# Economic Data Analysis II

- Based on An Introduction to Statistical Learning with R (by James, G., Witten, D., Hastie, T., Tibshirani, R.) [Check here or here]
- References
  - Christian Kleiber and Achim Zeileis, *Applied Econometrics with R*, Springer-Verlag, New York, 2008.
  - Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *Introduction to Statistical Learning with Applications in R*, Springer 2013.

# Regression

- Linear Regression (ISLR Chapter 3)
  - Estimation: Least Squares
  - Prediction:
  - Evaluation: RSS, $R^2$
  - Hypothesis Testing: t, F, $\chi^2$
  - Interpretation:
    Marginal Effects, Elasticity

$$Y_i = X_i\beta + \varepsilon_i,\ i = 1, 2, \ldots, N$$

$$\hat{\beta} = \min_{\beta} \arg \sum_{i=1}^{N} (Y_i - X_i\beta)^2$$

$$\hat{Y}_i = X_i\hat{\beta}, \quad \hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^2$$

# Regression

- Linear Regression (Continued)
  - Extensions
    - Non-linear Variable Transformation: Quadratic, Polynomials, and Interactions
    - Including Qualitative Variables
  - Bias-Variance Trade-Off

# Classification

- Classification Problem (ISLR Chapter 4)
  - E(Y|X) = Pr(Y=1 or 0|X)
  - Logistic Regression
    - Estimation: Maximum Likelihood
    - Prediction
    - Interpretation
  - Probit Model

$$P_i = \int_{-\infty}^{X_i\beta} \tfrac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz$$

$$\log\left(\frac{P_i}{1-P_i}\right) = X_i\beta + \varepsilon_i$$

$$P_i = E(Y_i = 1 \mid X_i) = \mathrm{P}(X_i) = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$$

$$\hat{\beta} = \max_{\beta} \arg \prod_{i,Y_i=1} \mathrm{P}(X_i) \prod_{i,Y_i=0} \left(1-\mathrm{P}(X_i)\right)$$

$$\hat{P}(X_i) = \frac{e^{X_i\hat{\beta}}}{1+e^{X_i\hat{\beta}}}$$

$$\frac{\partial \hat{P}(X_i)}{\partial X_i} = \hat{P}(X_i)(1-\hat{P}(X_i))\hat{\beta}$$

# Classification

- Classification (Continued)
  - Extension: Multiclass Logistic Regression (Mutinomial Regression)

$$P_{ik} = E(Y_i = k \mid X_i) = P_k(X_i) = \frac{e^{X_i \beta^k}}{\sum_{l=1}^{K} e^{X_i \beta^l}}$$

$$\log(P_{ij} / P_{ik}) = X(\beta^j - \beta^k)$$

# Classification

- Classification (Continued)
  - Bayes Theorem for Classification
  - Discriminant Analysis
    - Linear Discriminant Analysis
    - Quadratic Discriminant Analysis

# Model Comparison

- Training vs. Testing
  - Estimate the model with the training data set, and compute prediction error for the testing data set
  - Training Error Statistics vs. Testing Error based on MSE or Misclassification Rate
- Model Comparison is based on Training Error Statistics
  - Cp, Adj-$R^2$
  - AIC, BIC

$$AIC = -2\log L + 2p$$

$$BIC = -2\log L + \log(N)p$$

# Model Validation

- Model Validation is based on Testing Error MSE
  - Simple Validation-Set Approach
    - Randomly split the sample into two sets (halves or fractions): training set and testing set
    - Compute and compare MSE for the testing set
  - Cross Validation
    - K-fold cross validation

# Cross Validation

- Cross Validation (ISLR Chapter 5)
  - K-fold Cross Validation

| | | | | |
|---|---|---|---|---|
| **Test** | **Train** | **Train** | **…** | **Train** |

  - |-------------------- N -------------------------|
  - |-- $N_k$ --|: the number of observations
  - The predicted $\hat{Y}_i^k$ is obtained from the estimated model with the $k^{th}$ part's observations removed

# Cross Validation

- Cross Validation (Continued)
  - K-fold Cross Validation
    - Randomly divide the sample into K equal-sized parts. Leave out part k, fit the model to the other K-1 parts of combined
    - Obtain the prediction errors for the left-out $k^{th}$ part, and compute CV as

$$CV = \sum_{k=1}^{K} \frac{N_k}{N} MSE_k, \; where \; MSE_k = \frac{1}{N_k} \sum_{i=1}^{N_k} (Y_i - \hat{Y}_i^k)^2$$

      - CV tends to be biased upward

# Cross Validation

- Cross Validation (Continued)
  - K-fold Cross Validation
    - If K=N, this is N-fold or leave one out cross validation (LOOCV)
    - Special Case: OLS

$$CV = \sum_{i=1}^{N} \left( \frac{Y_i - \hat{Y}_i}{1 - h_i} \right)^2, \; where\ h = diag[X(X^{'}X)^{-1}X^{'}]$$

# Model Selection

- Model Selection (ISLR Chapter 6)
  - Stepwise Regression
  - Shrinkage Methods
    - Least Absolute Angle (LAR) Regression
  - Projection Methods
    - Involving variable transformation but not necessary for variable selection
      - PCA: Principal Components Analysis
      - Factor Analysis

# Model Selection

- Variable Selection
  - All or Best Subsets Selection
    - $2^p$ Complexity
    - Overfitting or Multicolinearity
  - Stepwise Selection
    - Forward Selection (allow p>N)
    - Backward Selection (require N>p)
      - Search through 1+p(p+1)/2 models, but the best model is not guaranteed

# Model Selection

- Variable Selection (Continued)
  - Shrinkage Method
    - Least Absolute Regression (LAR): A regression technique constrains or regularizes the regression estimates: Fit a regression model with all predictors, but the estimated coefficients are shrunken toward zeros relative to the least squares estimates
    - Ridge Regression
    - The LASSO

# Variable Selection

- Ridge Regression

$$\min_{\beta_1,\ldots\beta_p} \arg \sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 \quad s.t. \quad \sum_{j=1}^{p}\beta_j^2 \le s$$

  – Equivalently,

$$\min_{\beta_1,\ldots\beta_p} \arg \sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

$$where\ \lambda \ge 0,\ and\ \lambda\sum_{j=1}^{p}\beta_j^2 = shrinkage\ l_2 - penality$$

# Variable Selection

- Ridge Regression (Continued)
  - The intercept $\beta_0$ is not restricted
  - The predictors $X_{ij}$ should be standardized for the ridge regression
  - The tuning parameter $\lambda$ is determined separately by cross-validation
  - Ridge regression includes all predictors in the final model. This method can not be used for variable selection

# Variable Selection

- The LASSO (Least Absolute Shrinkage and Selection Operator)

$$\min_{\beta_1,\dots\beta_p} \arg \sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 \quad s.t. \quad \sum_{j=1}^{p}|\beta_j| \le s$$

  – Equivalently,

$$\min_{\beta_1,\dots\beta_p} \arg \sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$where\ \lambda \ge 0,\ and\ \lambda\sum_{j=1}^{p}|\beta_j| = shrinkage\ l_1 - penality$$

$$as\ \lambda \to \infty,\ \beta_j \to 0\ \ for\ some\ \ j$$

# Variable Selection

- The LASSO (Continued)
  - The intercept $\beta_0$ is not restricted
  - The predictors $X_{ij}$ should be standardized for the ridge regression
  - The tuning parameter $\lambda$ is determined separately by cross-validation
  - This method can perform variable selection to achieve a sparse model