Midterm Review

D1 (a) Since we are minimizing SSR. $\sum_{i=1}^{n} E_i^2$

$$RSS = S(A,B) = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - A - BX_i)^2$$

① $$\frac{\partial S(A,B)}{\partial (A)} = \sum_{i=1}^{n} (-1)(2)(Y_i - A - BX_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (Y_i - A - BX_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} A + \sum_{i=1}^{n} BX_i$$

$$\Rightarrow \sum_{i=1}^{n} Y_i = nA + B\sum_{i=1}^{n} X_i$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} Y_i = A + B\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \Rightarrow \bar{Y} = A + B\bar{X}$$

$$\Rightarrow A = \bar{Y} - B\bar{X}$$

(b) From ①, $\sum_{i=1}^{n} (-1)(2)(Y_i - A - BX_i) = 0$

$$\sum_{i=1}^{n} (Y_i - A - BX_i) = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i \Rightarrow \Sigma Y_i = \Sigma \hat{Y}_i$$

(c) $$\frac{\partial S(A,B)}{\partial (B)} = \sum_{i=1}^{n} (-1)(2)\, X_i (Y_i - A - BX_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} X_i (Y_i - \bar{Y} + B\bar{X} - BX_i) = 0 \Rightarrow \sum_{i=1}^{n} X_i (Y_i - \bar{Y}) = B\sum_{i=1}^{n} X_i (X_i - \bar{X})$$

$$\Rightarrow B = \frac{\sum_{i=1}^{n} X_i (Y_i - \bar{Y})}{\sum_{i=1}^{n} X_i (X_i - \bar{X})} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\Rightarrow B = \sum_{i=1}^{n} k_i Y_i \quad \text{where } k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Therefore B is linear combination of Y_i

$$E[B] = E\left[\sum_{i=1}^{n} k_i Y_i\right] = \sum_{i=1}^{n} k_i E[Y_i] = \sum_{i=1}^{n} k_i (\alpha + \beta X_i)$$

$$= \sum_{i=1}^{n} \alpha k_i + \sum_{i=1}^{n} \beta X_i k_i = \alpha \cdot \frac{\sum_{i=1}^{n} (X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} + \beta \cdot \frac{\sum_{i=1}^{n} (X_i - \bar{X}) X_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$= 0 \cdot \alpha + \beta \cdot \frac{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \beta \cdot 1 = \beta$$

$\therefore$ B is unbiased estimator of β

(d) $Regss = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} (A + BX_i - A - B\bar{X})^2$

$= \sum_{i=1}^{n} (BX_i - B\bar{X})^2 = \sum_{i=1}^{n} B^2 (X_i - \bar{X})^2$

$= \underbrace{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right)}_{S_{XX}} B = B S_{XX}^2$

D2

	X_i #叶 class	Y_i #叶 stats class	$X_i Y_i$	X_i^2	$\hat{Y}$	$E_i = Y_i - \hat{Y}$
i	1	1	1	1	0.6	0.4
	2	1	2	4	1.3	-0.3
	3	2	6	9	2	0
	4	2	8	16	2.7	-0.7
	5	4	20	25	3.4	0.6

$\bar{X} = \frac{(1+2+3+4+5)}{5} = 3$

$\bar{Y} = \frac{(1+1+2+2+4)}{5} = 2$

(a) Sme $B = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{(1+4+9+16+25) - 5 \cdot 3 \cdot 2}{(1+4+9+16+25) - 5 \cdot (3)^2}$

$= \frac{37 - 5(3)(2)}{55 - 5(3)^2} = 0.7$

$A = \bar{Y} - B\bar{X} = 2 - 0.7 \cdot 3 = 2 - 2.1 = -0.1$

Mode is $\hat{Y} = -0.1 + 0.7X$

We calculate $\hat{Y} = -0.1 + 0.7X_i$ and fill in the table

(b) $RSS = \sum (Y_i - \hat{Y}_i)^2 = 0.4^2 + 0.3^2 + 0^2 + 0.7^2 + 0.6^2 = 1.1$

(c) A: the # of stats class a person took is -0.1 when the person does not take any CS class

B: with every unit increase in # of CS class the student take, (one more class in CS)

the person will have 0.7 unit increase in Stats classes taken.

A does not make sense since a person cannot take negative number of classes

(d) Conduct t-test

Null Hypothesis: $\beta = 0$, there is no linear relationship

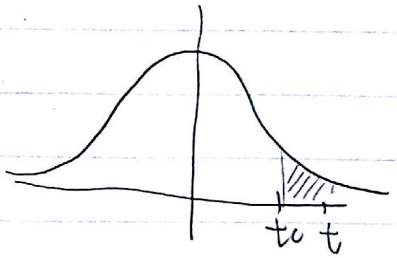
Alternative Hypothesis $\beta \neq 0$. there exist a linear relationship

$$RMS = \frac{RSS}{n-2} = \frac{1.1}{5-2} = \frac{1.1}{3} = 0.3667$$

$$t = \frac{B}{\sqrt{\frac{RMS}{SXX}}} \qquad SXX = \sum_{i=1}^{5} xi^2 - 5\bar{X}^2 = 10$$

$$t = \frac{0.7}{\sqrt{\frac{0.3667}{10}}} = \frac{0.7}{\sqrt{0.03667}} = 3.65546$$

By R_1 we know $t_{c,\alpha=0.025,df=3} = 3.1824$



Since $t = 3.65546 > t_c = 3.1824$

Therefore we reject the null hypothesis there is a linear relationship between # of CS class taken and # of Stats classes taken

(e) ANOVA Table

source of variation	sum of squares	df	mean square	F
Regression	Regss = 4.9	1	RegMS = $\frac{Regss}{1}$ = 4.9	F = $\frac{RegMS}{RMS}$ = $\frac{4.9}{0.3667}$ = 13.3624
Residual	RSS = 1.1	3	RMS = 0.3667	
Total	TSS = 6			

Regss = $B^2 SXX$ = $0.7^2 \cdot 10$ = 4.9

(f) $\because F_{cdf=1,3,} = 10.13$ for $\alpha = 0.05$

Since $F = 13.3624 > F_C = 10.13$

Thus, we reject H_0 and there is a linear relationship between # of CS classes taken and # of stats classes taken

(g) Yes. same result from α and f

(h) $t^2 = (3.65546)^2 = 13.3624 = F$

They are approximately equal

(i) $R^2 = \frac{Regss}{Tss} = \frac{4.9}{6} = 0.81667 \approx 0.82$

Therefore, 82% of the variation in the # stats classes taken can be explained by the # of CS classes taken by our model.

(j) Since $t_{0.025, df=3} = 3.1824$

Since Confidence interval for β is $[B - t_{\alpha/2, df}\sqrt{\frac{RMS}{Sxx}}, B + t_{\alpha/2, df}\sqrt{\frac{RMS}{Sxx}}]$

$t_{\alpha/2, df}\sqrt{\frac{RMS}{Sxx}} = t_{0.025, df=3}\sqrt{\frac{0.3667}{10}} = 3.1824\sqrt{\frac{0.3667}{0.1}} = 6.0942$

$\Rightarrow 0.7 - 0.60942 \le \beta \le 0.7 + 0.60942$

$\Rightarrow 0.09058 \le \beta \le 1.30942 \Rightarrow 0.1 \le \beta \le 1.3$

$\Rightarrow$ 95% confidence interval for β is $[0.1, 1.3]$

3) F-test on poverty & minority for crime. $\alpha = 0.05$

(a) H_0: $\beta_1 = \beta_2 = 0$

H_A: at least one of the β_1, β_2 is not 0

from the code output given, p-value $= 4.249 \times 10^{-9} < \alpha$

Then p-value is significant. We reject H_0, and conclude that at least one of poverty and minority will affect Crime rate.

(b) t-test on poverty & crime $\alpha = 0.05$

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

From code output, we have p-value is $0.075 > \alpha$

p-value is not significant. We fail to reject H_0 and conclude that poverty itself does not significantly affect crime.

(c) t-test on minority & crime $\alpha = 0.05$

$H_0: \beta_2 = 0$

$H_A: \beta_2 \neq 0$

From code output, we have p-value is $6.86 \times 10^{-8} < \alpha$

p-value is significant, we reject the H_0 and conclude that minority itself significantly affect crime

(d) We can see that adjusted R-square = 0.445, and Fstat value is fairly low., thus there is not a strong linear relationship between minority, & poverty on crime. Also we can see that Residual squared error is 18.62 and the fitted model is not very accurate.

(e) $\hat{R}^2 = 0.445 = 44.5\%$,

Thus 44.5% variation in the crime can be attributed to poverty and minority

(f) I would use R-adjusted square instead of R-square since we $\hat{R}^2$ can eliminate the artificial inflation in R^2 when the explanatory variables are co-linear correlated.

(g) $\hat{Y} = 58.670 + 1.195X$

$\hat{Y}$: Crime, X: minority

No poverty since poverty is not significant.

(h) When X = minority = 25

$\hat{Y} = 58.670 + 1.195(25) = 88.545$

(i) Since calculated from the original dataset, we can calculate the table below (covariance table)

	Crime	poverty	Minority
Crime	1	0.56	0.65
poverty	(0.36)	1	0.73
Minority	0.65	(0.73)	1

Yes, since we can see that

① The covariance between minority and poverty is very high

② Since covariance between poverty and crime is very low, they are not very correlated, however since crime and minority has high covariance, thus, the high covariance between poverty and minority may cause the increase in the correlation between poverty and crime

~~Yes~~ minority $\xrightarrow{\text{affect}}$ poverty $\xrightarrow{\text{affect}}$ crime

Thus, there is pair-wise correlation between minority and poverty.