Modeling Assignment 3: Statistical Inference in Multiple Linear Regression

Assignment Overview

This assignment has two parts, the first part, Mechanics and Computations, is intended to be sure that you understand the mechanics of hypothesis testing and the information provided from a typical regression analysis. This first part is computational in nature and does not require the use of R, though may require the use of a calculator. The second part, Application, asks you to begin to apply statistical inference using regression models with the AMES data that you worked with during Modeling Assignment #1. You will use R for all descriptive statistics, graphs, and fitting regression models.

In this assignment we will review model output from R and perform hypothesis specifications and computations related to statistical inference for linear regression. You are expected to show all work in your 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. Throughout this assignment keep all decimals to four places, i.e. X.xxxx. You are expected to use correct notation and terminology, as well as to be clear, complete and concise with all interpretation of results.

Assignment Document

Results should be presented, labeled, and discussed in the numerical order of the questions given. Please use MS-WORD or some other text processing software to record and present your answers and results. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document you submit to be graded MUST be submitted in pdf format. Please use the naming convention: ModelAssign3_YourLastName.pdf.

PART 1: MECHANICS AND COMPUTATIONS (30 points)

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) (3 points) How many observations are in the sample data?

(2) (3 points) Write out the null and alternate hypotheses for the t-test for Beta1.

(3) (3 points) Compute the t- statistic for Beta1. Conduct the hypothesis test and interpret the result.

(4) (3 points) Compute the R-Squared value for Model 1, using information from the ANOVA table. Interpret this statistic.

(5) (3 points) Compute the Adjusted R-Squared value for Model 1. Discuss why Adjusted R-squared and the R-squared values are different.

(6) (3 points) Write out the null and alternate hypotheses for the Overall F-test.

(7) (3 points) Compute the F-statistic for the Overall F-test. Conduct the hypothesis test and interpret the result.

Model 2

Now let's consider the following R output for an alternative 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) (3 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.

(9) (3 points) Write out the null and alternate hypotheses for a nested F-test using Model 1 and Model 2.

(10) (3 points) Compute the F-statistic for a nested F-test using Model 1 and Model 2. Conduct the hypothesis test and interpret the results.

PART II: APPLICATION (20 points)

For this part of the assignment, you are to use the AMES Housing Data you worked with during Modeling Assignment #1. Each question is worth 5 points.

Model 3

(11) Based on your EDA from Modeling Assignment #1, focus on 10 of the continuous quantitative variables that you though/think might be good explanatory variables for SALESPRICE. Is there a way to logically group those variables into 2 or more sets of explanatory variables? For example, some variables might be strictly about size while others might be about quality. Separate the 10 explanatory variables into at least 2 sets of variables. Describe why you created this separation. A set must contain at least 2 variables.

(12) Pick one of the sets of explanatory variables. Run a multiple regression model using the explanatory variables from this set to predict SALEPRICE(Y). Call this Model 3. Conduct and interpret the following hypothesis tests, being sure you clearly state the null and alternative hypotheses in each case:

a) all model coefficients individually

b) the Omnibus Overall F-test

Model 4

(13) Pick the other set (or one of the other sets) of explanatory variables. Add this set of variables to those in Model 3. You are preparing to fit a multiple regression model with this combined set of explanatory variables – call this Model 4. You should note that Model 3 is nested within Model 4. Fit the multiple regression model using the explanatory variables from the combined set of explanatory variables to predict SALEPRICE(Y). In other words, fit Model 4. Conduct and interpret the following hypothesis tests, being sure you clearly state the null and alternative hypotheses in each case:

a) all model coefficients individually

b) the Omnibus Overall F-test

Nested Model

(14) Write out the null and alternate hypotheses for a nested F-test using Model 3 and Model 4, to determine if the set of additional variables added to Model 3 to make Model 4 variables are useful for predicting SALEPRICE(Y). Your hypotheses must use symbols. Compute the F-statistic for this nested F-test and interpret the results.