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	168.9481	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.)

$$Degrees\ of\ freedom\ Corrected\ Total\ = Sample\ size - 1$$

71 = Sample size – 1

Sample size a.k.a observations in data = 72

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

$$\mathbf{H0:}\ \boldsymbol{\beta 1 = 0}$$

$$\mathbf{H1:}\ \boldsymbol{\beta 1 \neq 0}$$

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

$$t_1 = \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})}$$

$$t_1 = \frac{2.18604}{0.41043}$$

$$\boxed{t_1 = 5.3262}$$

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

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Total}}$$

$$R^2 = 1 - \frac{630.35953}{2756.36857}$$

$$\boxed{R^2 = 0.7713}$$

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

$$R^2adj = 1 - \frac{SS_{Res}/n-p}{SS_{Total}/n-1}$$

$$R^2adj = 1 - \frac{MS_{Res}}{MS_{Total}}$$

$$R^2adj = 1 - \frac{9.40835}{38.82209}$$

$$\boxed{R^2adj = 0.7577}$$

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

$$H0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H1: \beta_0 \neq 0, \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \beta_4 \neq 0$$

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

$$F = \frac{MS_{Reg}}{MS_{Res}}$$

$$F = \frac{531.50226}{9.40835}$$

$$\boxed{F = 56.4926}$$

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	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.

Model 2 nests Model 1; Model 1's predictors are a subset of Model 2's predictors

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

$$\mathbf{H0}: \beta_5 = \beta_6 = 0$$
$$\mathbf{H1}: \beta_5 \neq 0, \beta_6 \neq 0$$

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

$$F_0 = \frac{[SSE_{RM} - SSE_{FM}]/[dim_{FM} - dim_{RM}]}{SSE_{FM}/[n - dim_{FM}]}$$

$$F_0 = \frac{[630.35953 - 572.60911]}{[7-5]} \Big/ \frac{572.60911}{[72-7]}$$

$$\boxed{F_0 = 3.2778}$$

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.

$$AIC = n \ln\left(\frac{SS_{Res}}{n}\right) + 2p$$

$$AIC_{Model\ 1} = 72 * \ln(630.35953\ /72) + 2*5$$

$$\boxed{AIC_{Model\ 1} = 166.2129}$$

$$AIC_{Model\ 2} = 72 * \ln(572.60911\ /72) + 2*7$$

$$\boxed{AIC_{Model\ 2} = 163.2947}$$

(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?)

$$BIC = n \ln\left(\frac{SS_{Res}}{n}\right) + p\ln(n)$$

$$BIC_{Model\ 1} = 72 * \ln(630.35953\ /72) + 7*\ln(72)$$

$$\boxed{BIC_{Model\ 1} = 177.5963}$$

$$BIC_{Model\ 2} = 72 * \ln(572.60911\ /72) + 7*\ln(72)$$

$$\boxed{BIC_{Model\ 2} = 179.2313}$$

The above results (Schwarz formula) are different from that of SAS, because SAS uses the Sawa criterion.

(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.)

$$C_p = \frac{SS_{Res}}{\widehat{\sigma^2}} - n + 2p$$

$$C_{p\ (model\ 1)} = \frac{630.35953}{9.40835} - 72 + 2 * 5$$

$$\boxed{C_{p\ (model\ 1)} = 5}$$

$$C_{p\ (model\ 2)} = \frac{572.60911}{8.80937} - 72 + 2 * 7$$

$$\boxed{C_{p\ (model\ 2)} = 7}$$

Here we use Mean square error as the estimate for variation. In turn we are assuming a zero bias model. We know since model 1 is a reduced model, it likely to have some degree of bias. Hence Cp does not make sense. Likewise for model 2 we do not know what the variance is, we use the mean square error as an estimate, assuming a zero bias model. Which is also not always correct.

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

Variable	DF	Parameter Estimate	Standard Error	t Value = Parameter Estimate/Standard Error	t Value (SAS output) in question
Intercept	1	11.33027	1.99409	5.6819	5.68
X1	1	2.18604	0.41043	5.3262	
X2	1	8.2743	2.33906	3.5374	3.54
X3	1	0.49182	0.26473	1.8578	1.86
X4	1	-0.49356	2.29431	-0.2151	-0.22

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

Model1				
Source	DF	Sum of Squares	Mean Square = Sum of Squares / DF	Mean Square (SAS output in question)
Model	4	2126.00904	531.5023	531.50226
Error	67	630.35953	9.4084	9.40835
Corrected Tota	71	2756.36857	38.8221	

Model2				
Source	DF	Sum of Squares	Mean Square = Sum of Squares / DF	Mean Square
Model	6	2183.75946	363.9599	363.95991
Error	65	572.60911	8.8094	8.80937
Corrected Tota	71	2756.36857	38.8221	

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

$$Root\ MSE = \sqrt{Mean\ Square\ Error}$$

$$Root\ MSE_{Model\ 1} = \sqrt{9.40835} = 3.0673$$

$$Root\ MSE_{Model\ 2} = \sqrt{8.80937} = 2.9681$$