Tutorial Notes 3 of MATH3424

1 Summary of course material

1.1 Covariance, Correlation Coefficient

- Covariance of X and Y is defined by

$$\mathrm{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} \tag{1}$$

- Correlation and coefficient between X and Y is given by

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(Y, X)}{s_y s_x} \tag{2}$$

1.2 Simple Linear Regression Model

- A simple linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, \cdots, n$$

Core assumption:

$$\epsilon_1, \cdots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

- Parameter estimation (least square estimates/unbiased/standard error/distribution):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

- Fitted values and residuals

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad e_i = y_i - \hat{y}_i$$

- Measuring the quality of fit:

$$\mathrm{SST} = \sum (y_i - \bar{y})^2 \quad \mathrm{SSR} = \sum (y_i - \bar{y})^2 \quad \mathrm{SSE} = \sum (y_i - \bar{y})^2$$

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

- Hypothesis test and confidence intervals: $H_0: \beta_1 = \beta_1^0$ v.s. $H_1: \beta_1 \neq \beta_1^0$

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} \qquad \hat{\beta}_1 \pm t_{(n-2, \alpha/2)} \times s.e.(\hat{\beta}_1)$$

if $|t| > t_{(n-2, \alpha/2)}$, reject H_0

- No intercept model: difference

2 Questions

1. Consider a simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $i = 1, \cdots, n$.

   - You are given 5 pairs of (x_i, y_i) where y_4 is missing and the fitted line passes through the point $(3, 1.65)$. Find c and then determine $\sum_{i=1}^{5}(y_i - \bar{y})^2$.

     $(\bar{x}, \bar{y})$

x_i	1	2	3	4	5
y_i	0.25	1.75	1.5	c	2.5

   - $n=25$

     Given the following statistics from 25 pairs of (x_i, y_i):

     $$\bar{x} = 0, \quad \hat{\sigma}^2 = 100, \quad \hat{\beta}_0 = 3$$

     $\alpha = 0.02$

     determine the length of a 98% confidence interval for β_0.

   - Given the following statistics from 10 pairs of (x_i, y_i): $n=10, \ \alpha = 0.05$

     $$\sum_{i=1}^{10}(x_i - \bar{x})^2 = 400, \quad \sum_{i=1}^{10}(y_i - \bar{y})^2 = 425, \quad \sum_{i=1}^{10}(\hat{y}_i - \bar{y})^2 = 225$$

     S_{xx} $\quad$ SST $\quad$ SSR

     Calculate the test statistic for testing the hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ by t test. Write down your conclusion clearly. Set the significance level at $\alpha = 0.05$

Sol: 1. (a) $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} i = 3$

$\bar{y} = 1.65 = \dfrac{0.25 + 1.75 + 1.5 + 2.5 + c}{5} \Rightarrow c = 1.65 \times 5 - 6 = 2.25$

(b) $\hat{\beta}_0 \pm t_{(23, 0.01)} \times s.e.(\hat{\beta}_0)$ $\quad \approx 2.5$

$s.e.(\hat{\beta}_0) = \hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \dfrac{\hat{\sigma}}{\sqrt{n}} = \dfrac{10}{5} = 2.$

$3 \pm 2.5 \times 2$

length of 98% CI $\approx 2 \times 2.5 \times 2 \approx 10.$

(c) $t = \dfrac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$, $\quad SSE = SST - SSR = 200,$

$\quad \sum_{i=1}^{11}(\hat{y}_i - y_i)^2.$

$\hat{\sigma}^2 = \dfrac{SSE}{n-2} = \dfrac{200}{8} = 25$

$s.e.(\hat{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \dfrac{\hat{\sigma}}{\sqrt{S_{xx}}} = \dfrac{5}{20} = \dfrac{1}{4}$

$$\underline{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\underline{\hat{\beta}_0} + \underline{\hat{\beta}_1 x_i} - \underline{\hat{\beta}_0} - \underline{\hat{\beta}_1 \bar{x}})^2$$

$$= \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 \underline{S_{xx}}$$

$$\hat{\beta}_1^2 = \frac{SSR}{S_{xx}} = \frac{225}{400}$$

$$|\hat{\beta}| = \sqrt{\hat{\beta}_1^2} = \frac{15}{20} = \frac{3}{4}$$

$$|t| = \frac{\frac{3}{4}}{\frac{1}{4}} = 3 \quad > \quad t_{(8,\, 0.025)} = 2.306$$

So we reject H_0.

2. Consider a linear model, for $i = 1, \cdots, 3$ $\quad S(\beta_0,\beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

$$y_i = \beta_0 + i\beta_1 + \epsilon_i$$

i	1	2	3
x_i	1	2	3

$\bar{x} = 2$

where ϵ_i follows independent normal distribution with mean 0 and variance $i\sigma^2$.

- Find the least squares estimates of β_0 and β_1 in terms of y_i
- Find the $Var(\hat{\beta}_0)$ and $Var(\hat{\beta}_1)$.

Sol: (a)

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{3}(i-2)y_i}{\sum_{i=1}^{3}(i-2)^2} = \frac{1}{2}(y_3 - y_1).$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{1}{3}(y_1+y_2+y_3) - (y_3-y_1) = \frac{4}{3}y_1 + \frac{1}{3}y_2 - \frac{2}{3}y_3.$$

(b)

$$Var(\hat{\beta}_0) = Var\left(\frac{4}{3}y_1 + \frac{1}{3}y_2 - \frac{2}{3}y_3\right) = \frac{16}{9}Var(y_1) + \frac{1}{9}Var(y_2) + \frac{4}{9}Var(y_3) = \frac{16}{9}\sigma^2 + \frac{2}{9}\sigma^2 + \frac{12}{9}\sigma^2 = \frac{10}{3}\sigma^2$$

$$Var(\hat{\beta}_1) = \frac{1}{4}(3\sigma^2 + \sigma^2) = \sigma^2$$

$$Cov(aX+bY, cZ) = a\,Cov(X, cZ) + b\,Cov(Y, cZ)$$
$$= ac\,Cov(X,Z) + bc\,Cov(Y,Z)$$

3. Consider the simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $i = 1, \cdots, n$. Show that $Cov(\bar{y}, \hat{\beta}_1) = 0$ and $Cov(e_i, \hat{\beta}_1) = 0$.

$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

population covariance

Proof: $d_i = \dfrac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ $\quad \sum x_i = n\bar{x}$

$$Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\frac{1}{n}\sum_{i=1}^n y_i, \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = Cov\left(\frac{1}{n}\sum_{i=1}^n y_i, \sum_{i=1}^n d_i y_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^n Cov(y_i, d_i y_i) = \frac{1}{n}\sum_{i=1}^n d_i \underbrace{Cov(y_i, y_i)}_{Var(y_i) = \sigma^2} = \frac{\sigma^2}{n}\sum_{i=1}^n d_i = 0$$

$$Cov(e_i, \hat{\beta}_1) = Cov(y_i - \hat{y}_i, \hat{\beta}_1) = Cov(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \hat{\beta}_1)$$

$$= Cov(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i, \hat{\beta}_1) = Cov(y_i - (\bar{y} + \hat{\beta}_1 (x_i - \bar{x})), \hat{\beta}_1)$$

$$= Cov\left(y_i - (\bar{y} + d_i \sum_{j=1}^n (x_j - \bar{x}) y_j), \sum_{i=1}^n d_i y_i\right)$$

$$= \underbrace{Cov\left(y_i, \sum_{i=1}^n d_i y_i\right)}_{d_i Cov(y_i, y_i)} - \underbrace{Cov\left(\bar{y}, \sum_{i=1}^n d_i y_i\right)}_{=0} - \underbrace{Cov\left(d_i \sum_{j=1}^n (x_j - \bar{x}) y_j, \sum_{j=1}^n d_j y_j\right)}$$

$= d_i \sigma^2$

$$d_i \sum_{j=1}^n (x_j - \bar{x}) d_j \underbrace{Cov(y_j, y_j)}$$

$$d_i \sigma^2 \sum_{i=1}^n \frac{(x_j - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

$$= d_i \sigma^2$$

$$= d_i \sigma^2 - d_i \sigma^2 = 0$$