Tutorial Notes 4 of MATH3424

1 Summary of course material

1.1 Multiple Linear Regression

- Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Suppose there are n observations, each of them can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \cdots, n$$

- Matrix notation:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Then the observed data can written as

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

- Parameter estimation (Least square estimator):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

If $\epsilon_1, \cdots, \epsilon_n$ are i.i.d. with common variance σ^2, an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p-1} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p-1}$$

- Centering and Scaling:

Centering does not affect the regression coefficients except that the estimate of constant term $\hat{\beta}_0'$ is always 0.

Scaling will change the values of the regression coefficient.

2 Questions

1. Using the following summary statistics:

$n = 20$, $\sum_{i=1}^{20} x_{i1} = 114$, $\sum_{i=1}^{20} x_{i2} = -136$, $\sum_{i=1}^{20} y_i = 222$,

$\sum_{i=1}^{20} x_{i1}^2 = 860$, $\sum_{i=1}^{20} x_{i1}x_{i2} = -1025$, $\sum_{i=1}^{20} x_{i2}^2 = 1228$, $\sum_{i=1}^{20} x_{i1}y_i = 1537$,

$\sum_{i=1}^{20} x_{i2}y_i = -1824$, $\sum_{i=1}^{20} y_i^2 = 2950$,

$S_{x_1x_1} = 210.2$ $S_{x_1x_2} = -249.8$ $S_{x_2x_2} = 303.2$ $S_{x_1y} = 271.6$

$S_{x_2y} = -314.4$ $S_{yy} = 485.8$

and

$$\begin{pmatrix} 860 & -1025 \\ -1025 & 1228 \end{pmatrix}^{-1} = \begin{pmatrix} 0.2251146 & 0.1879010 \\ 0.1879010 & 0.1576535 \end{pmatrix}$$

$$\begin{pmatrix} 210.2 & -249.8 \\ -249.8 & 303.2 \end{pmatrix}^{-1} = \begin{pmatrix} 0.227525 & 0.187453 \\ 0.187453 & 0.157737 \end{pmatrix}$$

to fit a model of y on x_1 and x_2, i.e., do the following regression model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

(a) **Assume that $\boldsymbol{\beta_0 = 2}$.**

i. Find the least squares estimates of the unknown parameters β_1 and β_2, then write down the fitted line.

ii. Find the Residual Sum of Squares and the unbiased estimate of the unknown parameter σ^2.

(b) **Assume that $\boldsymbol{2\beta_1 = \beta_2}$.**

· Find the least squares estimates of the unknown parameters β_1 and β_2, then write down the fitted line.

(c) **Assume that $\boldsymbol{\beta_0, \beta_1, \beta_2}$ are all unkown.**

· Find the least squares estimates of the unknown parameters β_0, β_1 and β_2, then write down the fitted line.

2. Consider a situation in which the regression data set is divided into two parts as follows. The model is given by

$$\begin{aligned} y_i &= \beta_0^{(1)} + \beta_1 x_i + \epsilon_i, \quad \text{for } i = 1, \cdots, n_1 \\ &= \beta_0^{(2)} + \beta_1 x_i + \epsilon_i, \quad \text{for } i = n_1 + 1, \cdots, n_1 + n_2 \end{aligned}$$

In other words there are two regression lines with common slope. Using the centered model,

$$\begin{aligned} y_i &= \beta_0^{(1)^*} + \beta_1 (x_i - \bar{x}_1) + \epsilon_i, \quad \text{for } i = 1, \cdots, n_1 \\ &= \beta_0^{(2)^*} + \beta_1 (x_i - \bar{x}_2) + \epsilon_i, \quad \text{for } i = n_1 + 1, \cdots, n_1 + n_2 \end{aligned}$$

where $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$ and $\bar{x}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} x_i$.

Show that the least squares estimate of β_1 is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_1) y_i + \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2) y_i}{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2 + \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2)^2}$$