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

Prabhat Thakur

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?

Ans: From the ANOVA table, the model is fitted using 4 predictors (p) and has 67 degrees of freedom (df).
Model's degree of freedom is calculated by df = n-p-1, so the number of observations can be calculated by **n = df + p + 1**. Substituting df and p value in above formula, we get **n = 67 + 4 + 1 = 72.**

There are 72 **observations** in the sample data.

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

Ans: For t-test hypotheses of an individual coefficient, we can form null and alternate hypotheses as below.
NH: $\beta 1 = 0$, this null hypothesis states that the coefficient β1 is zero and the variable x1 has no meaningful contribution to the prediction of the response variable.
AH: $\beta 1 \neq 0$, this alternate hypothesis states that the coefficient β1is not zero and thus has a statistically significant effect on the prediction of the response variable.

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

Ans: From the ANOVA table, the t-statistic of a coefficient can be calculated by coefficient Estimate / coefficient Std Error.

For β1 coefficient, t-value = 2.186 / 0.4104 = **5.3265**

If we extend Q2, t-test for β1, significance of a coefficient can be tested by evaluating p- value for the calculated t-value. P-value can be calculated using t-value and degree of freedom, for β1, the two-tailed p- value is less than 0.0001 which is extremely statistically significant. Similar p-value for β1 is given in the above coefficient table. We can use the p-value to complete our t-test for β1. Since, the p-value is low and thus statistically significant, we reject the null hypothesis that β1 = 0.

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

Ans: The R-squared value is calculated by ratio of the model sum of squares to the total sum of squares.
R-squared = Model SS/ Total SS

For Model 1, R-squared = 2126 / 2756.37 = **0.7713**
This R-squared value also matches with above model 1 summary statistics.

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

Ans: Since we already calculated R-squared value for Model 1, traditional formula for expressing the adjusted R-squared in terms of the ordinary R-squared is given by:

$$R_{adj}^2 = R^2 - \frac{(1 - R^2) * p}{n - p - 1}$$

n= number of observations, p= no of predictors

For Model 1:

$$R_{adj}^2 = 0.7713 - \frac{(1 - 0.7713) * 4}{72 - 4 - 1}$$

$$R_{adj}^2 = \mathbf{0.7577}$$

This Adjusted R-squared value also matches with above model 1 summary statistics.

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

Ans: Model 1 is in the following form **E(y) =** β0+ β1x1+ β2x2+ β3x3+ β4x4

For overall F-test hypotheses of a model, we can form null and alternate hypotheses as below.
NH: β1 = β2 = β3 = β4 = 0 -> Reduced Model (RM) is **E(y) =** β0, null hypothesis states that coefficients β1 , β2 , β3 , and β4 is zero and the variable x1, x2, x3, and x4 has no meaningful contribution to the prediction of the response variable.

AH: β1 or β2 or β3 or β4 ≠ 0, Full model (FM) is **E(y) =** β0+ β1x1+ β2x2+ β3x3+ β4x4, this alternate hypothesis states that at least one predictor coefficient is not zero and thus has a statistically significant effect on the prediction of the response variable.

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

Ans: The F-statistic is given by Model Mean Square Due to Regression (MSR) / Mean Square Due to Error/Residual (MSE).
From ANOVA table above, Overall F-stat = 531.5 / 9.41 = **56.4825**

If we extend Q6, Overall F-test model 1, significance of a model can be tested by evaluating p- value for the calculated F-value. We can use the p-value to complete our Overall F-test. Since, the p-value is low and thus statistically significant, we reject the null hypothesis that all predictor coefficients β1 = β2 = β3 = β4 = 0.

<u>Model 2:</u> 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.

Ans: Model 1 is nested in Model 2. All the predictors in Model1 are also used in Model 2. Another way to define their relationship is Model 1 is reduced model of Model 2 and has less predictors compared to Model 2. Also, since Model 2 has more predictors, its R-squared value is higher than model 1.

Partial F test is used to test 'nested' models, which means to test if Model 1 (reduced model) is better than Model 2

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

Ans: Regression functions for Model 1 and Model 2 are as follows
Model 1 E(y) = β0+ β1x1+ β2x2+ β3x3+ β4x4
Model 2 E(y) = β0+ β1x1+ β2x2+ β3x3+ β4x4+ β5x5+ β6x6

Null and alternate hypotheses for a nested F-test using Model 1 and Model 2 can be written as below:
NH: β5 = β6 = 0; -> Model 2 Reduced Model (RM) is E(y) = β0+ β1x1+ β2x2+ β3x3+ β4x4 which is same as Model 1. The null hypothesis states that coefficients β5 and β6 of Model 2 is zero and the variable x5, and x6 has no meaningful contribution to the prediction of the response variable, hence Model 2 is not statically significant compared to Model 1.

AH: β5 ≠ 0 or β6 ≠ 0; -> Model 2 Full model (FM) is **E(y) =** β0+ β1x1+ β2x2+ β3x3+ β4x4+ β5x5+ β6x6, this alternate hypothesis states that at least one coefficient of the additional predictors x5 and x6 in Model 2 is not zero and thus has a statistically significant effect on the prediction of the response variable.

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

Ans: To Compute the F-statistic for a nested F-test in other words to see whether the reduced model is adequate, we use the ratio

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/(p+1-k)}{\text{SSE(FM)}/(n-p-1)}\ .$$

SSE (RM) = Model 1 Sum Square of Error/residual = 630.36
SSE (FM) = Model 2 Sum Square of Error/residual = 572.6091
n = number of observations = 72
p = # of predictors in Model 2 = 6
k = # of parameters in Model 1 = 5

Let put these values in the above formula of F-value.
F = ([630.36 - 572.6091]/ (6+1-5))/ (572.6091 / (72-6-1))
F= (57.7509/2)/ (572.6091 / 65) = 28.8754 / 8.8094

F= 3.2778

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.

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.