Student id: 1291822

ASSIGNMENT 2

1) For the case $y = \beta_0 + \beta_1 x + \epsilon$, we saw in class that

$$Var\, b = (X^T X)^{-1} \sigma^2$$

where

$$X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$$

but

Here, $y = \beta_0 + \beta_1 (x - \bar{x}) + \epsilon$

$\therefore$ Here $X' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 - \bar{x} & x_2 - \bar{x} & \cdots & x_n - \bar{x} \end{bmatrix}^T$

where $\bar{x} = \dfrac{\Sigma x_i}{n}$

$$Var\, b = (X'^T X')^{-1} \sigma^2$$

Now, $X'^T X' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 - \bar{x} & x_2 - \bar{x} & \cdots & x_n - \bar{x} \end{bmatrix} \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix}$

Non diagonal terms $= (x_1 - \bar{x})1 + (x_2 - \bar{x})1 + \ldots + (x_n - \bar{x})1$
$= \Sigma x_i - n\bar{x} = \Sigma x_i - \Sigma x_i = 0$

The non-diagonal terms of $Var\, b = 0$
$\Rightarrow b_1$ & b_0 are uncorrelated.

2) Here $X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 8 & 8 & 8 & 12 & 12 & 12 & 12 & 14 & 14 & 16 & 16 & 16 & 16 & 20 & 20 \end{bmatrix}^T$

$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$, $\epsilon = [\epsilon_1 \ \epsilon_2 \ \dots \ \dots \ \epsilon_{15}]^T$

$y_{15\times1} = [10.36 \ \ 9.52 \ \ 9.34 \ \ 20.97 \ \ 21.35 \ \dots \dots \ 43.69 \ \ 37.22]^T$

a) $y = X\beta + \epsilon$ is the linear model

where best estimate of β, $b = (X^TX)^{-1}X^Ty$

b) Least square estimate we get $b = [-11.08 \ \ 2.62]^T$

c) Sample variance, $s^2 = \dfrac{(y-Xb)^T(y-Xb)}{15-2}$

$= \dfrac{138}{13} = 10.6$

d) For 18 years of formal education, income $= -11.08 + 2.62 \times 18$
$= 36.1$

3 a) $(1-\alpha)$ confidence interval for $\beta_1 = b_1 \pm t_{\alpha/2}\, S\sqrt{C_{22}}$

$S = \sqrt{10.6} = 3.26$

$C_{22} = 0.00477$

$b_1 = 2.62$

$\therefore$ interval $= 2.62 \pm t_{\alpha/2}\, 0.225$

Given interval $= (2.21929, 3.017580)$

$\Rightarrow 3.017580 - 2.21929 = 2 \times t_{\alpha/2} \times 0.225$

$\Rightarrow t_{\alpha/2} = 1.77$

For 13 degrees of freedom $\frac{\alpha}{2} = 0.05 \Rightarrow \alpha = 0.1$

$\Rightarrow$ Interval has 90% confidence.

b) Under $C\beta = \delta^*$

$$\frac{(Cb-\delta^*)^T [C(X^TX)^{-1}C^T]^{-1}(Cb-\delta^*)/r}{S^2}$$

follows $F_{r,n-p}$ distribution.

For $\beta = \begin{bmatrix} -11.08 \\ 2.62 \end{bmatrix}$

$C = I_{2\times2}$, $\delta^* = \begin{bmatrix} -11.08 \\ 2.62 \end{bmatrix}$

$\Rightarrow \dfrac{(\beta-\delta^*)^T(X^TX)(\beta-\delta)}{2\times 10.6}$ follows $F_{2,13}$

$\Rightarrow$ 95% joint region

$$\begin{pmatrix} \beta_0+11.08 & \beta_1-2.62 \end{pmatrix} \begin{bmatrix} 15 & 204 \\ 204 & 2984 \end{bmatrix} \begin{pmatrix} \beta_0+11.08 \\ \beta_1-2.62 \end{pmatrix} \leq F^{0.95}_{2,13} \times 2 \times 10.6$$

$$\Rightarrow 15(\beta_0+11.08)^2 + 408(\beta_1-2.62)(\beta_0+11.08) + 2984(\beta_1-2.62)^2 \leq 80.56$$

This gives us an elliptical region inside of which the (β_0, β_1) values correspond to 95% confidence.

c) $1-\alpha = 0.99, \Rightarrow \alpha = 0.01$

$x_0 = [1 \quad 18]$

99% Confidence interval for one prediction $= \hat{y}_0 \pm t_{n-p}^{\alpha/2} S\sqrt{1+x_0^T(X^TX)^{-1}x_0}$

$= 36.1 \pm t_{13}^{0.005} \times 3.26\sqrt{1+0.159}$

$= 36.1 \pm 3.012 \times 3.26 \times 1.076584$

$= 36.1 \pm 10.58101$

$= (25.52, 46.68)$

4) a) $X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 7.2 & 10 & 9 & 5.5 & 9 & 9.8 & 14.5 & 8 \\ 8.7 & 9.4 & 10.0 & 9 & 12 & 11 & 12 & 13.7 \\ 5.5 & 4.4 & 4 & 7 & 5 & 6.2 & 5.8 & 3.9 \end{bmatrix}^T$

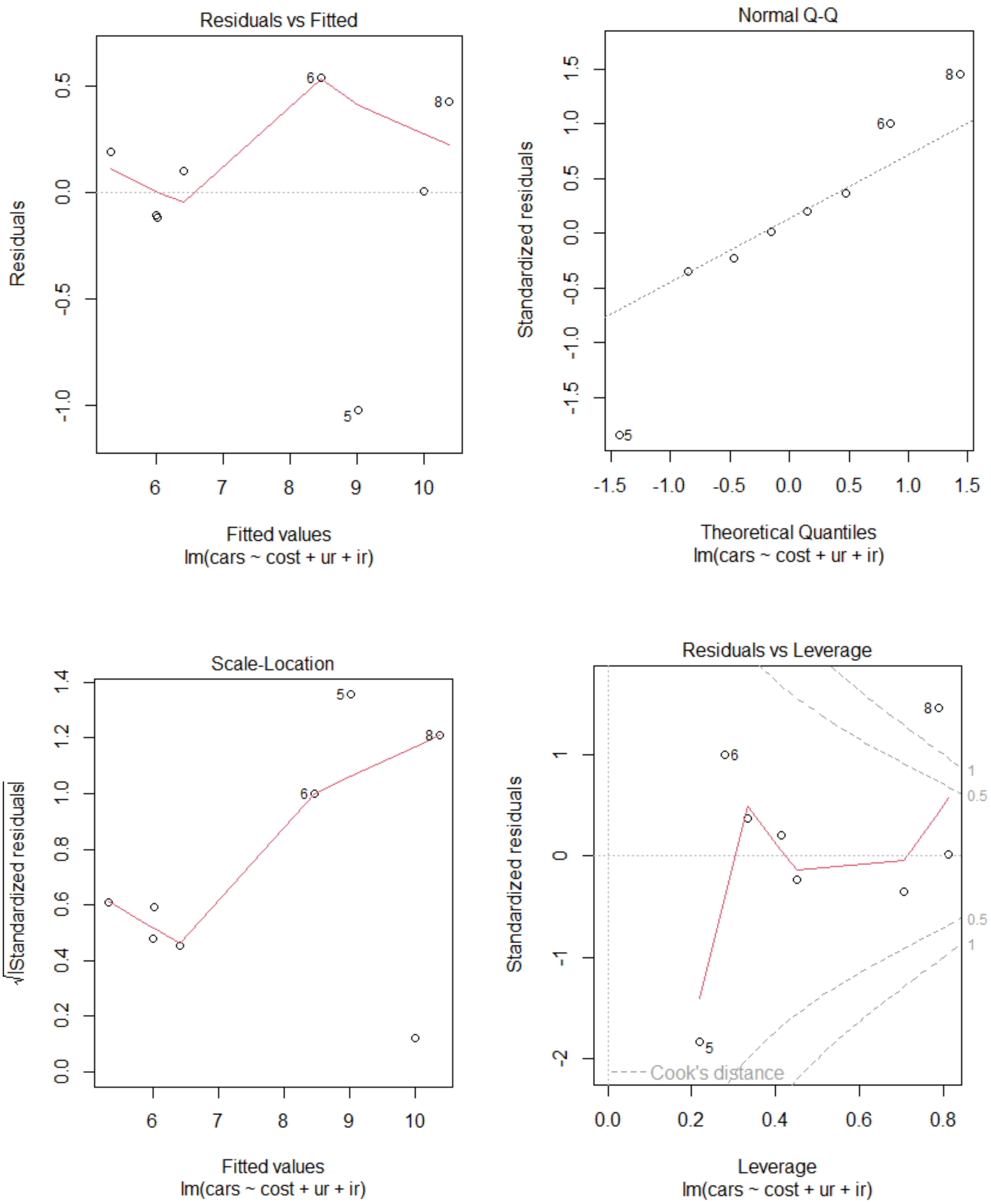
$y = [5.5 \quad 5.9 \quad 6.5 \quad 5.9 \quad 8 \quad 9 \quad 10 \quad 10.8]^T$

Using $b = (X^TX)^{-1}X^Ty$ we get,

$b = [-7.40 \quad 0.121 \quad 1.117 \quad 0.386]^T$

$SSR = 1.058$

Estimate of σ^2, $\hat{\sigma}^2 = \dfrac{SSR}{8-4} = 0.396$


Residuals vs Fitted
Residuals
Fitted values
lm(cars ~ cost + ur + ir)
Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(cars ~ cost + ur + ir)
Scale-Location
√|Standardized residuals|
Fitted values
lm(cars ~ cost + ur + ir)
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(cars ~ cost + ur + ir)

b) i) From the Residuals V/s fitted values plot we see that, points 5, 6 & 8 show large residuals with point 5 showing very large residual.

Points are too less to determine any significant trend.

(ii) From the normal Q-Q plot we again see points 5, 6 & 8 being outliers & point 5 being an extreme one, the distribution looks long-tailed which may imply residuals don't follow normal distributions though, data points are too less for any conclusion.

(iii) In $\sqrt{\text{Standardized residual}}$ V/s Fitted values plot we see that points 5, 6 & 8 are outliers, we also observed that the variance in $\sqrt{SR}$ increases with fitted values indicating variance in ϵ is not constant.

(iv) In SR V/s Leverage plot we see points 5, 6 & 8 having high SR but points 5 & 6 have low leverage in contrast to point 8 which has very high leverage & Cook's distance.
Points show high variance at low & high leverages while moderate variation at medium leverage.

c) $H = X(X^TX)^{-1}X^T$, for 5^{th} point

Leverage $= H_{55}$

$= \underline{\underline{0.22}}$

Standardized residual $= \dfrac{\text{residual}}{\sqrt{\sigma^2(1-H_{55})}} = \dfrac{-1.0228}{\sqrt{0.396(1-0.22)}}$

$= \underline{\underline{-1.84}}$

Cook's distance, $D_5 = \dfrac{1}{4} \times (-1.84)^2 \times \dfrac{0.22}{(1-0.22)} = \underline{\underline{0.239}}$

d) 90% prediction $\Rightarrow \alpha = 0.10$

$x_0 = [1 \; 7 \; 8.6 \; 5]$

$$\text{Interval} = \hat{y}_0 \pm t_4^{0.05} \times 0.629 \times \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

$$= 4.98 \pm 1.6$$

$$= (3.38, 6.58)$$

= (3,380 to 6,580 cars sold)

e) $SSRes = \text{sum}(e^2) = 1.582$

$SSReg = \hat{y}^T y = 502.178$

We see that $SSReg >> SSRes$

For F-test: $MSReg = SSReg/p = 502.178/4 = 125.54$

$||^{ly}$ $MSRes = SSRes/(n-p) = 1.582/4 = 0.395$

$$Fstat = \frac{MSReg}{MSRes} = 317.40$$

Fstat follow F-statistic with (4,4) degrees of freedom. It's 99% threshold is at 15.98 and since Fstat > 15.98 we reject that $\beta = 0$. $\Rightarrow$ Our model is relevant.

f) $H_0: \beta_2 = 1$ v/s $\beta_2 \neq 1$

t-test :

$$t_{stat} = \frac{|b_2 - \beta_2|}{s\sqrt{C_{33}}} = \frac{|1.117 - 1|}{0.629 \times \sqrt{0.0622}} = 0.749$$

The tstat has 4 degrees of freedom : p-value = 0.248 *, ∴ we cannot reject the null-hypothesis ($\beta_2 = 1$) at 0.05 level.

F-test :

$$F_{stat} = \frac{R(\beta_2 | \beta_0, \beta_1, \beta_3)/1}{s^2} \quad \text{follows } F_{1,4} \text{ statistic}$$

For only one parameter, $F_{stat} = t^2_{stat} = 0.749^2 = 0.5604$

We cannot reject the null hypothesis at 0.05 level (critical value = 7.71).

g) $\beta_1 = \beta_3 = 0$; $r = 2$

$$R(\beta_1, \beta_3 | \beta_0, \beta_2) = (b_1 \;\; b_3) \, A_{11}^{-1} \begin{pmatrix} b_1 \\ b_3 \end{pmatrix}$$

$$A_{11} = \begin{bmatrix} C_{22} & C_{24} \\ C_{42} & C_{44} \end{bmatrix} \quad \text{where } C_{ij} \text{ corresponds to } C = (X^T X)^{-1}$$

$$= \begin{bmatrix} 0.0245 & -0.00194 \\ -0.00194 & 0.1353 \end{bmatrix}$$

We get, $R(\beta_1, \beta_3 | \beta_0, \beta_2) = 2240$

$$F_{stat_{2,4}} = \frac{R(\beta_1, \beta_3 | \beta_0, \beta_2)/2}{0.396} = \frac{1120}{0.396} = 2828$$

This rejects null hypothesis ~~of~~ $\beta_1 = \beta_3 = 0$ at 0.01 level (critical value = 18)