Regression Analysis
Assignment #1

설명변수가 1개인 아래의 단순선형회귀(simple linear regression)를 고려하자.

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1, 2, \ldots, n)$$

이를 다중선형회귀모형(multiple linear regression)에서 다룬 벡터-행렬표현법으로 바꾸면 아래의 모형으로 생각할 수 있다.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

여기서

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

이다. 오차에 대한 분포는 $e_i \sim N(0, \sigma^2)$이며 서로 독립임을 가정한다.

1. $\mathbf{X}^T\mathbf{X}$를 계산하라. 그리고 $(\mathbf{X}^T\mathbf{X})^{-1}$을 계산하라.

2. $\mathbf{X}^T\mathbf{y}$를 계산하라.

3. 다중선형회귀모형에 의하면 회귀계수의 추정량 $\mathbf{b} = (b_0, b_1)^T$는 $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^T\mathbf{X}^T\mathbf{y}$로 주어진다. 이를 계산하여

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad b_0 = \bar{y} - b_1\bar{x}$$

   임을 보여라. 여기서 $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$이고 $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$이다.

4. 다중선형회귀모형에 의하면 회귀계수의 추정량 $\mathbf{b} = (b_0, b_1)^T$의 분산 및 공분산을 나타내는 행렬은

$$Var(\mathbf{b}) = \begin{pmatrix} Var(b_0) & Cov(b_0, b_1) \\ Cov(b_1, b_0) & Var(b_1) \end{pmatrix} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

   으로 주어짐이 알려져 있다. $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$을 계산하여

$$Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad Var(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

   이 됨을 증명하라. 또한 $Cov(b_0, b_1)$은 어떻게 주어지는지 명시하라.