UNIVERSITI TUNKU ABDUL RAHMAN

ACADEMIC YEAR 2019/2020

JANUARY 2020 TRIMESTER

FINAL ASSESSMENT

ANSWER SCRIPT

Candidate is required to fill in ALL the information below:

Name : (as stated in Student Identity Card)	Ngu Yi Hui		
Faculty /Institute/ Centre:	FSc	Programme :	Statistical Computing And Operations Research
Index No. (in numbers) :	A00082DBSCF	Index No. (in words) :	A Zero Zero Zero Eight Two DBSCF
Course Code :	UDPS2223	Course Description :	Applied Regression Analysis
Submission Date :	5th MAY 2020	Submission Time :	9:30am - 12pm

QUESTION NUMBER	**FOR EXAMINER'S USE ONLY**	
	MARKS	
	Internal	**External**
Q1		
Q2		
Q3		
Q4		
TOTAL MARKS		

UNIVERSITI TUNKU ABDUL RAHMAN

Wholly owned by UTAR Education Foundation
(Co. No. 578227-M)
DU012(A)

DECLARATION STATEMENT

I, Ngu Yi Hui (Name), Student ID No. 18ADB01438, hereby solemnly and fully declare and confirm that during my programme of study at Universiti Tunku Abdul Rahman, I shall abide and comply with all the rules, regulations and lawful instructions of Universiti Tunku Abdul Rahman and endeavour at all times to uphold the good name of the University.

I hereby declare that my submission for this Final Assessment is based on my original work, not plagiarised from any source(s) except for citations and quotations which have been duly acknowledged. I am fully aware that students who are suspected of violating this pledge are liable to be referred to the Examination Disciplinary Committee of the University.

Programme:	Statistical Computing And Operations Research
(Digital) Signature:	HUI
Student's I.C / Passport No.:	991110-14-6378
Index No:	A00082DBSCF
Date of Submission:	5^{th} MAY 2020

Q1.

(a) Define x = amount of protein intake per day (in grams)

y = diastolic blood pressure.

$\Sigma x = 762$ $\Sigma y = 66.7$ $\bar{y} = 7.411$

$\Sigma x^2 = 64868$ $\Sigma xy = 5758.2$ $\bar{x} = 84.667$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 64868 - \frac{762^2}{9} = 352$$

$$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 5758.2 - \frac{(762)(66.7)}{9} = 110.933$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{110.933}{352} = 0.315$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 7.411 - 0.315(84.667) = -19.259$$

$\therefore$ Linear Regression : $\hat{y} = -19.259 + 0.315x$

(b) When no intercept term, to minimize SSE

$$\Sigma(y_i - \beta_i x_i)^2 = \Sigma y_i^2 - 2\Sigma y_i x_i \beta_i + \Sigma \beta_i^2 x_i^2$$

$$\frac{\partial}{\partial \beta} = -2\Sigma y_i x_i + 2\Sigma x_i^2 \beta = 0$$

$$\beta \Sigma x^2 = \Sigma xy$$

$$\hat{\beta} = \frac{\Sigma xy}{\Sigma x}$$

Obtained $\Sigma xy = 255274$, $\Sigma x^2 = 125068$

$$\hat{\beta} = \frac{255274}{125068} = 2.041 = a_1$$

$\therefore y = 2.041x$

For example, we fit a regression line which the dependent variable is height of children and the independent variable is age of children. When the age of children is equal to zero, his height should not be a negative value and should be zero value. Thus, we will fit a regression line pass through origin which the model equation $y = a_1 x$ is adequate.

Q2.

(a) Define x_1 = number of bidders

x_2 = age of the clocks

y = auction price

(i) $\hat{y} = -1338.951 + 85.953x_1 + 12.741x_2$

(ii) $R^2 = \frac{SSR}{TSS} = \frac{4283062.960}{4799789.5} = 0.8923$

$R_a^2 = 1 - \frac{n-1}{n-k-1}(1-R^2) = 1 - \frac{31}{29}(1-0.8923) = 0.8849$

$R^2 = 0.8923$, shows that 89.23% of the variation of the auction price (y) is interpreted by number of bidders (x_1) and age of clocks (x_2) in the model.

(iii) $H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$

$t^* = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{12.741}{0.905} = 14.078$

$t_{0.025,29} = 2.045$

Since $|t^*| > t_{0.025,29}$, reject H_0.

∴ At $\alpha = 0.05$, there is significant evidence to conclude that the age of clock (x_2) is related to auction price (y) in the model, when number of bidders (x_1) is already in the model.

(b) The coefficient estimators should be standardized, because the sample size is small ($n=17$) and the unit of x_2 seems larger than x_1.

Besides that, since the degree of freedom (df) of sum square error (SSE) is 15, means that the number of variable in the model is only one (17-15-1). Thus, we will choose a better variable between the two variables which explained the more and is significant to the dependent variable.

Q3.

(a) Procedure :

① Initially, there is no regressor variable in the model except intercept.

② We calculate the t-statistics and p-values for all the independent variables. Then select the variable that has the largest t-statistics, which is x_3 in this case. We compared its t-value with t_{IN}, and found that $|t| = 7.23 > t_{IN} = 1.677$. Therefore, we add x_3 into our model.

③ Next, we calculate the t-statistics and p-values for the rest independent variables again. x_1 has the largest t-statistics among the others, and its t-value is larger than t_{IN}, given that x_3 is already in the model. $|t| = 7.25 > t_{IN} = 1.678$. Thus, we add x_1 into our model.

④ Furthermore, we calculate the t-statistics and p-values for the rest independent variables again. x_2 has the largest t-statistics among the others, and its t-value is higher than t_{IN}, given that x_3 and x_1 already in the model. $|t| = 6.58 > t_{IN} = 1.679$. Hence, we add x_2 into the model.

⑤ Then, we calculate the t-statistics and p-values for the rest independent variables again. We found that both t-values of x_4 and x_5 are smaller than $t_{IN} = 1.679$. So, x_4 and x_5 are not allowed to be added into the model. We stop the procedure and obtain the final model.

∴ Final model: $\hat{y} = 2.331 + 3.288x_1 + 3.251x_2 + 1.038x_3$

Q3.

(b) (i) Model A: $\hat{y} = a + bX_1$

Model B: $\hat{y} = a + bX_1 + cX_2 + dX_3$

Firstly, since Model B includes more predictors than Model A, it will definitely explain more variation of y compared with Model A.

Furthermore, $R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$

$SSE_{Model\ A} = \Sigma(Y_i - a - bX_1)^2$

$SSE_{Model\ B} = \Sigma(Y_i - a - bX_1 - cX_2 - dX_3)^2$

Since SSE for Model B minus more contents than Model A, the $SSE_{Model\ B}$ smaller than $SSE_{Model\ A}$. And cause the $R_A^2 \leq R_B^2$.

(ii) If predictors x_2 and x_3 are having a negative relationship with y and the predictor x_1 is having a positive relationship with y, the b in the model B can be increased.

Q4.

(a) $\Sigma h_{ii} = 4$

$h_{44} = 4 - (0.231 + 0.235 + 0.711 + 0.173 + 0.4 + 0.882 + 0.213 + 0.499 + 0.428)$

$= 0.228$

Determine leverage points:

cut-off point $= \frac{2p}{n} = \frac{2(4)}{10} = 0.8$

$h_{77} > 0.8$

∴ Point 7 is a leverage point.

Determine influence points:

Point 1: $P(F(4,6) \leq 0.012) = 0.00037$

Point 2: $P(F(4,6) \leq 0.012) = 0.00037$

Point 3: $P(F(4,6) \leq 0.255) = 0.10347$ (>0.1)

Point 4: $P(F(4,6) \leq 0.011) = 0.00031$

Point 5: $P(F(4,6) \leq 0.008) = 0.00017$

Point 6: $P(F(4,6) \leq 0.033) = 0.0027$

Point 7: $P(F(4,6) \leq 1.9) = 0.77017$ (>0.5)

Point 8: $P(F(4,6) \leq 0.01) = 0.00026$

Point 9: $P(F(4,6) \leq 0.06) = 0.00843$

Point 10: $P(F(4,6) \leq 0.039) = 0.00372$

∴ Point 3 is an influence point with small influence.
Point 7 is an influence point with major influence.

(b) No, because the plot shows not much problem on linearity, normality and constant variance assumptions.