Basic Information and Instructions for Test - II

Maximum Total Points - 100

Time 1:30 minutes

Remark: Need calculators, tables and computer with R.

The test is open book and open notes (or any printed materials)

You will be required to submit the exam paper and email the r-program associated with it.

Exam Materials:

Test will be based on the materials covered in class. Primary coverage are from Lecture Series 4, 5 and 6 along with the other supported materials like examples.

Remark: Answer key along with the expected correction rules will be provided for the following sample test.

Applied Regression **Test - 2 - Sample** **Spring**

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.

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

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

Source	Sum of Squares	Deg of Freedom	Mean Squares	F - Stat	P -value
Regression					
Error					
Total					

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

2. A data file regarding different aircrafts contains 33 observations with the following variables: y = operating cost (dollars-in 1000 per hour), x_1 = number of seats, x_2 = speed (mph), x_3 = flight range (miles), and x_4 = fuel consumption (gallons per hour). Considering operating cost per hour as the dependent variable, the full model (call it model-1) was tried and its output (from R) is given below.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x4	1	150.24	150.24	288.72	0.0000
x3	1	4.84	4.84	9.31	0.0049
x1	1	0.96	0.96	1.85	0.1850
x2	1	0.41	0.41	0.79	0.3816
Residuals	28	14.57	0.52		

(a) Now the researcher wants consider a second model (call it model-2)by dropping both (x_1, x_2) at the same time. Write the appropriate hypothesis and conduct an appropriate test to check whether addition of (x_1, x_2) (at the same time) on top of (x_4, x_3) was a significant improvement at 5% level?

(b) Find the $adj - R^2$ for model-2.

(c) Is the model-2 significant at 1% level?

(d) Is variable x_3 is significant in model-2 at 5% level?

3. The attached data set has 10 independent variables (X1-X10) and one dependent variable (Y). Use forward ($\alpha = 0.05$), backward($\alpha = 0.10$) and stepwise (alpha.enter=0.5, and alpha.remove=0.10) selection procedure to find the best models.

(a) Which variable is added at the first step of the forward selection procedure and why?

(b) Which variable is dropped at the first step of the Backward Selection process and why?

(c) Calculate the $R^2_{prediction}$ for the final model and show that it is equal to the output from R.