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 SAS 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 SAS 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.**

Grading Note: These problems will be graded 'up or down', i.e. there is no partial credit. This practice is how this assignment is graded, how the small computations in the quizzes are graded (since they are automated), and how the small computations on the final exam are graded.

Model 1: Let's consider the following SAS output for a regression model which we will refer to as Model 1.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2126.00904	531.50226		<.0001
Error	67	630.35953	9.40835		
Corrected Total	71	2756.36857			

Root MSE	3.06730	R-Square	
Dependent Mean	37.26901	Adj R-Sq	
Coeff Var	8.23017		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.33027	1.99409	5.68	<.0001
X1	1	2.18604	0.41043		<.0001
X2	1	8.27430	2.33906	3.54	0.0007
X3	1	0.49182	0.26473	1.86	0.0676
X4	1	-0.49356	2.29431	-0.22	0.8303

Number in Model	C(p)	R-Square	AIC	BIC	Variables in Model
4	5.0000	0.7713	166.2129	177.5963	X1 X2 X3 X4

(1) (5 points) How many observations are in the sample data? (Hint: The answer needs to be computed. It is not simply a value listed on this page.)

To determine this, we need to calculate the observations as n = df + p + 1. The variables indicate that df = degrees of freedom, p = predictor variables, and n = number of observations. So n = 67 + 4 + 1, which brings us to 72 observations.

(2) (5 points) Write out the null and alternate hypotheses for the t-test for Beta1.

Full explicit model: $y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + \varepsilon$

Reduced Model: $y = B_0 + B_2X_2 + B_3X_3 + B_4X_4 + \varepsilon$ while defining the null and alternative hypotheses as:

H_0: Beta 1 = 0, which states that the null hypothesis for Beta 1 is 0 and X1 has no meaningful contribution to the prediction of the response variable

H_a: Beta 1 ≠ 0, which states that the alternative hypothesis for Beta 1 is not 0 and that it has statistically significant effect on the prediction of the response variable

(3) (5 points) Compute the t- statistic for Beta1.

T-statistic calculation: Estimate / Standard Error

Beta 1 T-Statistic: 2.18604/.41043 = 5.3262

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

R-Squared value calculation: Sum Squared of Residuals (SSR) / Total Sum of Squares (SSTO)

R-Squared for Model 1: 2126.00904 / 2756.36857 = .7713

This is validated by the R-Squared represented in the output above.

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

Adjusted R-Squared calculation: $R^2 - (1 - R^2) * \frac{p}{(n-p-1)}$

Adjusted R-Squared for model 1: $.7713 - (1 - .7713) * \frac{4}{(72-4-1)} = .7577$

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

Test: hypothesis that all predictor variables have no explanatory influence which would leave each coefficient equal to zero.

Full Model (H_a): $y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + \varepsilon$; $where\ B_1\ or\ B_2\ or\ B_3\ or\ B_4 \neq 0$

Reduced Model (H_0): $y = B_0 + \varepsilon; where\ B_1 = B_2 = B_3 = B_4 = 0$

There needs to be confirmation that at least one coefficient does not equal 0.

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

F-Stat calculation: Mean Square due to Regression (MSR) / Mean Square due to Error (MSE)

F-Stat for Model 1: 531.50226/9.40835 = 56.4926

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

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	2183.75946	363.95991	41.32	<.0001
Error	65	572.60911	8.80937		
Corrected Total	71	2756.36857			

Root MSE	2.96806	R-Square	0.7923
Dependent Mean	37.26901	Adj R-Sq	0.7731
Coeff Var	7.96388		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.39017	2.89157	4.98	<.0001
X1	1	1.97132	0.43653	4.52	<.0001
X2	1	9.13895	2.30071	3.97	0.0002
X3	1	0.56485	0.26266	2.15	0.0352
X4	1	0.33371	2.42131	0.14	0.8908
X5	1	1.90698	0.76459	2.49	0.0152
X6	1	-1.04330	0.64759	-1.61	0.1120

Number in Model	C(p)	R-Square	AIC	BIC	Variables in Model
6	7.0000	0.7923	163.2947	179.2313	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.

If we look at the reduced models of both model 1 (M1) and model 2 (m2), it appears model 1 nests model 2. Reduced model 1 has less predictors than model 2. Model 1 also excludes predictors, which

would make this a unique case. To see if model 1 would fit better than model 2, the F-Test would be performed.

Model 1 and model 2 would be compared based on their full and reduced models to see if their independent variables are statistically significant. The p-values for each model suggests which variables can be determined to fit positively to the regression.

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

Full Model: $y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6 + \varepsilon$

Reduced Model: $y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + \varepsilon$

H_0: $B_5 = B_6 = 0$

H_a: $B_5 \neq = or\ B_6 \neq 0$**; if p-values are shown to have statistical significance, then it the null hypothesis will be rejected. This means that X5 and X6 should be included in the model due to their significance in their exploratory power. If this happens, then model 2 will be chosen over model 1.**

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

F-Stat calculation: Mean Square due to Regression (MSR) / Mean Square due to Error (MSE)

$$\frac{MSR}{MSE} = \frac{\frac{SSR}{P}}{\frac{SSE}{n-p-1}}$$

Model 1:

$$F = \frac{\frac{.7713}{4}}{\frac{1-.7713}{72-4-1}} = 56.4901$$

Model 2:

$$F = \frac{\frac{.7923}{6}}{\frac{1-.7923}{72-6-1}} = 41.3252$$

Here are some additional questions to help you understand other parts of the SAS output.

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

(12) (0 points) Compute the BIC values for both Model 1 and Model 2. (Hint: Compute the BIC using the Schwarz BIC formula. Why does this value differ from the SAS value? What formula does SAS use?)

(13) (0 points) Compute the Mallow's Cp values for both Model 1 and Model 2. (Hint: This is a trick question. Do these values make sense? Why might they not make sense? Consult your LRA book.)

(14) (0 points) Verify the t-statistics for the remaining coefficients in Model 1.

(15) (0 points) Verify the Mean Square values for Model 1 and Model 2.

(16) (0 points) Verify the Root MSE values for Model 1 and Model 2.