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.

Zeeshan Latifi Assignment 4

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			

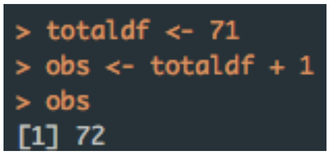
Root MSE	3.06730	**R-Square**	
Dependent Mean	37.26901	**Adj R-Sq**	
Coeff Var	8.23017		

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.)

Total number of observations are 72, total degrees of freedom from the ANOVA table plus 1.



```
> totaldf <- 71
> obs <- totaldf + 1
> obs
[1] 72
```

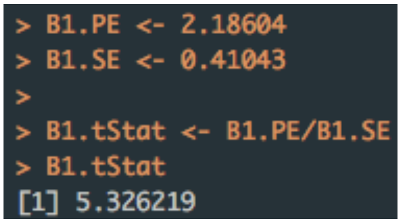
(2) (5 points) Write out the null and alternate hypotheses for the t-test for Beta1.

$H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$

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

The t-statistic is computed by the parameter estimate divided by the standard error for β_1. The answer is 5.326219

$$t_i = \frac{\widehat{\beta_\iota}}{se(\widehat{\beta_\iota})}$$

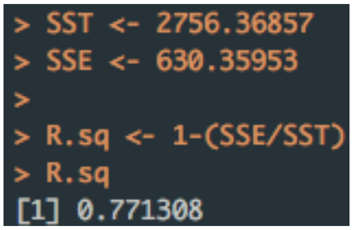


```
> B1.PE <- 2.18604
> B1.SE <- 0.41043
>
> B1.tStat <- B1.PE/B1.SE
> B1.tStat
[1] 5.326219
```

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

This is computed by taking the error sum of squares (SSE) divided by the total sum of squares (SST). Subtract that value by 1. The answer is 0.7713

$$R^2 = 1 - \frac{SSE}{SST}$$



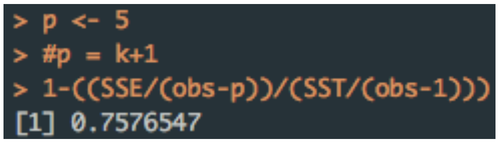
```
> SST <- 2756.36857
> SSE <- 630.35953
>
> R.sq <- 1-(SSE/SST)
> R.sq
[1] 0.771308
```

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

Computed by using the formula below:

$$1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

The answer is 0.7576



```
> p <- 5
> #p = k+1
> 1-((SSE/(obs-p))/(SST/(obs-1)))
[1] 0.7576547
```

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

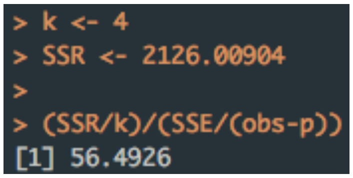
$$H_0: \beta_1 = \cdots = \beta_k = 0 \quad vs \quad H_1: \beta_i \neq 0 \;\; for\ some\ i \in \{1, \dots, k\}$$

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

Computed by the formula below:

$$F_0 = \frac{SSR/k}{SSE/(n-p)}$$

The answer is 56.4926



```
> k <- 4
> SSR <- 2126.00904
>
> (SSR/k)/(SSE/(obs-p))
[1] 56.4926
```

<u>Model 2:</u> 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.

Based on the predictor variables, it shows that Model 1 predictors are a subset of Model 2 predictor variables. Model 1 is the reduced model and Model 2 is the full model. Therefore, Model 2 nests Model 1. $Model\ 1 \subset Model\ 2$

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

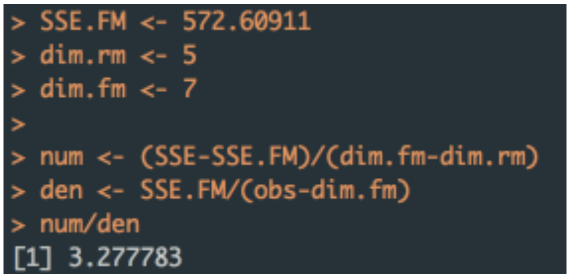
$H_0: \beta_1 \ldots = \beta_6 = 0 \quad vs \quad H_1: \beta_i \neq 0 \;\; for\ some\ i\ \in \{1 \ldots 6\}$

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

Computed by the formula below:

$$F_0 = \frac{[SSE(RM) - SSE(FM)]/[dim(FM) - \dim(RM)]}{SSE(FM)/[(n - \dim(FM))]}$$

The answer is 3.2777



```
> SSE.FM <- 572.60911
> dim.rm <- 5
> dim.fm <- 7
>
> num <- (SSE-SSE.FM)/(dim.fm-dim.rm)
> den <- SSE.FM/(obs-dim.fm)
> num/den
[1] 3.277783
```

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

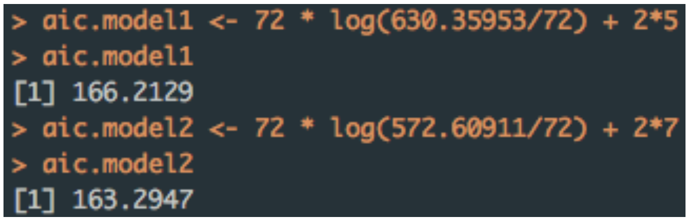
(11) (0 points) Compute the AIC values for both Model 1 and Model 2.

Computed by the formula below:

$$AIC = n * \log\left(\frac{SSE}{n}\right) + 2p$$

The answer for Model 1 is 166.2129

The answer for Model 2 is 163.2947



```
> aic.model1 <- 72 * log(630.35953/72) + 2*5
> aic.model1
[1] 166.2129
> aic.model2 <- 72 * log(572.60911/72) + 2*7
> aic.model2
[1] 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?)

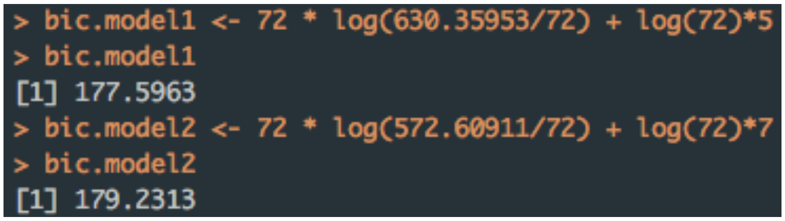
Computed by the formula below:

$$BIC = n * \log\left(\frac{SSE}{n}\right) + \log(n)\,p$$

The answer for Model 1 is 177.5963

The answer for Model 2 is 179.2313

The values are different due to the sample size, since n increases BIC will select with probability approaching 1.



```
> bic.model1 <- 72 * log(630.35953/72) + log(72)*5
> bic.model1
[1] 177.5963
> bic.model2 <- 72 * log(572.60911/72) + log(72)*7
> bic.model2
[1] 179.2313
```

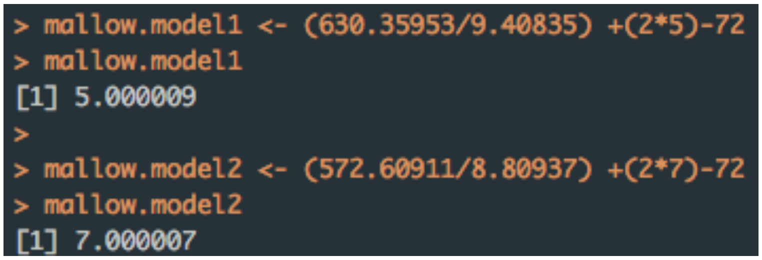
(13) (0 points) Compute the Mallow's Cp values for both Model 1 and Model 2. (Hint: This is a trick question. Do these values make sense? Why might they not make sense? Consult your LRA book.)

Computed by the formula below:

$$C_{pc} = \frac{SSE}{\hat{\sigma}^2} + 2pc - n$$

The answer for Model 1 is 5.0000

The answer for Model 2 is 7.0000



```
> mallow.model1 <- (630.35953/9.40835) +(2*5)-72
> mallow.model1
[1] 5.000009
>
> mallow.model2 <- (572.60911/8.80937) +(2*7)-72
> mallow.model2
[1] 7.000007
```

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

$$t_i = \frac{\widehat{\beta_\iota}}{se(\widehat{\beta_\iota})}$$

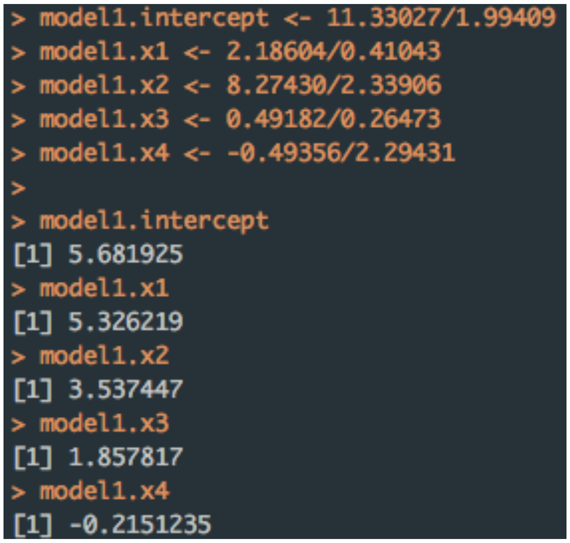
Intercept: 5.6819

X1: 5.3262

X2: 3.5374

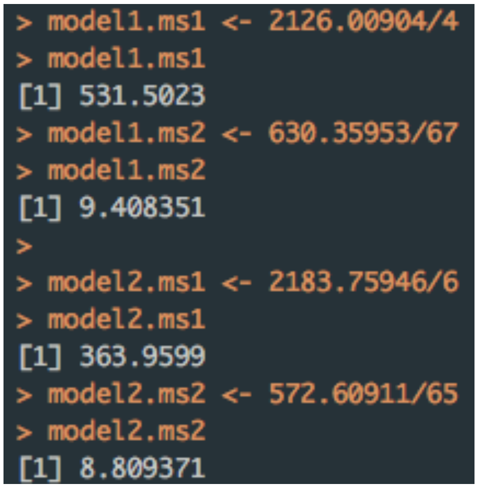
X3: 1.8578

X4: -0.2151



```
> model1.intercept <- 11.33027/1.99409
> model1.x1 <- 2.18604/0.41043
> model1.x2 <- 8.27430/2.33906
> model1.x3 <- 0.49182/0.26473
> model1.x4 <- -0.49356/2.29431
>
> model1.intercept
[1] 5.681925
> model1.x1
[1] 5.326219
> model1.x2
[1] 3.537447
> model1.x3
[1] 1.857817
> model1.x4
[1] -0.2151235
```

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



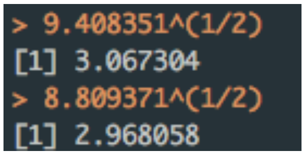
```
> model1.ms1 <- 2126.00904/4
> model1.ms1
[1] 531.5023
> model1.ms2 <- 630.35953/67
> model1.ms2
[1] 9.408351
>
> model2.ms1 <- 2183.75946/6
> model2.ms1
[1] 363.9599
> model2.ms2 <- 572.60911/65
> model2.ms2
[1] 8.809371
```

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

Take the square root of the mean square error for each model.

The answer for Model 1 is 3.0673

The answer for Model 2 is 2.9681



```
> 9.408351^(1/2)
[1] 3.067304
> 8.809371^(1/2)
[1] 2.968058
```