Assignment #4: Statistical Inference in Linear Regression (50 points)

This assignment will be made available in both pdf and Microsoft docx format. Answers should be typed into the docx file, saved, and converted into pdf format for submission. **Color your answers in green so that they can be easily distinguished from the questions themselves.**

Throughout this assignment keep all decimals to four places, i.e. X.xxxx.

Any computations that involve "the log function", denoted by log(x), are always meant to mean the natural log function (which will show as ln() on a calculator). The only time that you should ever use a log function other than the natural logarithm is if you are given a specific base.

In this assignment we will review model output from R and perform the computations related to statistical inference for linear regression. By performing this computations we are ensuring that we understand how the numbers in this R output are computed. **Students are expected to show all work in their computations. A good practice is to write down the generic formula for any computation and then fill in the values need for the computation from the problem statement.**

For internal use only

Model 1: Let's consider the following R output for a regression model which we will refer to as Model 1. (Note 1: In the ANOVA table, I have added 2 rows – (1) Model DF and Model SS - which is the sum of the rows corresponding to all the 4 variables (2) Total DF and Total SS - which is the sum of all the rows;

Note 2: The F test corresponding to the Model denotes the overall significance test. In R output, you will see that at the bottom of the Coefficients table)

ANOVA:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1974.53	1974.53	209.8340	< 0.0001
X2	1	118.8642568	118.8642568	12.6339	0.0007
X3	1	32.47012585	32.47012585	3.4512	0.0676
X4	1	0.435606985	0.435606985	0.0463	0.8303
Residuals	67	630.36	9.41		
Note: You can make the following calculations from the ANOVA table above to get Overall F statistic					
Model (adding 4 rows)	4	2126	531.50		<0.0001
Total (adding all rows)	71	2756.37			

Coefficients:				
	Estimate	Std. Error	t value	Pr(>t)
Intercept	11.3303	1.9941	5.68	<.0001
X1	2.186	0.4104		<.0001
X2	8.2743	2.3391	3.54	0.0007
X3	0.49182	0.2647	1.86	0.0676
X4	-0.49356	2.2943	-0.22	0.8303

Residual standard error: 3.06730 on 67 degrees of freedom	
Multiple R-sqaured: 0.7713, Adjusted R-squared: 0.7577	
F-statistic: on 4 and 67 DF, p-value < 0.0001	

Number of predictors	C(p)	R-square	AIC	BIC	Variables in the model
4	5	0.7713	166.2129	168.9481	X1 X2 X3 X4

(1) (5 points) How many observations are in the sample data?
The number of observations would be equal to the degrees of freedom +1. This would be equal to 72.

(2) (5 points) Write out the null and alternate hypotheses for the t-test for Beta1.
Null: $\beta 1 = 0$
Alternate: $\beta 1 \neq 0$

(3) (5 points) Compute the t- statistic for Beta1.
T = 2.186/.4104
T= 5.33

(4) (5 points) Compute the R-Squared value for Model 1, using ANOVA.

Model SS: 1974.53+118.8643+32.4701+.4356

Total SS: 1974.53+118.8643+32.4701+.4356+630.36

Model SS / Total SS = 2126.3/2756.6599

R squared = .7713

The R-Squared value for Model 1 is 77%

(5) (5 points) Compute the Adjusted R-Squared value for Model 1.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Adjusted R-Squared: 1-(1-.7713)((72-1)/(72-4-1) = .7576

(6) (5 points) Write out the null and alternate hypotheses for the Overall F-test.

Null: $\beta 1 = \beta 2 = \beta 3 = \beta 4 = 0$

Alternative: At least one $\beta \neq 0$

(7) (5 points) Compute the F-statistic for the Overall F-test.

F= (SSR/k) / (SSE/(n-k-1))

ModelSS: 1974.53+118.8643+32.4701+.4356

SSE: 630.36

(2126.3/4)/630.36/(72-4-1)

=56.5054

Model 2: Now let's consider the following R output for an alternate regression model which we will refer to as Model 2.

ANOVA:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1928.27000	1928.27000	218.8890	<.0001
X2	1	136.92075	136.92075	15.5426	0.0002
X3	1	40.75872	40.75872	4.6267	0.0352
X4	1	0.16736	0.16736	0.0190	0.8908
X5	1	54.77667	54.77667	6.2180	0.0152
X6	1	22.86647	22.86647	2.5957	0.112
Residuals	65	572.60910	8.80937		
Note: You can make the following calculations from the ANOVA table above to get Overall F statistic					
Model (adding 6 rows)	6	2183.75946	363.96	41.3200	<0.0001
Total (adding all rows)	71	2756.37			

For internal use only

Coefficients:				
	Estimate	Std. Error	t value	Pr(>t)
Intercept	14.3902	2.89157	4.98	<.0001
X1	1.97132	0.43653	4.52	<.0001
X2	9.13895	2.30071	3.97	0.0002
X3	0.56485	0.26266	2.15	0.0352
X4	0.33371	2.42131	0.14	0.8908
X5	1.90698	0.76459	2.49	0.0152
X6	-1.0433	0.64759	-1.61	0.112
Residual standard error: 2.968 on 65 degrees of freedom				
Multiple R-sqaured: 0.7923, Adjusted R-squared: 0.7731				
F-statistic: 41.32 on 6 and 65 DF, p-value < 0.0001				

Number of predictors	C(p)	R-square	AIC	BIC	Variables in the model
6	7	0.7923	163.2947	166.7792	X1 X2 X3 X4 X5 X6

(8) (5 points) Now let's consider Model 1 and Model 2 as a pair of models. Does Model 1 nest Model 2 or does Model 2 nest Model 1? Explain.

Two models are nested if one model contains all the terms of the other, and at least one additional term. Since Model 1 is $Y=\beta0 + \beta1(X1) + \beta2(X2) + \beta3(X3) + \beta4(X4)$ while Model 2 is $\beta0 + \beta1(X1) + \beta2(X2) + \beta3(X3) + \beta4(X4) + \beta5(X5) + \beta6(X6)$. Therefore, the models are nested based on the predictor variables associated with $\beta1$, $\beta2$, $\beta3$ and $\beta4$.

(9) (5 points) Write out the null and alternate hypotheses for a nested F-test using Model 1 and Model 2.

Null: $\beta k+p = 0$

Alternative: At least one $\beta \neq 0$

(10) (5 points) Compute the F-statistic for a nested F-test using Model 1 and Model 2.

F = (SSER−SSEC)/ # of additional β 0 s SSEC/[n−(k+p+1)]

F = (630.36 – 572.6091)/2/[72-(4+2+1)]

F = 57.7509/2/62

F = .4657

Here are some additional questions to help you understand other parts of inference.

(11) (0 points) Compute the AIC values for both Model 1 and Model 2.

AIC = n * log(SSE/n) +2*p

Model 1 AIC = 72 * ln(630.36/72) + 2*5

=166.2176

Model 2 AIC = 72 * ln(572.6091/72) + 2*7

= 157.4004

For internal use only

(12) (0 points) Compute the Mallow's Cp values for both Model 1 and Model 2.

BIC = n*ln(SSE/n)+P*ln(n)

Model 1 BIC = 72 * ln(630.36/72)+5*ln(72)

= 177.5963

Model 2 BIC =72 * ln(572.6091/72)+7*ln(72)

= 179.2315

(13) (0 points) Verify the t-statistics for the remaining coefficients in Model 1.

The t-statistics are provided for each of the coefficients in Model 1 except for X1. To verify the others and calculate X1, the estimate will be divided by the std error.

Intercept: 11.3303/1.9941 = 5.6818

X1: 2.186 / .0410 = 53.2651

X2: 8.274 / 2.3391 = 3.5361

X3: .4918 / .2647 = 1.858

X4: -.4936 / 2.2943 = -.2151

(14) (0 points) Verify the Mean Square values for Model 1 and Model 2.

To verify the Mean Square values for Model 1 and 2, the sum squares will be divided by the degrees of freedom

Model 1:

X1: 1974.53/1 = 1974/53

X2: 118.8643/1 = 118.8643

X3: 32.4701/1 = 32.4701

X4: .4356/1 = .4356

Residuals: 630.36/67 = 9.4084

Model2:

X1: 1928.2700/1 = 1928.2700

X2: 136.9208/1 = 136.9208

X3: 40.7588/1 = 40.7588

X4: .1674/1 = 1674

X5: 54.7767/1 = 54.7767

X6: 572.6091/65 = 8.8094

(15) (0 points) Verify the Root MSE values for Model 1 and Model 2.

Model 1: square root of 9.41 = 3.0676

Model 2: square root of 8.0894 = 2.8442

For internal use only