Name: CHAN King Yeung
SID = 1155119394

STAT 3008

Assignment 3

Question 1

a) $\hat{\beta} = (X'X)^{-1}X'Y$

$$= \begin{pmatrix} 0.45002 & -0.242619 & 0.020761 \\ -0.242619 & 0.167181 & -0.015105 \\ 0.020761 & -0.015105 & 0.001389 \end{pmatrix} \begin{pmatrix} 1359.06 \\ 6322.83 \\ 47452.65 \end{pmatrix}$$

$$= \begin{pmatrix} 62.73 \\ 10.551 \\ -1.3792 \end{pmatrix} \Rightarrow \hat{\beta}_0 = 62.73,\ \hat{\beta}_1 = 10.551,\ \hat{\beta}_2 = -1.3792$$

b) $RSS = Y'Y - Y'X\hat{\beta}$

$$= 89882.2642 - \begin{pmatrix} 1359.06 & 6322.83 & 47452.65 \end{pmatrix} \begin{pmatrix} 62.73 \\ 10.551 \\ -1.3792 \end{pmatrix}$$

$$\approx 3362.946$$

$$\hat{\sigma} = \sqrt{\frac{3362.946}{21}}$$

$$\approx 12.65$$

c) $\left.\frac{d\hat{y}}{dx}\right|_{x^*} = 0$

$$0 = \hat{\beta}_1 + 2\hat{\beta}_2 x^*$$

$$x^* = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$$

point estimate of $x^* = 3.825$

d)

	df	SS	MS	F_0	p-value
Regression	2	9559.1481	4779.5741	29.8462	7.2704×10^{-7}
Residual	21	3362.946	160.1403		
Total	23	12922.0941			

e) Choice of EV values are not reasonable. Polynomial regression relies on the data points in the center of data points to obtain a reliable $\hat{\sigma}^2$. However, data are close to each other and allocated in 2 sides. Only 1 data point in the middle. It is difficult to find out the optimal value of x.

Question 3

a)

	variable(s)	AIC
step 1:	null	-68.1
	(x_1)	-151
	(x_2)	-121.8
	(x_3)	-148.8
	(x_4)	-66.7

$\therefore$ select x_3 since $-148.8 < -68.1$

step 2:	(x_3)	-148.8
	$(x_1 \mid x_3)$	-149.3
	$(x_2 \mid x_3)$	-448.1
	$(x_4 \mid x_3)$	-148.1

$\therefore$ select x_2 since $-448.1 < -148.8$

step 3:	(x_2, x_3)	-448.1
	$(x_1 \mid x_2, x_3)$	-608.2
	$(x_4 \mid x_2, x_3)$	-459.5

$\therefore$ select x_1 since $-608.2 < -448.1$

step 4:	(x_1, x_2, x_3)	-608.2
	$(x_4 \mid x_1, x_2, x_3)$	-7317.6

$\therefore$ select x_4 since $-7317.6 < -608.2$

$\Rightarrow$ the parsimonious model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

b)

	variable(s)	BIC
step 1:	(x_1, x_2, x_3, x_4)	-7304.6
	(x_1, x_2, x_3)	-597.8
	(x_1, x_2, x_4)	-7306.7
	(x_1, x_3, x_4)	-138.6
	(x_2, x_3, x_4)	-449.1

$\therefore$ discard x_3 since $-7306.7 < -7304.6$

step 2:	(x_1, x_2, x_4)	-7306.7
	(x_1, x_2)	-601.3
	(x_1, x_4)	-143.1
	(x_2, x_4)	-112.5

$\Rightarrow$ the parsimonious model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + e$$

c)

$$-7317.6 = n\ln\left(\frac{RSS}{n}\right) + 2(5) \quad \ldots (1)$$

$$-7304.6 = n\ln\left(\frac{RSS}{n}\right) + 5\ln(n) \quad \ldots (2)$$

by (1) and (2), $n = 99.4843$

≈ 100

d) Yes, the AIC in model 16 is -7317.6 based on x_1, x_2, x_3 and x_4, and the AIC in model 12 is -7317.1 based on x_1, x_2 and x_4. The change in AIC or RSS is not very significant, that is the variation explained by x_3 is not very much while having x_1, x_2 and x_4. Thus, x_3 are highly correlated with x_1, x_2, and x_4.