Economic Data Analysis Using R

- Introduction to R
- Getting Started - Using Rstudio IDE
- Economic Data
- Data Visualization – Using Graphs
- Data Analysis I
- Data Analysis II

Data Analysis I

- Descriptive Statistics
- Correlation and Covariance
- Analysis of Variances (AOV, ANOVA)
  – Using contingency tables
  – AOV with one category variable
  – AOV with two category variables

Data Analysis I

- Hypothesis Testing
  – DGP ~ Non IID
  – One-Variable Testing (t-test)
  – Two-Variable Testing (paired t-test)

Economic Data Analysis Using R

- Introduction to R
- Getting Started - Using Rstudio IDE
- Economic Data
- Data Visualization – Using Graphs
- Data Analysis I
- Data Analysis II

Data Analysis II

- Cross Sections Data
  - Hypothesis Testing
    - Homoscedasticity
    - Normality
  - Linear Regression
    - Ordinary Least Squares
    - Quantile Regression
      - Least Absolute Deviation
    - Maximum Likelihood

Data Analysis II

- Time Series Data
  - Hypothesis Testing
    - Durbin-Watson
    - Box-Pierce / Ljung-Box
    - ACF/PACF
  - Transformation: Lag, Difference
  - ARIMA Model
- Panel Data
  - Multilevel Analysis (lml4)

Data Analysis II

- Based on An Introduction to Statistical Learning with R (by James, G., Witten, D., Hastie, T., Tibshirani, R.) [Check here]

Data Analysis II

- Regression (ISLR Chapter 3)
- Classification (ISLR Chapter 4)
- Cross Validation (ISLR Chapter 5)
- Model Selection (ISLR Chapter 6)
- Nonlinear Models (ISLR Chapter 7)

Data Analysis II

- Regression (ISLR Chapter 3)
  - Linear Regression
  - Extensions
    - Including Qualitative Variables
    - Including Polynomials and Interactions
  - Model Selection
    - Selection Criteria: C_p, AIC, BIC, Adj-R^2, CV
    - Forward/Backward Selection

Data Analysis II

- Classification (ISLR Chapter 4)
  - Logistic Regression
    - Logit and Probit
  - Bayes Theorem for Classification
  - Discriminant Analysis
    - Linear Discriminant Analysis
    - Quadratic Discriminant Analysis

Data Analysis II

- Cross Validation (ISLR Chapter 5)
  - Resampling Methods
  - Cross Validation
  - Bootstrapping

Data Analysis II

- Model Selection (ISLR Chapter 6)
  - Stepwise Regression
  - Ridge Regression
  - LASSO
  - PCA: Principal Components Analysis

Discriminant Analysis

- Based on Bayes' Theorem

$$\Pr(Y \mid X) = \frac{\Pr(X \mid Y)\Pr(Y)}{\Pr(X)}, \quad where\, Y = k\ (class), X = x$$

$$Let\ \Pr(Y = k) = \pi_k = prior\ \ probability$$

$$\Pr(X = x \mid Y = k) = f_k(x; \mu_k, \sigma_k^2) = normal\ density$$

$$= \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(x - \mu_k)^2}{2\sigma_k^2} \right]$$

$$Since\ \Pr(X = x) = \sum_l \pi_l f_l(x; \mu_l, \sigma_l^2),$$

$$\Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x; \mu_k, \sigma_k^2)}{\sum_l \pi_l f_l(x; \mu_l, \sigma_l^2)} = p_k(x)$$

Discriminant Analysis

- Discriminant Function

$$\textit{Comparing class } k = 1,2 \textit{ with } p_k(x) = \frac{\pi_k f_k(x;\mu_k,\sigma_k^2)}{\pi_1 f_1(x;\mu_1,\sigma_1^2) + \pi_2 f_2(x;\mu_2,\sigma_2^2)}$$

$$\Leftrightarrow \pi_1 f_1(x;\mu_1,\sigma_1^2) \quad vs. \quad \pi_2 f_2(x;\mu_2,\sigma_2^2)$$

$$\Leftrightarrow \log(\pi_1) + \log(f_1(x;\mu_1,\sigma_1^2)) \quad vs. \quad \log(\pi_2) + \log(f_2(x;\mu_2,\sigma_2^2))$$

$$\textit{where } \log(f_k(x;\mu_k,\sigma_k^2)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_k^2) - \frac{(x-\mu_k)^2}{2\sigma_k^2}$$

$$\textit{Define the discriminant function}: \delta_k(x) = \log(\pi_k) - \frac{1}{2}\log(\sigma_k^2) - \frac{(x-\mu_k)^2}{2\sigma_k^2}$$

$$\textit{Comparing class } k = 1,2 \textit{ with } \delta_k(x)$$

$$\Leftrightarrow \delta_1(x) \quad vs. \quad \delta_2(x) \quad (\delta_k(x) \textit{ is quadratic in } x)$$

Discriminant Analysis

- Linear Discriminant Analysis

$$\textit{Assume } \sigma_k^2 = \sigma^2 \ \forall k = 1, 2, \textit{ we have}$$

$$\log(f_k(x; \mu_k, \sigma^2)) = -\frac{1}{2}\left[\log(2\pi) + \log(\sigma^2) + \frac{x^2}{\sigma^2}\right] + x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

Then the discriminant function is linear in x:

$$\delta_k(x) = \log(\pi_k) + x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{\sigma^2}$$

Comparing class $k = 1, 2$ *with* $\delta_k(x)$:

$$\log(\pi_1) + x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{\sigma^2} \quad vs. \quad \log(\pi_2) + x\frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{\sigma^2}$$

Discriminant Analysis

- Linear Discriminant Analysis

(μ_k, σ^2) *can be estimated from* $X \Rightarrow (\hat{\mu}_k, \hat{\sigma}^2)$

π_k *is the prior probability of* $Y \Rightarrow \hat{\pi}_k$

Compute $\hat{\delta}_k(x) = \log(\hat{\pi}_k) + x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{\hat{\sigma}^2}\ \forall k = 1, 2,$ *and compare.*

From $\hat{\delta}_k(x)$, *the estimated* $\Pr(Y = k \mid X = x) = \dfrac{\exp(\hat{\delta}_k(x))}{\sum_l \exp(\hat{\delta}_l(x))}$

  – Bayes classifier assigns an observation X=x to the class Y=k for which the discriminant function is largest.

Discriminant Analysis

- Quadratic Discriminant Analysis
  - Without equal variance assumption, we have

$$\delta_k(x) = \log(\pi_k) - \frac{1}{2}\log(\sigma_k^2) - \frac{(x-\mu_k)^2}{2\sigma_k^2}$$

Discriminant Analysis

- Generalization to Multivariate Case
  - Assumes $X \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, the QDA classifier for X=x

$$\delta_k(\mathbf{x}) = \log(\pi_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)$$

  - When $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, we have the LDA classifier for X=x

$$\delta_k(\mathbf{x}) = \log(\pi_k) + \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$$