APPLIED REGRESSION **TEST-II** **SPRING 2020**

Name: _____ **Instructor: D. Kushary**

Information and Instruction:

(1) Test is open book and open notes and you can you computer.

(2) Time is 1 hour 30 minutes (strictly).

(3) YOUR ANSWER PAPER SHOULD BE A PDF FILE AND FILE-NAME SHOULD BE TEST2_LASTNAME_FIRSTNAME.PDF

(4) YOUR R-PROGRAM NEEDS TO BE SUBMITTED IN CANVAS OR VIA EMAIL - FILE NAME SHOULD BE TEST2_LASTNAME_FIRSTNAME.R)

1. A researcher is interested in predicting the time requires to service vending machines including traveling time. The variables impacting the total time were (1) Number of cases stocked (X_1) and the distance traveled (X_2). Twenty five data points were used and a multiple linear regression was fitted and the following results were found:

$$\hat{Y} = -2.3412 + 1.6159 \times X_1 + 0.0144 \times X_2 \quad \& \quad S = 3.2595$$

During the residual analysis 9th point seems to have some issues. Given the following:
$y_9 = 79.24, \hat{y}_9 = 71.8203, h_{99} = 0.4983$

(a) Find R-student for the 9th observation and judge whether it should be considered an outlier or not using the test at 5% level.

To find R-Student, we first need to find $S^2_{(i)}$ for i=9. Now

$$S^2_{(i)} = \frac{(n-p)MSE - e_i^2/(1-h_{ii})}{n-p-1} = \frac{(25-3) \times (3.2595)^2 - (79.24 - 71.8203)^2/(1-0.4983)}{25-3-1} = 5.90$$

Now

$$t_9 = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}} = \frac{(79.24 - 71.8203)}{\sqrt{5.90 \times (1-0.4983)}} = 4.31$$

As $|t_9| = 4.31 > t_{0.05/2,25-3-1} = 2.080$, we consider it as an outlier.

b) Find the Cook's D for the 9th point ? (10 points)

To find Cook's D, we find the studentized residual first and then Cook's D.

$r_9 = \frac{e_9}{\sqrt{MSE \times (1-h_{ii})}} = \frac{(79.24-71.8203)}{\sqrt{3.2595^2 \times (1-0.4983)}} = 3.2138$

Then Cook's D $D_9 = \frac{r_9^2}{p} \times \frac{h_{99}}{1-h_{99}} = \frac{3.2138^2}{3} \times \frac{0.4983}{(1-0.4983)} = 3.42$

c) Find also the standardized Press residual for 9th observation.

Standardized Press residual = Studentized residual = $r_9 = 3.2138$ (from part (b))

2. A researcher fitted a full MLR model (Model-1) using 13 observations where the response, y, is the yield of what in kg/ha. The regressors are:
X_1 : the amount of rain in mm for the period October to April.
X_2 : is the number of days in the growing season.
X_3 : is the amount of rain in mm during the growing season.
X_4 : is the water use in mm for the growing season.
X_5 : is the pan evaporation in mm during the growing season.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	$\hat{\beta}_i$
X_1	1	11517225.11	11517225.11	20.13	0.0028	8.72
X_2	1	134897.52	134897.52	0.24	0.6421	6.57
X_3	1	32645.77	32645.77	0.06	0.8180	-10.54
X_4	1	990824.49	990824.49	1.73	0.2296	13.03
X_5	1	486488.95	486488.95	0.85	0.3871	-3.25
Residuals	7	4004641.24	572091.61			

(a) Is the above model significant at 5% level?

To test the model, we set the hypothesis as

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ vs H_1 : at least one of them is non-zero

$F - stat = \frac{13162081.84/5}{4004641.24/7} = 4.60 \ > \ F_{0.05,5,7} = 3.97$ (p-value = 0.0353)

Hence we reject the null hypothesis and conclude that the model is significant at 5% level.

(b) Next, the researcher tried the Model-2 $Y = \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_3 +$ Error, Is Model-2, significant at 5% level?

To test the Model-2, we set the hypothesis as

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs H_1 : at least one of them is non-zero

$F - stat = \frac{(11517225.11+134897.52+32645.77)/3}{(990824.49+486488.95+4004641.24)/9} = 6.39 \ > \ F_{0.05,3,9} = 3.86$ (p-value = 0.0131)

Hence we reject the null hypothesis and conclude that the model is significant at 5% level.

(c) Is variable X_3 significant in Model-2 at 5% level?

To test X_3 in Model-2, we set the hypothesis as

$H_0 : \beta_3 = 0 \; vs \; H_1 : \beta_3 \neq 0$

$F - stat = \frac{32645.77)/1}{(990824.49+486488.95+4004641.24)/9} = 0.0536 \; < \; F_{0.05,1,9} = 5.12$ (p-value = 0.8221)

Hence we do not reject the null hypothesis and conclude that X_3 is not significant at 5% level.

(WRITE APPROPRIATE R-PROGRAM TO SOLVE Q3 AND SUBMIT IT)

3. The attached data set has 5 independent variables (x1-x5) and one dependent variable (y). Use forward ($\alpha = 0.05$), backward($\alpha = 0.05$)

(a) Let say X_m is the last variable added in the model in the forward selection procedure above. Now, if α value is changed from the 0.05 to 0.01 in forward selection, X_m will still be added to the model (no need to run with the new alpha)?

The last variable added to the model is X_3. As the p-value from the R-output for X_3 is $0.033 > 0.01$, the X_3 would not have been added.

(b) Looking at the final model chosen by the backward elimination, which variable is the is the next candidate to be dropped if we change the α value and what value of α is that.

The next variable to be dropped is X_3 as the p-value is the highest among them. As the p-value for X_3 is 0.0219, for any $\alpha < 0.0219$ the variable X_3 will be dropped.

(c) Did the forward and the backward choose the same model? If not, which one you prefer and why.

COMPARISON OF THE MODELS

	(Intercept)	x1	x2	x3	x4	SSE	RSQ	adjR2	Cp	BIC
1	1.00	0.00	0.00	1.00	0.00	169.52	0.18	0.15	15.89	1.40
4	1.00	1.00	1.00	1.00	1.00	91.90	0.56	0.47	5.00	-4.25

It is easily seen the model with four variables is better in every respect (e.g adj-Rsqr, Cp).