Regression Analysis 2023-04-01

Q2 (a) $\bar{X} = 0$. so $\sum_i (X_i - \bar{X})^2 = \sum_i X_i^2 = 2m$.

$$\hat{\beta} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i X_i Y_i}{2m} = \frac{1}{2}\left[-\bar{Y}_{(1)} + \bar{Y}_{(3)}\right]$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \bar{Y} = \frac{1}{3}\left(\bar{Y}_{(1)} + \bar{Y}_{(2)} + \bar{Y}_{(3)}\right)$$

(b)

$$\mathrm{Cov}\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \sigma^2 (X^T X)^{-1} = \sigma^2 \left(\begin{bmatrix} 1 & -1 \\ \vdots & -1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}^T \begin{bmatrix} 1 & -1 \\ \vdots & -1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \right)^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{3m} & 0 \\ 0 & \frac{1}{2m} \end{bmatrix}$$

(c) The line connects $(-1, \bar{Y}_{(1)})$ and $(1, \bar{Y}_{(3)})$ is

$$Y = \frac{\bar{Y}_{(1)} + \bar{Y}_{(3)}}{2} + \frac{\bar{Y}_{(3)} - \bar{Y}_{(1)}}{2} X$$

Same slope but different intercept

Q3 (a)

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{2m} \\ Y_{2m+1} \\ \vdots \\ Y_{4m} \end{bmatrix} = \underbrace{\begin{bmatrix} -1 & Z_1 \\ \vdots & \vdots \\ -1 & Z_{2m} \\ 1 & Z_{2m+1} \\ \vdots & \vdots \\ 1 & Z_{4m} \end{bmatrix}}_{X} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{2m} \\ \varepsilon_{2m+1} \\ \vdots \\ \varepsilon_{4m} \end{bmatrix}$$

(b) $\hat{\beta} = (X^T X)^{-1} X^T Y$

$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

(c) We need $\sigma^2 (X^T X)^{-1}$ to be diagonal.

$$X^T X = \begin{bmatrix} 4m & \sum_{i=2m+1}^{4m} Z_i - \sum_{i=1}^{2m} Z_i \\ \sum_{i=2m+1}^{4m} Z_i - \sum_{i=1}^{2m} Z_i & 4m \end{bmatrix}$$

We need $\sum_{i=2m+1}^{4m} Z_i = \sum_{i=1}^{2m} Z_i$ to have independence.

Besides, we need $\varepsilon | X \sim N(0, \sigma^2 I)$

Q4 (a) OLS minimizes $(y - X\beta)^T (y - X\beta)$.

$\hat{\beta} = (X^T X)^{-1} X^T y$

$Cov(\hat{\beta}) = (X^T X)^{-1} X^T V(y|X) X (X^T X)^{-1} = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}$

(b) GLS minimizes $(y - X\beta)^T \Sigma^{-1} (y - X\beta)$

$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y$

(c) GLS is BLUE

Q5 (a) $\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 3.3 \\ 1.14 \end{bmatrix}$

(b) $\tilde{\beta} = \hat{\beta} - (X^T X)^{-1} L^T [L(X^T X)^{-1} L^T]^{-1} L\hat{\beta}$ with $L = [1, 1]$

$= \begin{bmatrix} -3.72 \\ 3.72 \end{bmatrix}$

(c) We can use a t-test with $a = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

$$\frac{a^T \tilde{\beta} / \sqrt{a^T (X^T X)^{-1} a}}{\sqrt{\hat{e}^T \hat{e} / (n-p)}} \sim t(n-p) \quad \text{where } n = 6, \; p = 2.$$

Q6 (a) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. $x_i = \begin{cases} 0, & i = 1, \dots, n \\ 1, & i = n+1, \dots, 3n \end{cases}$

(b) $\hat{\sigma}_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$.

Q7 (a) Persons in intensive care $= y$

Number of inhabitants $= x$

$y = \beta_0 + \beta_1 x + \beta_2 1(\text{sweden}) + e.$

we need $e | x$, Sweden to be normal

$e | x$, Germany to be normal

(b) $H_0: \beta_2 \geq 0 \quad H_1: \beta_2 < 0$

(c) We can use t test. Same t test statistic for testing $H_0: \beta_2 = 0 \quad H_1: \beta_2 \neq 0$. But it is one sided t test.

Q8 (a) $E(\text{LifeExpF}) = \beta_0 + \beta_1 \text{ppgdp} + \beta_2 \log(\text{fertility}) + \beta_3 \text{pctUrban}$

We need LifeExpF given all regressors be normal.

(c) t tests. $H_0: \beta_i = 0, \quad H_1: \beta_i \neq 0$

F test: $H_0: \beta_1 = \beta_2 = \beta_3 = 0 \quad H_1$: some non zero.

(d) $$\frac{(RSS_0 - RSS_1)/3}{RSS_1/195} = 168.2$$

$$\frac{(RSS_0 - RSS_2)/1}{RSS_2/197} = 402.5$$

Hence $RSS_0 = \left(\frac{3 \cdot 168.2}{195} + 1\right) RSS_1 = \left(\frac{402.5}{197} + 1\right) RSS_2$

Then

$$F = \frac{(RSS_2 - RSS_1)/2}{RSS_1/195} = 17.45$$

The critical value is $F_{0.95}(2, 195) = 0.38$.

Hence model 1 fits the data better if the normality assumption is plausible

Otherwise, we can use bootstrap.