Section for Applied Statistics and Data Analysis

TA: Cong Mu

Office Hour: Wednesday 10:00AM - 12:00PM

September 13, 2019

Overview

1. Some Statistics
   - Regression Formulation
   - Least Squares Estimation
   - Goodness of Fit
2. Some Programming
   - Example with `gala` Dataset
   - Exercise with `teengamb` Dataset

General Formulation

Suppose we have n observations and p variables.

$$y = X\beta + \epsilon,$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1}.$$

Estimating Parameters

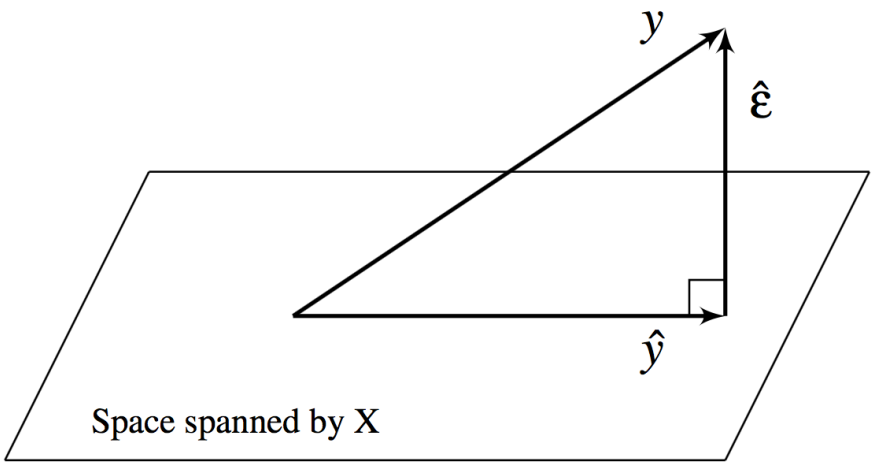

Figure 2.1 Geometrical representation of the estimation β. The data vector Y is projected orthogonally onto the model space spanned by X. The fit is represented by projection $\hat{y} = X\hat{\beta}$ with the difference between the fit and the data represented by the residual vector $\hat{\varepsilon}$.

(Figure from Linear Models with R)

Least Squares Formulation

Recall

$$y = X\beta + \epsilon \qquad \text{where} \qquad \epsilon = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_n \end{bmatrix}^\top.$$

To minimize the sum of the squared errors

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon^\top \epsilon = (y - X\beta)^\top (y - X\beta).$$

$$\frac{\partial L}{\partial \beta} = 0 \qquad \Longrightarrow \qquad \left(X^\top X\right) \beta = X^\top y.$$

$$\widehat{\beta} = \left(X^\top X\right)^{-1} X^\top y, \qquad \widehat{y} = X\widehat{\beta} = X\left(X^\top X\right)^{-1} X^\top y.$$

Some Examples

- $y = \mu + \epsilon$

$$X = \mathbf{1}, \beta = \mu \quad \Longrightarrow \quad \widehat{\beta} = \left(\mathbf{1}^\top \mathbf{1}\right)^{-1} \mathbf{1}^\top y = \frac{1}{n} \mathbf{1}^\top y = \bar{y}.$$

- $y = \beta_0 + \beta_1 x + \epsilon \quad \Longleftrightarrow \quad y = \beta_0 + \beta_1 \bar{x} + \beta_1 (x - \bar{x}) + \epsilon$

$$X = \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix}_{n \times 2} \quad \Longrightarrow \quad X^\top X = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{bmatrix}_{2 \times 2}.$$

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0' \\ \widehat{\beta}_1 \end{bmatrix}_{2 \times 1} = \begin{bmatrix} \bar{y} \\ \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix}_{2 \times 1}.$$

Some Properties

Gauss-Markov Theorem

In a linear regression model in which the errors are uncorrelated with equal variances and expectation value of zero, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator. i.e.

$$\mathbb{E}\left[\widehat{\beta}_{OLS}\right] = \beta.$$

$$Var\left[\widehat{\beta}_{OLS}\right] \leqslant Var\left[\widehat{\beta}\right] \qquad \text{for} \qquad \widehat{\beta} \in \left\{\widehat{\beta} \mid \mathbb{E}\left[\widehat{\beta}\right] = \beta\right\}.$$

Coefficient of Determination

- **Sum of Squares Total**

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

- **Sum of Squares Regression**

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

- **Sum of Squares Error**

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

$$SSTO = SSR + SSE, \qquad R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Data - gala

- **Background**
  The number of species found on the various Galapagos Islands. There are 30 cases (Islands) and 6 variables in the dataset.
- **Variables**
  - `Species`: the number of species found on the island
  - `Area`: the area of the island (km2)
  - `Elevation`: the highest elevation of the island (m)
  - `Nearest`: the distance from the nearest island (km)
  - `Scruz`: the distance from Santa Cruz Island (km)
  - `Adjacent`: the area of the adjacent island (km2)

Thanks for listening!