Section for Applied Statistics and Data Analysis

TA: Cong Mu

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

September 20, 2019

Overview

1. Some Statistics
   - Linear Combinations of Two Normals
   - Least Squares Estimation
   - Goodness of Fit

2. Some Programming
   - Exercises in Faraway

Review

Recall

For random variables $X_1, X_2, \cdots, X_n$

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right].$$

$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var\left[X_i\right] + \sum_{i \neq j} Cov\left[X_i, X_j\right].$$

Review

Recall

Given $f_{X_1,X_2}(x_1,x_2)$ with $Y_1 = g_1(X_1,X_2)$ and $Y_2 = g_2(X_1,X_2)$

$$f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(x_1,x_2)\,|J(x_1,x_2)|^{-1}.$$

where

$$J(x_1,x_2) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix} = \frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2}\frac{\partial g_2}{\partial x_1}$$

and

$$x_1 = h_1(y_1,y_2),\; x_2 = h_2(y_1,y_2)$$

Linear Combinations of Two Normals

Claim

Let $U = aX_1, V = bX_2$ where $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Then $U \sim \mathcal{N}(a\mu_1, a^2\sigma_1^2)$ and $V \sim \mathcal{N}(b\mu_2, b^2\sigma_2^2)$.

Question

Assume X_1 and X_2 are independent. What can you say about $Y = U + V$?

Question

How to show a random variable follows some distribution ?

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

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

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

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

Goodness of Fit

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

where

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

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

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

Exercises in Faraway Chapter 2

- **Exercise 2:** `uswages` dataset
- **Exercise 3:** artificial data
- **Exercise 4:** `prostate` dataset

Thanks for listening!