APPLIED REGRESSION **TEST-II** **SPRING 2022**

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

Information and Instruction:

(1) Test is open book and open notes and you can use your 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 - FILE NAME SHOULD BE TEST2_LASTNAME_FIRSTNAME.R)

1. A study was conducted to understand the relationship of amount of body fat (Y) to several possible predictor variables, based on a sample of 20 healthy females 25- 34 years old. The possible predictor variables are triceps- skinfold thickness (X1), thigh circumference (X2), and midarm-circumference (X3).

   The attached data set has 3 independent variables (X1-X3) and one dependent variable (Y). , backward($\alpha = 0.05$)

   (a) Use forward selection procedure (use $\alpha = 0.05$) and write down the final estimated model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-23.6345	5.6574	-4.18	0.0006
X2	0.8565	0.1100	7.79	0.0000

   The estimated model is $\hat{Y} = -23.6345 + 0.8565 \times X2$

   (b) Use backward selection procedure (use $\alpha = 0.05$) and write down the final estimated model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.7916	4.4883	1.51	0.1486
X1	1.0006	0.1282	7.80	0.0000
X3	-0.4314	0.1766	-2.44	0.0258

   The estimated model is $\hat{Y} = 6.7916 + 1.006 \times X1 - 0.4314 \times X3$

(c) Run the model with all three independent variables and write down the ANOVA table and use the ANOVA table to test whether X2 is significant or not at 5%. (You can run only one model (ie. ONLY ONE LM-STATEMENT IN R-program) and you can only use the ANOVA TABLE TO ANSWER THE QUESTION)

It is only possible to test for X2 from ANOVA table if X2 is last variable in model. Hence you need to write the code as Mdl=lm(y $\sim$ X1 + X3 + X2, data=data1)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	352.27	352.27	57.28	0.0000
X3	1	37.19	37.19	6.05	0.0257
X2	1	7.53	7.53	1.22	0.2849
Residuals	16	98.40	6.15		

To test significance of X2 we test $H_0 : \beta_2 = 0 \quad vs \quad H_1 : \beta_2 \neq 0$

F-stat = 1.22 and p-value = 0.2849 $> \alpha = 0.05$

Hence we do not reject the null hypothesis at 5% level.

2. A hospital administrator wished to study the relationship between patient satisfaction (Y) and patient's age (X1, in years), severity of illness (X2, an index), and anxiety level (X3 , an index). The ANOVA tables is given below

ANOVA TABLE

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	8275.39	8275.39	81.80	0.0000
X3	1	763.42	763.42	7.55	0.0088
X2	1	81.66	81.66	0.81	0.3741
Residuals	42	4248.84	101.16		

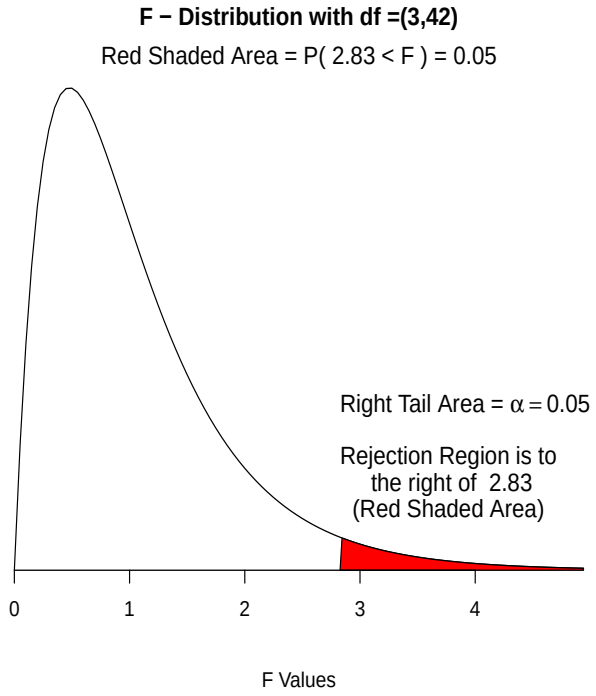
(a) Is the above model significant at 5% level? (Graph Needed)

Ans: We set the hypothesis as
$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad vs$
H_1 : at least one of them is non-zero

$$\text{F-stat} = \frac{(8275.39+763.42+81.66)/3}{4248.84/42} = 30.05$$

But $F_{0.05,3,42} = 2.827$(qf(0.95,3,42) in R)

Hence we reject H_0 as
$F - Stat = 30.05 > F_{0.05,3,42} = 2.827$,
and conclude that the model is significant at 5% level.



(b) Now if the administrator decides to drop the variable X2 and rerun the model. Will the new model have higher adjusted-R^2? (Compute and compare)

Ans: SST = 13369.31 (addition of all the sum of Squares)

For Model-1 (with all 3 variables) $adj - R^2 = 1 - \frac{(4248.84/42)}{13369.31/45} = 0.6595$

For Model-2 (after dropping X2) $adj - R^2 = 1 - \frac{(4248.84+81.66)/43}{13369.31/45} = 0.661$

Hence the Model-2 has higher $adj - R^2$.

(c) Is X3 significant in the new model (after dropping X2) at 1% level?

We need to test $H_0 : \beta_3 = 0 \quad vs \quad H_1 : \beta_3 \neq 0$

F-stat $= \frac{763.42/1}{(4248.84+81.66)/43} = 7.58 \quad > \quad F_{0.01,1,43} = 7.264$

Hence we reject the null hypothesis and conclude that X3 is significant at 1% level.

Extra credit

(d) In Model-1 (with all 3 variables), addition of X2 on top of X1 and X3 (i.e Model-2) reduced the SSE (of model-2) by 81.66 (from the above table) which was not enough reduction and as a result X2 was insignificant at 5% level. But suppose the reduction in SSE (of model-2) was C and then X2 became significant at 5% level. What is the minimum value of C?

Solution: If X2 would have been significant in Model-1 then

$$F - Stat = \frac{C/1}{(4248.84+81.66-C)/42} > F_{0.05,1,42} = 4.07$$

$$\Rightarrow\ 42C > 4.07 \times (4330.5 - C) \Rightarrow (42 + 4.07) \times C > 4.07 \times 4330.5 \Rightarrow C > 382.57$$

Hence the minimum value of C would have been 382.57.

(e) If the estimated Model-1 is $\hat{Y} = 158.49 - 1.14 \times X1 - 0.442 \times X2 - 13.47 \times X3$
Find a 95% confidence interval of β_2 (the coefficient of X2).

Solution: We see from the table that the F-stat = 0.81 for testing $H_0 : \beta_2 = 0\ vs\ H_1 : \beta_2 \neq 0$
Hence the t-stat $= \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = -\sqrt{F - stat} = -\sqrt{0.81} \Rightarrow \frac{0.442}{S_{\hat{\beta}_2}} = 0.9 \Rightarrow S_{\hat{\beta}_2} = \frac{0.442}{0.9} = 0.4911$
So the 95% CI for β_2 is $\hat{\beta}_2 \pm t_{0.05/2,42} \times S_{\hat{\beta}_2} = -0.442 \pm 2.018 \times 0.4911 = (-1.433, 0.549)$

3. (Continuation of Question 2) Residual analysis was performed on the full model (with independent variables as X1, X2 and X3) and following data came up (Observation number is 17):

$SSE = 4248.84, \quad \bar{y} = 61.56 \quad Y_{17} = 79, \quad e_{17} = 16.61, \quad h_{17} = 0.1195$

$\hat{Y}_{17} = Y_{17} - e_{17} = 79 - 16.61 = 62.39$

(a) Find R-student for the 17th observation and judge whether it should be considered an outlier or not using the test at 5% level.

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

$$S^2_{(i)} = \frac{(n-p)MSE - e_i^2/(1-h_{ii})}{n-p-1} = \frac{4248.8 - 16.61^2/(1-0.1195)}{46-4-1} = 95.9869$$

Now

$$t_9 = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}} = \frac{16.61}{\sqrt{95.9869 \times (1-0.1195)}} = 1.8067$$

As $|t_{17}| = 1.8067 \not> t_{0.05/2,46-4-1} = 2.01$, we do not consider it as an outlier.

(b) Find the Cook's D for the 17th point ?

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

$r_i = \frac{e_i}{\sqrt{MSE \times (1-h_{ii})}} = \frac{16.61}{\sqrt{101.16 \times (1-0.1195)}} = 1.76$

Then Cook's D $D_i = \frac{r_i^2}{p} \times \frac{h_{ii}}{1-h_{ii}} = \frac{1.76^2}{4} \times \frac{0.1195}{(1-0.1195)} = 0.1051$

(c) Find the DFFITS for the 17th point ?

$$DFFITS_i = \left(\frac{h_{ii}}{1-hii}\right)^{1/2} \times t_i = \left(\frac{0.1195}{1-0.1195}\right)^{1/2} \times 1.8067 = 0.6656$$

EXTRA CREDIT

(d) Regardless of your own decision in part (a), the administrator dropped the 17th observation. Can you find the R^2 for the new model after you dropped the 17th observation?

To find R^2 we need to find $SSE_{(17)}$ and $SST_{(17)}$

Note that $\bar{y}_{(17)} = \frac{46\times\bar{y}-y_{(17)}}{45} = \frac{46\times 61.56-79}{45} = 61.1724$

$SST = \sum y_i^2 - n \times \bar{y}^2 \Rightarrow SST + n \times \bar{y}^2 = \sum y_i^2$

Hence $SST_{(17)} = (SST + n \times \bar{y}^2 - y_{(17)}^2) - (n-1) \times \bar{y}_{(17)}^2$

$= (13369.31 + 46 \times 61.56^2 - 79^2) - 45 \times 61.1724^2 = 13058.65$

Now $S_{(17)}^2 = S_{(17)}^2 \times (n-1-k-1) = 95.9869 \Rightarrow SSE_{(17)} = 95.9869 \times 41 = 3935.46$

Hence the $R_{(17)}^2 = 1 - \frac{SSE_{(17)}}{SST_{(17)}} = 1 - \frac{3935.46}{13058.65} = 0.6986$