predicted positive classes to all actual positive classes.

    $$\text{Recall (R)} = TP/(TP + FN)$$

  - **Accuracy** is the percentage of correctly predicted classes out of total predictions.

    $$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN)$$

  - **F1 score** is the harmonic mean of precision and recall.

    $$\text{F1 score} = (2 * P * R)/(P + R)$$

    - ✓ F1 score is more appropriate (than accuracy) when <u>unequal class distribution</u> is in the dataset and it is necessary to measure the equilibrium of precision and recall.

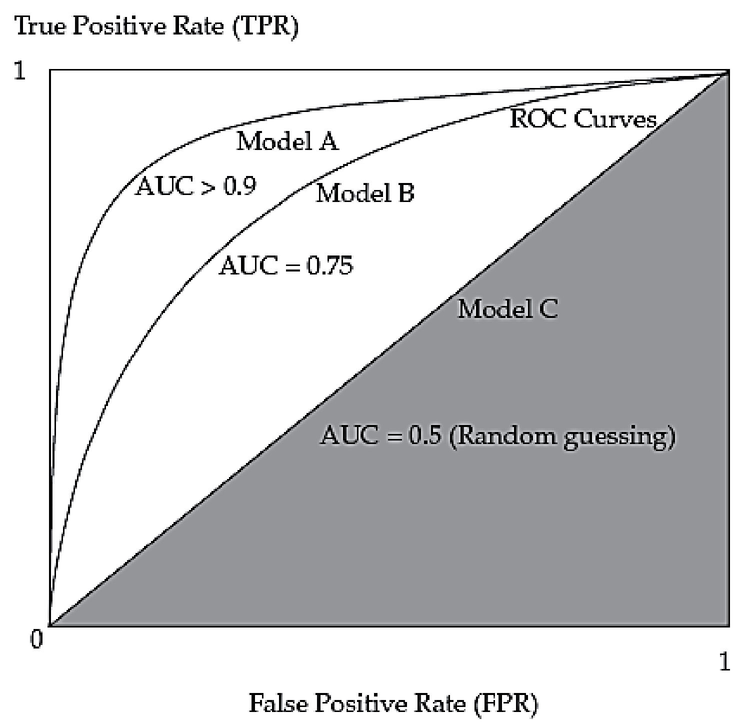
Performance Evaluation

➢ **2) Receiver Operating Characteristic (ROC).** This technique for assessing model performance involves the plot of a curve showing the trade-off between the false positive rate (x-axis) and true positive rate (y-axis) for various cutoff points.

False positive rate (FPR) = FP/(TN + FP)

True positive rate (TPR) = TP/(TP + FN)

- The **shape of the ROC curve** provides insight into the model's performance.
- A more **convex curve** indicates **better** model performance.
- Area **under the curve (AUC)** is the metric that measures the area under the ROC curve.
- An **AUC close to 1.0** indicates near **perfect** prediction, while an **AUC of 0.5** signifies **random guessing**.



Performance Evaluation

- **3) Root mean square error (RMSE).** This is useful for data predictions that are continuous, such as regression models. The RMSE is a single metric summarizing the prediction error in a sample.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - actual_i)^2}{n}}$$

Model Tuning

- It is necessary to <u>find an optimum tradeoff between bias and variance errors</u>, such that the model is neither underfitting (Bias error) nor overfitting (Variance error).
- **Parameters** are <u>estimated by the model</u> using an optimization technique on the training sample.
- **Hyperparameters** are <u>specified by ML engineers</u>, and are independent of the training sample.
- **Tuning** involves altering the hyperparameters until a desirable level of model performance is achieved.
  - For each specification of hyperparameter(s), a **confusion matrix** is prepared.
  - If there are multiple hyperparameters in the model, one can use a **grid search**.
    - ✓ A grid search is an automated process of selecting the best combination of hyperparameters.

2. Unstructured Data Analysis

➢ Unstructured, texted-based data is more suitable for human use. The five steps involved need to be modified(the first four) in order to analyze unstructured, text-based data:

1. **Text problem formulation**. The analyst will determine the problem and identify the exact inputs and output of the model.
2. **Data collection (curation)**. This is determining the sources of data to be used (e.g., web scouring, specific social media sites).
3. **Text preparation and wrangling**. This requires preprocessing the stream(s) of unstructured data to make it usable by traditional structured modeling methods.
   - ✓ Unstructured data can be in the form of text, images, videos, and audio files.
4. **Text exploration**. This involves test visualization as well as text feature selection and engineering.
5. **Model training**

2.1 Text Preparation (Cleansing)

- Text cleansing involves the following steps:

1. **Remove HTML tags.** Text collected from web pages has embedded HTML tags, which may need to be removed before processing.
2. **Remove punctuations.** Text analysis usually does not need punctuations, so these need to be removed as well. Some punctuations (e.g., %, $) may be needed for analysis, and if so, they are replaced with annotations (i.e., dollarSign, percentSign) for model training.
3. **Remove numbers.** When numbers (or digits) are present in the text, they should be removed or substituted with an annotation /number/.
4. **Remove white spaces.** Extra formatting-related white spaces (e.g., tabs, indents) do not serve any purpose in text processing and are removed.

Text Preparation (Cleansing)

Original text from a financial statement as shown on a webpage

CapEx on the normal operations remained stable on historicallylow levels, $800,000 compared to $1.2 million last year.

Quarter 3, so far, is 5% sales growth quarter-to-date, and year-to-date, we have a 4% local currency sales development.

Raw text after scraping from the source

<p> CapEx on the normal operations remained stable on historically low levels, $800,000 compared to $1.2 million last year. <b/><b/> Quarter 3, so far, is 5% sales growth quarter-to-date, and year-to-date, we have a 4% local currency sales development </p>

(1)

Text after removing html tags

CapEx on the normal operations remained stable on historically low levels, $800,000 compared to $1.2 million last year.
Quarter 3, so far, is 5% sales growth quarter-to-date, and year-to-date, we have a 4% local currency sales development.

(2)

Text after removing and replacing punctuations

CapEx on the normal operations remained stable on historically low levels /dollarSign/ /800000 compared to /dollarSign/12 million last year /endSentence/ Quarter 3 so far is 5 /percentSign/ sales growth quarter-to-date and year-to-date we have a 4 /percentSign/ local currency sales development /endSentence/

(3)

Text after replacing numbers

CapEx on the normal operations remained stable on historically low levels /dollarSign/ /number / compared to/dollarSign/ /number/ million last year /endSentence/ Quarter/ number/ so far is /number/ /percentSign/sales growth quarter-to-date and year-to-date we have a /number/ / percentSign/ local currency sales development /endSentence/

(4)

Text after removing extra white spaces

CapEx on the normal operations remained stable on historically low levels/dollarSign//number /compared to/dollarSign//number/million last year/endSentence/ Quarter/number/so far is /number//percentSign/sales growth quarter-to-date and year-to-date we have a/number// percentSign/local currency sales development/endSentence/

Text Wrangling (Preprocessing)

- To begin with text processing, tokens and tokenization need to be defined.
  - A **token** is equivalent to a word 断句拆词
  - **Tokenization** is the process of splitting a given text into separate tokens
    - ✓ Tokenization can be performed at word or character level, but it is most commonly performed at word level.

	Cleaned Texts	Tokens
Text 1	The man went to the market today	The / man / went / to / the / market / today
Text 2	Market values are increasing	Market / values / are / increasing
Text 3	Increased marketing is needed	Increased / marketing / is / needed
Text 4	There is no market for the product	There / is / no / market / for / the / product

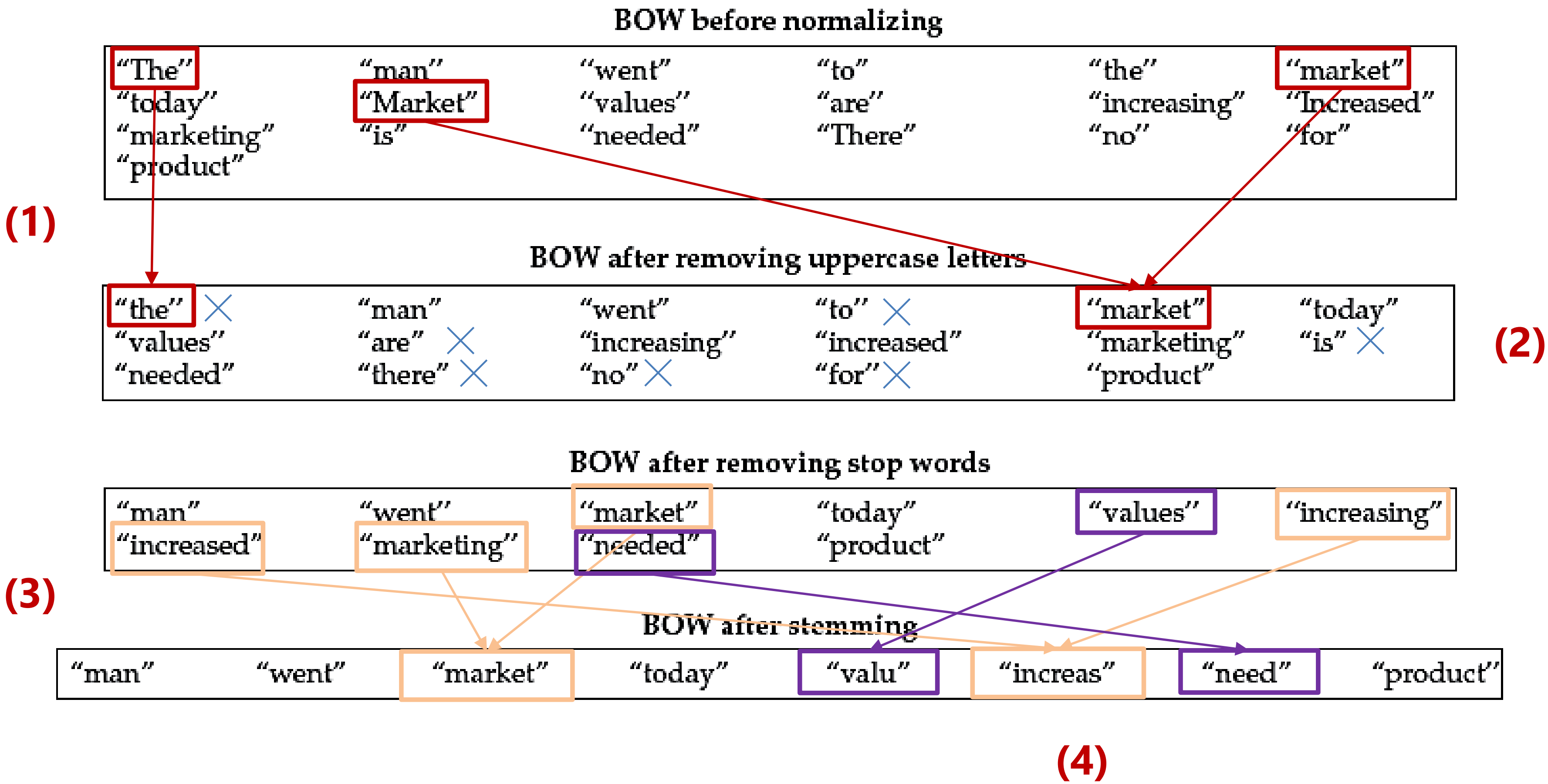
Text Wrangling (Preprocessing)

- **Step 1** in text preprocessing: **normalization**

1. **Lowercasing.** So as to not discriminate between “market” and “Market”.
2. **Removal of stop words.** Stop words are such commonly used words as “the,” “is,” and “a.” Stop words do not carry a semantic meaning for the purpose of text analyses and ML training.
3. **Stemming.** This is a rules-based algorithm that converts all variations of a word into a common value. For example, integrate, integration, and integrating are all assigned a common value of integrat.
4. **Lemmatization.** This involves the conversion of inflected forms of a word into its lemma (i.e., morphological root).
    - Lemmatization is similar to stemming but is more computationally advanced and resource intensive. It is an algorithmic approach and depends on the knowledge of the word and language structure.

Text Wrangling (Preprocessing)

- **Step 2** in text preprocessing: **bag-of-words (BOW)** procedure
  - It simply collects all the words or tokens without regard to the sequence of occurrence.



Text Wrangling (Preprocessing)

- If the sequence of text is important, **N-grams** can be used to represent word sequences.
  - ✓ **Terminology**: A two-word sequence is a **bigram**, a three-word sequence is **trigram**, and so forth.
  - ✓ Consider the sentence, "The market is up today."
    - ◆ Bigrams of this sentence include "the_market," "market_is," "is__up," and "up_today." BOW is then applied to the bigrams instead of the original words.
    - ◆ N-gram implementation will affect the normalization of the BOW because stop words will not be removed.

Text Wrangling (Preprocessing)

- **Step 3** in text preprocessing: build a **document term matrix (DTM)**.
  - In this matrix, each text document is a row, and the columns are represented by tokens. The **cell value** represents **the number of occurrences** of a token in a document (i.e., row).

DTM

	man	went	market	today	valu	increas	need	product
Text 1	1	1	1	1	0	0	0	0
Text 2	0	0	1	0	1	1	0	0
Text 3	0	0	1	0	0	1	1	0
Text 4	0	0	1	0	0	0	0	1

2.2 Text Exploration - EDA

- Various text statistics are used to explore, summarize and analyze text data.
  - **Term frequency**: number of times the word appears in the text.
- **Visualization** such as **word cloud** can be applied.

Text Exploration – Feature Selection

- **Feature selection** involves selecting a subset of tokens in the BOW.
  - Reduction in BOW size makes the model more **parsimonious** and **reduces feature-induced noise**.
    - ✓ High- and low-frequency words (noise) are often eliminated, resulting in a more concise BOW.
      - High-frequency words tend to be stop words (if not removed during the data wrangling phase) or common vocabulary words.
      - Low-frequency words may be irrelevant.

Feature Selection

- **Feature selection methods** include:
  - **1) Frequency** measures can be used for vocabulary pruning to remove noise features by filtering the tokens with very high and low TF values across all the texts.
  - **2) Chi-square.** This test is used to rank tokens by their usefulness to each class in text classification problems.
    - Tokens with the **highest chi-square test statistic values occur more frequently** in texts associated with a particular class and therefore can be selected for use as features for ML model training
  - **3) Mutual information (MI)** measures how much information is contributed by a token to a class of texts.
    - ✓ If the token appears **in all classes**, it is **not considered a useful** discriminant, and its MI equals **0**.
    - ✓ Tokens associated with **only one or a few classes** would have MI approaching **1**.

Feature Engineering

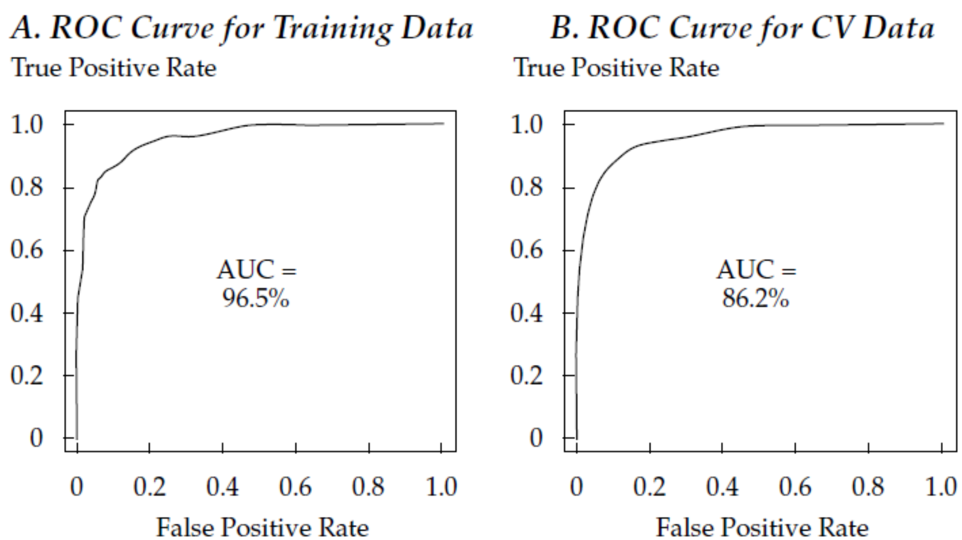
- **Techniques of FE include**:
  - **1) Numbers.** Tokens with standard lengths are identified and converted into a token such as /numberX/.
    - ✓ Four-digit numbers may be associated with years and are assigned a value of /number4/.
  - **2) N-grams.** Multi-word patterns that are particularly discriminative can be identified and their connection kept intact.
  - **3) Name entity recognition (NER).** NER algorithms search for token values, in the context it was used, against their internal library and assign a NER tag to the token.
    - ✓ For example, Microsoft would be assigned a NER tag of ORG and Europe would be assigned a NER tag of Place. NER object class assignment is meant to make the selected features more discriminatory.
  - **4) Parts of speech (POS).** This uses language structure dictionaries to contextually assign tags (POS) to text.
    - ✓ For example, Microsoft would be assigned a POS tag of NNP (indicating a proper noun), and the year 1969 would be assigned a POS tag of CD (indicating a cardinal number).

Model Training

- **Logistic regression is applied on the final training DTM for model training.**
  - As a binary classification model, it uses maximum likelihood estimation, output from the logistic model is a probability value ranging from 0 to 1.
  - We can use machine learning model to analyze text information sentiment into positive or negative.
    - ✓ y = 1 for sentences having positive sentiment.
    - ✓ y = 0 for sentences having negative sentiment.
  - If p value for a sentence is 0.90, there is a 90% likelihood that the sentence has positive sentiment. Theoretically, the sentences with $p > 0.50$ likely have positive sentiment.
- The threshold value is a cutoff point for p values, and the ideal threshold p value is influenced by the dataset and model training.
  - <u>p value resulting in the highest model accuracy is selected as the ideal threshold p value.</u>

Model Training

- After using Training sample to develop the model, we use validation sample and test sample for tuning and evaluating the model.



  - The AUC is 96.5% on training data and 86.2% on cross-validation data.
- As the model is overfitted, least absolute shrinkage and selection operator (LASSO) regularization is applied to the logistic regression.

问题反馈

- 如果您认为金程**课程讲义/题库/视频**或其他资料中**存在错误，欢迎您告诉我们**，所有提交的内容我们会在最快时间内核查并给与答复。
- **如何告诉我们?**
  - 将您发现的问题通过电子邮件告知我们，具体的内容包含：
    - ✓ 您的姓名或网校账号
    - ✓ 所在班级
    - ✓ 问题所在科目（若未知科目，请提供章节、知识点）和页码
    - ✓ 您对问题的详细描述和您的见解
  - 请发送电子邮件至：academic.support@gfedu.net
- **非常感谢您对金程教育的支持，您的每一次反馈都是我们成长的动力。**后续我们也将开通其他问题反馈渠道（如微信等）。