KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DEPARTMENT OF MATHEMATICS AND STATISTICS
Term 162

STAT 212 BUSINESS STATISTICS II
Third Major Exam
Allowed time 75 min
Wednesday May 10, 2017

Name:_____ ID #:_____ Section #:_____ Srl #:__

Important Note:

1) You must **show all work** to obtain full credit for questions on this exam.
2) **DO NOT round** your answers at each step. Round answers only if necessary at your final step to **4 decimal places**.

Question No	Full Marks	Marks Obtained
Q1	18	
Q2	9	
Q3	15	
Q4	18	
Total	60	

Question One (3+5+2+1+2+2+3 points):

A real estate builder wishes to determine how house size (House) is influenced by family income (X1=Income) and family size (X2=Size). House size is measured in **hundreds of square feet** and income is **measured in thousands of dollars**. The builder randomly selected 50 families and ran the multiple regression. Partial Computer output is provided below:

```
Analysis of Variance

Source              DF           SS           MS          F          P
Regression          ?      37043.3236    18521.6618       ?        0.000

Residual Error      ?      14487.7627      308.2503

Total              49      51531.0863


Predictor       CoefSE          Coef             T          P
Constant        -5.5146         7.2273        -0.7630     0.4493

Income          0.4262          0.0392        10.8668     0.0000

Size            5.5437          1.6949         3.2708     0.0020
```

Multiple R=0.8479, R-Square=0.7189, Adjusted R-Square=0.7069,

Standard Error=17.5571, SSR(X1 |X2)=36400.6326, and SSR(X2 |X1)=3297.7917

a. Find the missing values in the output?

b. Test the claim "*Variable X2 does not significantly improve the model after variable X1 has been included*"

c. Which of the independent variables in the model is significant at the 5% level? Why?

d. Which of the following values for the level of significance is the smallest for which at most one explanatory variable is significant individually? (0.001, 0.01, 0.025, 0.05)

e. What annual income (in thousands of dollars) would an individual with a family size of 4 need to attain a predicted 10,000 square foot house?

f. At the 0.01 level of significance, what conclusion should the builder draw regarding the inclusion of Size in the regression model?

g. What fraction of the variability in house size is explained by income and size of family?

Question Two (5+4 *points*):

The following is the ANOVA summary table for a multiple regression model with two independent variables:

Source	d.f.	SS	MS	F
Regression	2	30		
Error	10	120		
Total	12	150		

If $SSR(X_1) = 20$ and $SSR(X_2) = 15$, answer the following:

a. Determine whether there is a significant relationship between Y and each independent variable at the 0.05 level of significance.

b. Compute the coefficients of partial determination for $X1$ given $X2$ is already included in the model and interpret its meaning.

Question Three (5+3+2+5 *points*):

A publishing company is attempting to develop a model that it can use to help predict textbook sales for books it is considering for future publication. The marketing department has collected additional data on six variables from a random sample of 15 books. These variables are as follows:

Y: Number of Volumes sold (1000's).

X1: Number of Pages.   X2: Number of Competing books.

X3: Advertising Budget ($1000's).   X4: Age of the Author.

X5: Production Expenditures ($1000's).   X6: Number of Reviewers.

Correlations: Y, X1, X2, X3, X4, X5, X6

	Y	X1	X2	X3	X4	X5
X1	0.622					
	0.013					
X2	0.355	0.501				
	0.194	0.057				
X3	0.62	0.091	0.384			
	0.014	0.746	0.158			
X4	0.485	-0.019	-0.113	0.265		
	0.067	0.947	0.687	0.340		
X5	0.896	0.67	0.27	0.539	0.438	
	0.000	0.006	0.331	0.038	0.103	
X6	0.66	0.377	0.291	0.355	0.528	0.737
	0.007	0.166	0.292	0.194	0.043	0.002

Best Subsets Regression: Y versus X1, X2, X3, X4, X5, X6

Response is Y

Vars	R-Sq	R-Sq(adj)	C-p	S	X1	X2	X3	X4	X5	X6
1	80.2	78.7	1.4	23.600					X	
1	43.6	39.3	24.4	39.868						X
2	82.9	80.0	1.7	22.854			X		X	
2	81.6	78.6	2.5	23.676		X			X	
3	84.1	79.8	3.0	23.003	X		X		X	
3	83.8	79.4	3.1	23.199			X	X	X	
4	87.2	82.1	3.0	21.640	X		X	X	X	
4	85.0	79.0	4.4	23.456		X	X	X	X	
5	87.2	80.1	5.0	22.799	X	X	X	X	X	
5	87.2	80.1	5.0	22.808	X		X	X	X	X
6	87.3	77.7	7.0	24.166	X	X	X	X	X	X

From the Minitab output above, answer the following:

a. Do you think that Number of pages and Production expenditures are directly (**positively**) related? **Test using 2.5 % level of significance and provide your final answers in the boxes below.**

The test hypotheses are:	H_0: H_A:
The p-value for the test is:	
The decision rule is:	If Reject H_0.
The decision is:	

b. Is there any potential collinearity problem with the regression analysis? **Explain your answer in detail by supporting it with available MINITAB analyses results.**

A potential collinearity exists?	**please circle:** Yes No
Explanation:	

c. If you are going to fit a regression model using the forward selection method, what is the first predictor to be used? **Why?**

The first predictor to be used is:	
Explanation:	

d. Based on the Best Subset Regression, what is the best model to be selected? **Justify your selection.**

The group of predictors to be used is:	
Explanation:	R^2-adj = which is
	C-p = which is
	S_ε = which is

Question Four (1+5+2+2+3+2+4 *points*):
In the previous question, the variables analyzed were:

Y: Number of Volumes sold (1000's).
X1: Number of Pages.   X2: Number of Competing books.
X3: Advertising Budget ($1000's).   X4: Age of the Author.
X5: Production Expenditures ($1000's).   X6: Number of Reviewers.

In addition to these variables, the type of book sold is incorporated in the model. The company produces only three mutually exclusive types of books which are: Biology, Statistics, or Physics, as follows:
X7: 1=If a Biology book, 0 otherwise.
X8: 1=If a Statistics book, 0 otherwise.

Regression Analysis: Y versus X1, X2, X3, X4, X5, X6, X7, X8

```
The regression equation is
Y = - 104 + 0.123 X1 - 0.55 X2 + 1.16 X3 + 1.34 X4 + 0.58 X5 + 1.61 X6 - 20.9 X7 - 29.6 X8

Predictor         Coef      SE Coef            T          P          VIF
Constant       -103.69        37.57        -2.76      0.033
X1             0.12302      0.09192         1.34      0.229          8.4
X2              -0.553        2.864        -0.19      0.853          4.0
X3              1.1649       0.7447         1.56      0.169          4.4
X4              1.3393       0.7707         1.74      0.133          2.0
X5               0.580        1.011         0.57      0.587         15.5
X6               1.613        6.837         0.24      0.821          5.8
X7              -20.95        19.09        -1.10      0.315          2.7
X8              -29.56        16.21        -1.82      0.118          1.6

S = 22.27        R-Sq = 91.9%      R-Sq(adj) = 81.0%


Analysis of Variance

Source               DF            SS           MS          F          P
Regression            8       33671.6       4209.0       8.48      0.009
Residual Error        6        2976.8        496.1
Total                14       36648.4
```

From the Minitab outputs above, answer the following questions:

a. How much of the total variation in the number of volumes sold is explained by the variation in these eight explanatory variables?

The amount of the total variation in the number of volumes sold is explained by the variation in these eight explanatory variables =	

b. Would you conclude that the model is significant at 5% level of significance? Conduct the **test and provide your final answers** below to arrive at this conclusion.

The test hypotheses are:	H0:	
	HA:	
The test statistic value =		
The critical value(s) =		
The p-value for the test =		
The decision rule (using the p-value approach) is:	If	Reject H0.
The decision is:		

c. Which of the predictors can be concluded to be significant in explaining the variation in the number of the volumes sold? Explain. **Using $\alpha = 0.05$.**

The predictor(s) that is significant in explaining the variation in the number of the volumes sold is(are)	
Explanation	

d. How can you explain the contradiction between the results in part (b) and part (c) above.

Explanation	

e. Find 90% confidence interval for the regression coefficient of Advertising Budget, and interpret this C.I.

A 90% C.I. for the regression coefficient of Advertising Budget is	
Interpretation	

f. Based on your solution in part (e), can you conclude that the Advertising Budget is significant? Why?

Is the Advertising Budget significant? (circle your answer)	Yes No
Justification	

g. Estimate the number of volumes sold for a book with each of the following given information and indicate which book the predicted value is for

(X1, X2, …, X8) = (204, 3, 30000, 50, 34000, 6, **1, 1**)	
Which book? (Biology, Statistics, or Physics)?	
(X1, X2, …, X8) = (204, 3, 30000, 50, 34000, 6, **0.5, 1**)	
Which book? (Biology, Statistics, or Physics)?	
(X1, X2, …, X8) = (204, 3, 30000, 50, 34000, 6, **1, 0**)	
Which book? (Biology, Statistics, or Physics)?	

With My Best Wishes