Department of Mathematics and Statistics
Semester 132

STAT310 Final Exam Thursday, May 22, 2014

Name: ID #:

Question	Full Marks	Marks Obtained
Q1	10	
Q2	14	
Q3	10	
Q4	6	
Computer part		
Q5	20	
Q6	20	
Total	80	

Question One: (3+3+4=10 pts)

Suppose X_1 is a numerical variable and X_2 is a dummy variable and the regression equation for a sample of n=20 is: $\hat{Y} = 6 + 4\,X_1 + \,2\,X_2$

a. Interpret the regression coefficient associated with variable X_1.

b. Interpret the regression coefficient associated with variable X_2.

c. Suppose that the test statistic for testing the contribution of variable X_2 is 3.27. At the 0.05 level of significance, is there evidence that variable X_2 makes a significant contribution to the model?

Question Two: (9+5=10 pts.)

The real data set in this question first appeared in Hald (1952). Interest centers on using variable selection to choose a subset of the predictors to model Y . Throughout this question we shall assume that the full model below is a valid model for the data

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \in \quad \ldots\ldots 1$$

Output from Minitab associated with different variable selection procedures based on model (1) appears on the following pages:

a. Identify the optimal model or models based on R^2 adj , AIC, AIC_C, BIC from the approach based on all possible subsets.

b. Which of the models should be taken into consideration using the Mallows' C_p statistic? **Explain?**

Values of R^2-adj , AIC, AIC_C and BIC for the best subset of each size

Subset size	Predictors	R^2-adj	AIC	AIC_C	BIC
1	X4	0.6450	58.8516	61.5183	59.9815
2	X1 , X2	0.9744	25.4200	30.4200	27.1148
3	X1, X2, X4	0 .9764	24.9739	33.5453	27.2337
4	X1, X2, X3, X4	0.9736	26.9443	40.9443	29.7690

Output 1

Regression Analysis: Y versus x4

```
The regression equation is
Y = 118 - 0.738 x4

Predictor     Coef  SE Coef      T      P
Constant   117.568    5.262  22.34  0.000
x4         -0.7382   0.1546  -4.77  0.001

S = 8.96390  R-Sq = 67.5%  R-Sq(adj) = 64.5%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  1831.9  1831.9  22.80  0.001
Residual Error  11   883.9    80.4
Total           12  2715.8
```

Output 3

Regression Analysis: Y versus x1, x2, x4

```
The regression equation is
Y = 71.6 + 1.45 x1 + 0.416 x2 - 0.237 x4

Predictor     Coef  SE Coef      T      P     VIF
Constant     71.65    14.14   5.07  0.001
x1          1.4519   0.1170  12.41  0.000   1.066
x2          0.4161   0.1856   2.24  0.052  18.780
x4         -0.2365   0.1733  -1.37  0.205  18.940

S = 2.30874   R-Sq = 98.2%   R-Sq(adj) = 97.6%

Analysis of Variance
Source          DF       SS      MS       F      P
Regression       3  2667.79  889.26  166.83  0.000
Residual Error   9    47.97    5.33
Total           12  2715.76
```

Output 2

Regression Analysis: Y versus x1, x2

```
The regression equation is
Y = 52.6 + 1.47 x1 + 0.662 x2

Predictor  Coef  SE Coef      T      P    VIF
Constant  52.577    2.286  23.00  0.000
x1        1.4683   0.1213  12.10  0.000  1.055
x2       0.66225  0.04585  14.44  0.000  1.055

S = 2.40634 R-Sq = 97.9% R-Sq(adj) = 97.4%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       2  2657.9  1328.9  229.50  0.000
Residual Error  10    57.9     5.8
Total           12  2715.8
```

Best Subsets Regression: Y versus x1, x2, x3, x4

```
Response is Y

                          Mallows             x x x x
Vars  R-Sq  R-Sq(adj)          Cp       S     1 2 3 4
   1  67.5       64.5       138.7  8.9639           X
   2  97.9       97.4         2.7  2.4063     X X
   3  98.2       97.6         3.0  2.3087     X X   X
   4  98.2       97.4         5.0  2.4460     X X X X
```

Output 4

Regression Analysis: Y versus x1, x2, x3, x4

```
The regression equation is
Y = 62.4 + 1.55 x1 + 0.510 x2 + 0.102 x3 - 0.144 x4

Predictor     Coef  SE Coef      T      P      VIF
Constant     62.41    70.07   0.89  0.399
x1          1.5511   0.7448   2.08  0.071   38.496
x2          0.5102   0.7238   0.70  0.501  254.423
x3          0.1019   0.7547   0.14  0.896   46.868
x4         -0.1441   0.7091  -0.20  0.844  282.513

S = 2.44601   R-Sq = 98.2%   R-Sq(adj) = 97.4%

Analysis of Variance
Source          DF       SS      MS       F      P
Regression       4  2667.90  666.97  111.48  0.000
Residual Error   8    47.86    5.98
Total           12  2715.76
```

Question Three: (10 pts)

An economist is analyzing the incomes of professionals (physicians, dentists, and lawyers). He realizes that an important factor is the number of years of experience. However, he wants to know if there are differences among the three professional groups. He takes a random sample of 125 professionals and estimates the multiple regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

where y = annual income (in $1,000)

x_1 = years of experience

x_2 = 1 if physician and 0 if not

x_3 = 1 if dentist and 0 if not

The computer output is shown below.

THE REGRESSION EQUATION IS $y = 71.65 + 2.07x_1 + 10.16x_2 - 7.44x_3$

Predictor	Coef	StDev	T
Constant	71.65	18.56	3.860
x_1	2.07	0.81	2.556
x_2	10.16	3.16	3.215
x_3	-7.44	2.85	-2.611

S = 42.6 R-Sq = 30.9%

Analysis of Variance

Source of Variation	df	SS	MS	F
Regression	3	98008	32669.333	18.008
Error	121	219508	1814.116	
Total	124	317516		

1. (3 pts.) Do these results allow us to conclude at the 1% significance level that the model is useful in predicting the income of professionals?

2. (3 pts.) Is there enough evidence at the 5% significance level to conclude that income and experience are linearly related?

3. (4 pts.) Is there enough evidence at the 10% significance level to conclude that dentists earn less on average than lawyers?

Question Four: (6 pts.)

Show that $\mathbf{e} = (\mathbf{I} - \mathbf{H}) \in$ where $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\epsilon}$ and $\widehat{\boldsymbol{\beta}} = (\acute{\mathbf{X}}\mathbf{X})^{-\mathbf{1}}\, \acute{\mathbf{X}}\, \mathbf{Y}$.

<u>Question Five:</u> (20 pts.) <u>HOUSE1</u>

You need to develop a model to predict the selling price of houses in a small city, based on assessed value, time in months since the house was reassessed, and whether the house is new (0=no, 1=yes). A sample of 30 recently sold single-family houses that were reassessed at full value one year prior to the study is selected and the results are stored in HOUSE1.

<u>Develop the most appropriate multiple regression model to predict selling price</u>.

- Be sure to Organize your outputs according to the following steps:

1. (3 pt) Fit a regression model that includes all independent variables under consideration and determine the VIF for each independent variable.
    - In case of VIF>5, eliminate the independent variable and proceed to step 3.
2. (1 pt) Perform a best-subset regression with the remaining independent variables.
3. (3 pt) List all candidate models with justification.
4. (4 pt) find the models listed in step 4, choose a best model.
5. (5 pt) Perform a complete analysis of the model chosen, including a residual analysis.
6. (1 pt) Depending on the results of the residual analysis, do we need to add quadratic and/or interaction terms, transform variables, and reanalyze the data. (**<u>do not transform</u>**)
7. (3 pt) Repeat step 3 using the stepwise method and compare both results.

Question Six: (20 pts) **GCROSLYN**

You are a real estate broker who wants to compare property values in Glen Cove and Roslyn (which are located approximately 8 miles apart). In order to do so, you will analyze the data that includes samples of houses from Glen Cove and Roslyn. Making sure to include the dummy variable for location (Glen Cove or Roslyn),

1. (3 pt) Develop a regression model to predict appraised value, based on the land area of a property, the age of a house, and location.
2. (3 pt) Find the 90% C.I. for the appraised value for a house with land= 0.228 acr, age=39 years, and located in Glen Cove.
3. (3 pt) Develop a model with all interaction terms.
4. (8 pt) Test the hypothesis that none of the interaction terms is significant in the model. (Write H0: and H1, Test statistic, critical value, conclusion)
5. (3 pt) Test the claim: "The interaction between land and age is significant in explaining the variation of appraised value". (Write H0: and H1, Test statistic, critical value/p-value, conclusion)

$$R_A^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k-1}\right)$$

Test statistic $F = \dfrac{\dfrac{SSR}{k}}{\dfrac{SSE}{n-k-1}} = \dfrac{MSR}{MSE}$

$$SSR(X_j \mid \text{All except } X_j) = SSR(\text{All}) - SSR(\text{All except } X_j)$$

$$r^2_{YX_j \bullet (\text{All except } X_j)} = \frac{SSR(X_j \mid \text{All except } X_j)}{SST - SSR(\text{All}) + SSR(X_j \mid \text{All except } X_j)}$$

$$t_{n-k-1} = \frac{b_i - 0}{s_{b_i}}$$

C.I. for the slope *β*$_i$ is $b_i \pm t_{\alpha/2} s_{b_i}$,.

Variance Inflationary Factor $VIF_j = \dfrac{1}{1 - R_j^2}$

***C*p statistic:** $C_p = \dfrac{(1-R_k^2)(n-T)}{1-R_T^2} - \left[n - 2(k+1)\right]$

Test statistic $F = \dfrac{\dfrac{SSR_{Full} - SSR_{Reduced}}{m}}{\dfrac{SSE}{n-k-1}} = \dfrac{\dfrac{SSE_{Reduced} - SSE_{Full}}{m}}{MSE}$