Name: ____________________________ **Instructor: D. Kushary**

1. An attempt was made to predict the success in the early university years using multiple linear regression model. One measure of success was the cumulative GPA after three years. The independent variables used in this model was high school grades in mathematics (HSM=X_1), science (HSS=X_2) and English (HSE=X_3). Summary of the calculations are given below.

$$n = 35 \quad y_7 = 5 \quad \hat{y}_7 = 3.95 \quad h_{77} = 0.09 \quad S = 0.7 \quad \text{Multiple-R} = 0.779$$

a) Given the above information for the 7th point, calculate R-student and discuss if it should be considered as outlier.

To find R-Student, we first need to find $S^2_{(i)}$ for i=7. Now

$$S^2_{(i)} = \frac{(n-p)MSE - e_i^2/(1-h_{ii})}{n-p-1} = \frac{(35-4)\times(0.7)^2 - (5-3.95)^2/(1-0.09)}{35-4-1} = 0.4659$$

Now

$$t_7 = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}} = \frac{(5-3.95)}{\sqrt{0.4659\times(1-0.09)}} = 1.6125$$

As $|t_7| = 1.6125 < t_{\alpha/2,35-4-1} = 2.042$, we don't call it outlier.

b) Find the adjusted coefficient of determination and interpret it.

$$adj - R^2 = 1-(1-R^2)\frac{(n-1)}{(n-k-1)} = 1-(1-(0.779)^2)\frac{(35-1)}{(35-3-1)} = 0.5688$$

About 56.88% of the variations in score can be explained by this model.

c) Complete the ANOVA table and test for the overall significance of the model. Use $\alpha = 0.05$.

$$S = \sqrt{\frac{SSE}{n-k-1}} = 0.7 \quad \Rightarrow \quad SSE = (0.7)^2 \times (35-3-1) = 15.19$$

$$R^2 = 1 - \frac{SSE}{SST} \quad \Rightarrow \quad (0.779)^2 = 1 - \frac{15.19}{SST} \quad \Rightarrow \quad SST = \frac{15.19}{1-(0.779)^2} = 38.64$$

Using these values we can create the ANOVA table as

Source	Sum of Squares	Deg of Freedom	Mean Squares	F - Stat	P -value
Regression	23.45	3	7.82	15.95	
Error	15.19	31	0.49		
Total	38.64	34			

To test the overall model we set as

$$H_o : \beta_1 = \beta_2 = \beta_3 = 0 \quad vs \quad H_1 : \text{not all of them are } 0$$

Now F-stat = 15.95 $\quad > \quad F_{0.05,3,30} = 2.92 > F_{0.05,3,31}$ we reject H_o at 5% level. Conclusion: The overall model is significant at 5% level.

d) If a model with only Math and Science (no English) with the same data had standard error $= 0.8$, can you determine if English is a significant variable in the model? Use $\alpha = 0.05$

Answer: F-test for the reduced model can determine that. Let's say Model-1 has all three variables and Model-2 has only Math and Science. Hence $SSE_1 = 15.19$ (from above ANOVA table), and $SSE_2 = (35 - 3) \times (0.8^2) = 20.48$

$H_o : \beta_3 = 0 \quad H_1 : \beta_3 \neq 0$

Hence $\quad F - stat = \frac{(SSE_2 - SSE_1)/1}{MSE_1} = \frac{20.48 - 15.19}{0.49} = 10.796$

As $F - stat = 10.796 > F_{0.05,1,30} = 4.17 > F_{0.05,1,31}$, we reject the null hypothesis and conclude that English is a significant variable in the Model-1 at 5% level.

2. (a) Let say β_i is the coefficient of X_i in the model.

Hence we want to test $H_0 : \beta_1 = \beta_2 = 0 \quad vs \quad H_1$: At least one of them is non zero

So $H_0 : Y = \beta_0 + \beta_4.X_4 + \beta_3.X_3 + e$ (Model-2) and
$H_1 : Y = \beta_0 + \beta_4.X_4 + \beta_3.X_3 + e + \beta_1.X_1 + \beta_2.X_2$ (Model-1)

The statistic to test is $F - stat = \frac{(SSE_2 - SSE_1)/2}{SSE_1/(n-k-1)} = \frac{(0.96+0.41)/2}{14.57/28} = 1.3173$

But $F - stat = 1.3173 \not> F_{0.05,2,28} = 3.34$, hence we fail to reject H_0 and conclude that addition of (X_1, X_2) (at the same time) on top of (X_4, X_3) does not improvement the model significantly at at 5% level?

(b) Observe that SST = (150.24 + 4.84 + 0.96 + 0.41 + 14.57) = 171.02 and SSE for Model-2 is $SSE_2 = 0.96 + 0.41 + 14.57 = 15.94$

Hence the $adj - R^2 = 1 - \frac{SSE_2/30}{SST/32} = 1 - \frac{0.5313}{5.344} = 0.9006$.

(c) To test whether model-2 is significant or not we set the hypothesis as follows:

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

$F - stat = \frac{SSR_2/2}{SSE_2/30} = \frac{(150.24+4.84)/2}{0.5313} = 145.94$

As $F - stat = 145.94 > F_{0.05,2,30} = 3.32$ we reject H_0 and conclude that the model-2 is significant at 5% level.

(d) To test whether x_3 is significant or not in model-2 we set up the test as

$H_0 : \beta_3 = 0$(say Model-3) $\quad vs \quad H_1 : \beta_3 \neq 0$

$F - stat = \frac{(SSE_3 - SSE_2)/1}{SSE_2/30} = \frac{(4.84)/1}{0.5313} = 9.1097$

As $F - stat = 9.1097 > F_{0.05,1,30} = 4.17$ we reject H_0 and conclude that the variable x_3 is significant at 5% level in model-2.

3. (a) X_5 is the first variable chosen because it has the highest correlation with Y. Another way to look at it among all the variables, it has the lowest SSE among all the SLR models with Y and any of the X's. (As seen in the program).

(b) First variable to be dropped is X_8 as it has the highest p-value.

(c) $R^2_{prediction} = 1 - \frac{PRESSSTAT}{SST} = 1 - \frac{0.0186}{2489.522} = 0.9999$ (Using R)

Remark: Observe that there is a different library(mixlm as oppose to MASS) and different coding used in the R-program comparing to the Lecture-6 R-program for selection. The difference is package 'MASS' uses the AIC criterion and a testing is needed at the last step as it does not use the same stopping rule of significance of the variable.