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

Andrew Knight

<u>Model 1:</u> 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?

From the ANOVA table, there are 67 degrees of freedom and 4 predictor variables. Number of observations can be found by n = df + p + 1. So, we have n = 67 + 4 + 1 = 72. We have 72 observations.

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

The Full Model could be written as Y = B0 + B1 X1 + B2 X2 + B3 X3 + B4 X4 + ε and

the Reduced Model as Y = B0 + B2 X2 + B3 X3 + B4 X4 + ε where

H0: Beta1 = 0, this null hypothesis states that the coefficient B1 is zero and the variable x1 has no meaningful contribution to the prediction of the response variable.

Ha: Beta1 ≠ 0, this alternate hypothesis states that the coefficient B1 is not zero and thus has a statistically significant effect on the prediction on the response variable.

Note: I may use the symbols ≠, <>, or != interchangeably to indicate 'not equal' in answers below.

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

The t-statistic is given by the Estimate / Std Error.

For Beta1: 2.186 / 0.4104 = 5.3265

The t-test would be used to find the resulting p-value of the error and determine if we should reject the null hypothesis or not. Based on this t-value, the p-value is low and thus statistically significant. When p is low, null must go. We reject the null hypothesis that B1 = 0.

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

The R-squared value is given by the Model1 SSR (Sum of Squares of the residuals) / SSTO (Total Sum of Squares). From the formula we get:

R-squared = 2126 / 2756.37 = 0.7713

This is verified by the linear model summary displayed above.

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

The Adjusted R-squared value is given by: R^2 – (1-R^2) * p / (n – p – 1)

For Model 1: R-squared (adj) = 0.7713 – (1 – 0.7713) * 4 / (72 – 4 – 1) = 0.7577

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

We are testing the hypothesis that all predictor variables have no explanatory influence and as such would list each coefficient as being equal to zero.

Reduced Model for H0: y = B0 + ε, where B1 = B2 = B3 = B4 = 0

Full Model for Ha: y = B0 + B1 x1 + B2 x3 + B3 x3 + B4 x4 + ε, where B1 or B2 or B3 or B4 != 0

We want to confirm for each of the coefficients that at least one is not zero.

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

The F-statistic is given by Mean Square Due to Regression (MSR) / Mean Square Due to Error (MSE)

From table above, Overall F-stat = 531.5 / 9.41 = 56.4825

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			

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.

Model 1 is nested in Model 2. The reduced Model 1 has less predictors than Model 2. Model 1 would be considered a special case of Model 2 because Model 1 excludes predictors. The F-test would be used to test if the reduced Model 1 is a better fit than Model 2.

The full and reduced models would state the comparison of the Model 1 and Model 2 in terms of the independent variables which are statistically significant. Based on the p-values for each we can determine which variables contribute positively to the regression fit.

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

The Full Model (FM): $y = B1 * x1 + B2 * x2 + B3 * x3 + B4 * x4 + B5 * x5 + B6 * x6 + \varepsilon$

The Reduced Model (RM): $y = B1 * x1 + B2 * x2 + B3 * x3 + B4 * x4 + \varepsilon$

The Null Hypothesis is $B5 = B6 = 0$

The Alt Hypothesis is B5 != 0 or B6 != 0, if the p-values are found to be statistically significant, then we would reject the null hypothesis which means the predictors x5 and x6 have significant explanatory power and thus should be included in the model. If this were the case, we would choose Model 2 over Model 1 due to its ability to better predict the response variable.

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

The F-Test formula is given by, $F = (SSR / p) / (SSE / (n - p - 1)) = MSR / MSE$, where

MSR = the mean square due to regression and

MSE = mean square due to error.

It can also be written as $F = R^2/p / ((1 - R^2p) / df)$, and using the values from tables above we get:

For Model 1:

$F = (0.7713 / 4) / ((1 - 0.7713) / 67) = .1928 / .003413 = 56.4901$

For Model 2:

$F = (0.7923 / 6) / ((1 - 0.7923) / 65) = .13205 / .003195 = 41.3252$

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.

(12) (0 points) Compute the BIC values for both Model 1 and Model 2.

(13) (0 points) Compute the Mallow's Cp values for both Model 1 and Model 2.

(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.