Part A.

b d c a b   a a a a c

a d d b c   b a c c b

2 i). $\begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}_{n\times 1}$

ii) $RSS = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = (Y-HY)'(Y-HY) = (Y'-Y'H')(Y-HY) = Y'Y - Y'HY - Y'H'Y + Y'H'HY$

$\xrightarrow[H'H=H]{H'=H} Y'(I-H)Y$

iii) $J'J = \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}^2 = J$ $J' = J$

$SYY = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 = (Y-JY)'(Y-JY) = Y'Y - Y'JY - Y'J'Y + Y'J'JY = Y'(I-J)Y$

iv) $HX = X \Rightarrow H\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1} \Rightarrow H\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix}_{n\times n} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix}_{n\times n} \Rightarrow HJ = J$

Similarly

$X'H = X' \Rightarrow JH = J$

V). $\hat{Y} = HY$ $J\hat{Y} = JHY = JY = \begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}_{n\times 1}$

also $J\hat{Y} = \begin{pmatrix} \sum_{i=1}^{n}\hat{Y}_i/n \\ \vdots \\ \sum_{i=1}^{n}\hat{Y}_i/n \end{pmatrix}_{n\times 1}$ $\Rightarrow \bar{Y} = \sum_{i=1}^{n}\frac{\hat{Y}_i}{n}$

Vi). $SSreg = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 = (HY-JY)'(HY-JY) = Y'HY - Y'HJY - Y'JHY + Y'JY$

$= Y'HY - Y'JY - Y'JY + Y'JY = Y'(H-J)Y$

Vii) $RSS + SSreg = Y'(I-H)Y + Y'(H-J)Y = Y'(I-J)Y = SYY$

19. For a simple linear regression model,

$$Y = \beta_0 + \beta_1 x + e \qquad e_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

Given the following X observations $\mathcal{A}_X = \{5, 3, 5, 4\}$ and Y observations $\mathcal{A}_Y = \{9, 3, 9, 4\}$. If Helen wants to test the null hypothesis against the alternative hypothesis at 10% significance level,

$$H_0 : \beta_0 = 0 \qquad H_1 : \beta_0 < 0$$

then what are the correct p-value and conclusion?

a) $p = 0.2994$ and cannot reject Null

b) $p = 0.1472$ and cannot reject Null

c) $p = 0.0736$ and reject Null

d) $p = 0.1472$ and reject Null

20. For the multiple regression model $Y = X\beta + e$, $e \sim N(0, \sigma^2)$, Let $\hat{\beta} = (X'X + kI)^{-1}X'Y$. What is $E(\hat{\beta})$?

a) β

b) $\beta - k(X'X + kI)^{-1}\beta$

c) $\beta - k(X'X + kI)\beta$

d) $\beta - kX'X\beta$

Part B: Write down the calculation steps and answers on the blank area. (50 Marks)

1. (5 marks) In a multiple regression analysis, MSE = 30, number of observation = 58, number of predictor = 4, and TSS = 2000. What is the R^2 of the regression?

① ② ③

$$R^2 = \frac{SSreg}{SYY} = \frac{SYY - MSE \cdot (n-(p+1))}{SYY} = 0.205$$

3. (25 marks) The results of thirteen countries in the 2014 Asian Olympic game is given below:

Country	Gold (G)	Silver (S)	Bronze (B)
China	151	108	83
Korea	79	71	84
Japan	47	76	77
Kazakhstan	28	23	33
Iran	21	18	18
Thailand	12	7	28
North Korea	11	11	14
India	11	10	36
Taiwan	10	18	23
Qatar	10	0	4
Uzbekistan	9	14	21
Bahrain	9	6	4
Hong Kong	6	12	24

i) (5 marks) Plot a scatterplot of G against S. Which country would have the most influence on the regression fit?

Plot

X-axis = S (1)

Y-axis = G (1)

points (2)

China

Japan

(0.5)

- largest x-value
- fitting with/without makes difference

(0.5)

ii) (5 marks) It is found that $SXY = 15614.56$, $SXX = 12920.78$, $SYY = 20686.73$

$$\frac{1}{n}\sum G_i = 31.1,\ \frac{1}{n}\sum S_i = 28.8,\ \frac{1}{n}\sum G_i^2 = 2558.5,\ \frac{1}{n}\sum S_i^2 = 1869.5,\ \frac{1}{n}\sum S_iG_i = 2096.8$$

Use these results to estimate the regression model $G = \beta_0 + \beta_1 S + e$, $e \sim N(0, \sigma^2)$. Plot this line on the above scatter plot diagram. Is this fitting appropriate?

$\hat{\beta}_1 = 1.1549$ (1)

$\hat{\beta}_0 = -2.1599$ (1)

$\hat{\sigma}^2 = 241.29042$ (1)

Plot (1)

Fitting Appropriate? (1)

iii) (5 marks) Is there linear relationship between the number of gold and silver medals at 5% significance level?

$$F = \frac{SSreg/1}{RSS/(n-2)} = 74.73390 > 4.84 = F_{0.05}(1,11)$$ (2) (1) (1)

$\therefore$ Regression is effective / H_0 is rej. (2)

or $|t| = 8.64488 > 2.201 = t_{0.025}(11)$ (1) (1)

formula (2)

iv) (5 marks) Write down the equation governing the 95% confidence ellipse for the parameter (β_0, β_1).

$$\frac{(\hat\beta-\beta)'(X'X)(\hat\beta-\beta)}{2\hat\sigma^2} \le 3.982 = F_{0.05}(2,11)$$ (2) (1)

$$\beta_0^2 + 1869.5\beta_1^2 + 57.6\beta_0\beta_1 - 62.20\beta_0 - 4193.76\beta_1 + 432.86 \le 0$$ (2)

v) (5 marks) If Country A got 20 silver medals, what is the 90% prediction interval of its number of gold medal?

$\hat\sigma = 15.53353$, $\quad \text{sepred}(y|x^*) = \hat\sigma\sqrt{1+\frac{1}{13}+\frac{(s_*-\bar s)^2}{SSS}} = 1.0405 \cdot 15.53353$ (1)

$= 16.16271$ (1)

$\tilde y_* = -2.1599 + 1.1549 \cdot 20 = 20.9381$ (1)

$\tilde y_* \pm t_{0.05}(11)\ \text{sepred}(y|x^*)$ (1)

$= 20.9381 \pm 1.796 \cdot 16.16271$

$= 20.9381 \pm 29.02823$

$= [\,0\,, \quad 49.9663\,]$ (1)

End of paper