

|  |
| --- |
| Week 3 Assignment - Computational: Statistical Inference in Linear Regression  *MSDS 410* |

This assignment has two parts, the first is intended to be sure that you understand the mechanics of hypothesis testing and the information provided from a typical regression analysis. The second part asks you to begin to apply statistical inference using regression models with the AMES data.

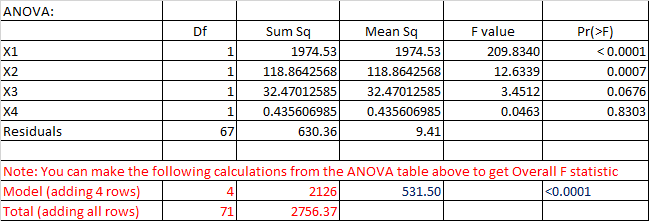
In this assignment we will review model output from R and perform hypothesis specifications and computations related to statistical inference for linear regression. Students are expected to show all work in their computations. A good practice is to write down the generic formula for any computation and then fill in the values need for the computation from the problem statement. Throughout this assignment keep all decimals to four places, i.e. X.xxxx. Students are expected to use correct notation and terminology, and to be clear, complete and concise with all interpretations of results.

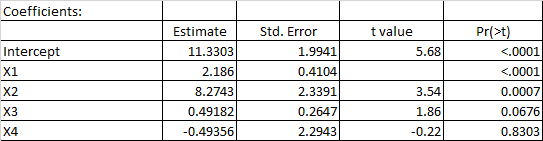
Any computations that involve “the log function”, denoted by log(x), ***are always meant to mean the natural log function (which will show as ln() on a calculator).*** The only time that you should ever use a log function other than the natural logarithm is if you are given a specific base.

**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)









1. (3 points) How many observations are in the sample data?

Residuals df = n – k - 1

67 = n – 4 – 1

n = 67 + 4 + 1 = 72

We have 72 observations

1. (3 points) Write out the null and alternate hypotheses for the t-test for Beta1.

In OLS regression the statistical inference for the individual regression coefficients can be performed by using a t-test. If there is a significant linear relationship between the independent variable X and the dependent variable Y, the beta1 will not equal zero. The null hypothesis states that the slope (beta 1) is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

* H0: Beta1 = 0; this is a null hypothesis; If it is true, x1 varaible is not a significant predictor to predict a response Y varaible
* Ha: Beta1 ≠ 0; this is an alternative hypothesis. If it is true, x1 varaible is a significant predictor to predict a response Y varaible

1. (3 points) Compute the t- statistic for Beta1. Conduct the hypothesis test and interpret the result.

t-statistic = Beta1 / SE

where b1 is the slope of the sample regression line, and SE is the standard error of the slope.

t-statistic = 2.186 / 0.4104 = 5.3265;

We can reject H0 based on the value of t-statistic and the given significance level. This decision can be made by using the p-value of the t-statistic or by using the critical value for the significance level. With df = 71, t-statistic = 5.3265, and significance level, we get p-value = 0.00001. When the p-value is low, null must go. We reject null hypthesis (beta1 = 0) and conclude that x1 varaible is a significant predictor to predict a response Y varaible.

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

R-Squared = SS Model / SS Total = 2126 / 2756.37 = 0.7713

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

Adjusted R-Squared = [R-Squared] – ((1 – [R-Squared]) \* [# of predictors] / [residuals df])

= 0.7713 – ((1 – 0.7713) \* 4 /67) = 0.7713 – 0.01365 = 0.7577

1. (3 points) Write out the null and alternate hypotheses for the Overall F-test.

Consider regression model Y = B0 + B1X1 + B2X2 + B3X3 + B4X4

The overall F-test for a regression effect is a joint hypothesis test that at least one of the predictor varaibles has a non-zero coefficient.

H0: B1 = B2 = B3 = B4 = 0; there is no relationship between predictors and a response Y varaible

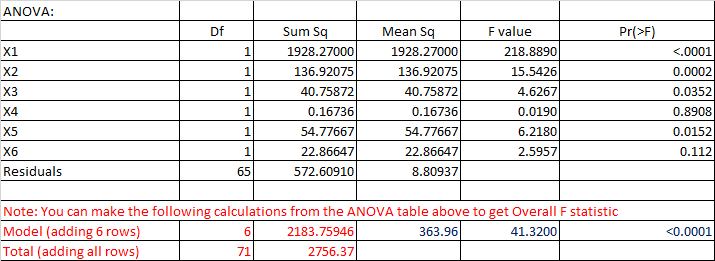
Ha: there is at least 1 inequality

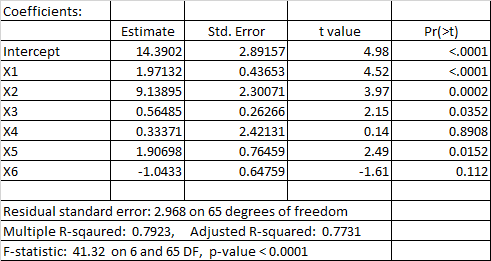
1. (3 points) Compute the F-statistic for the Overall F-test. Conduct the hypothesis test and interpret the result.

F-statistics = MS regression / MS residual = 531.5 / 9.41 = 56.4825

When there is a high F-statitics, there will be very small p-value. We can see that the F-statistics is high which we reject null hypotheses and conclude that this model is a significant and there is at least one predictor correlates to response variable Y.

**Model 2:** Now let’s consider the following R output for an alternate regression model which we will refer to as Model 2.







1. (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.

Model 1 is nested in Model 2 because Model 2 has more predictors than Model 1. We can also say that Model 1 is reduced (subset) model and Model 2 is full model. We can use a F-test for nested models to decide wether or not to include an additional predictor variable in the final model. Here are the models in equation forms:

* Reduced Model: Y = B0 + B1X1 + B2X2 + B3X3 + B4X4
* Full Model: Y = B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5+ B6X6

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

Ho: B5 = B6 = 0

Ha: there is at least 1 inequality

If p-value is small, I’ll be reducing error significantly with full model. Full model will be significant.

1. (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.

F- statistics = (([SS residuals of reduced model] - [SS residuals of full model]) / 2) / [mean sq. residuals of full model] = ((630.36 – 572.6091) / 2) / 8.8094 = 3.2778

So, F- statistics is small (p-value is big), do not reject null hypothesis and conclude that the reduced model is a good enough; do not need full model.

**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.

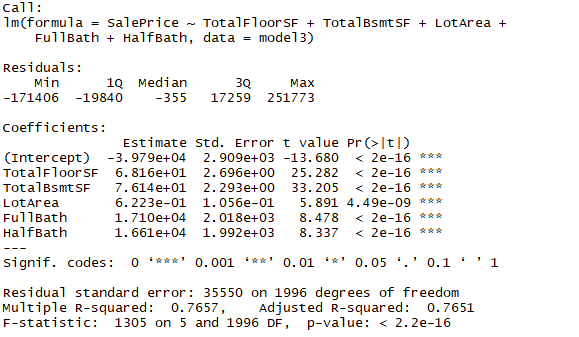
**Model 3:**

(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.

Yes, there is a way to logically group those varaibles into 2 or more sets of explanatory variables. I created this separation because I can see if those variables can be correlated to each other. We don’t want to see the predictors correlate to each other.

* Size: TotalFloorSF, TotalBsmtSF, LotArea, FullBath, HalfBath
* Quality: OverallQual, YearRemodel, QualityIndex, HouseAge, price\_sqft

1. 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:



1. all model coefficients individually

* H0: B1 = B2 = B3 = B4 = B5= 0; there is no relationship between predictors and a response Y varaible
* Ha: there is at least 1 inequality

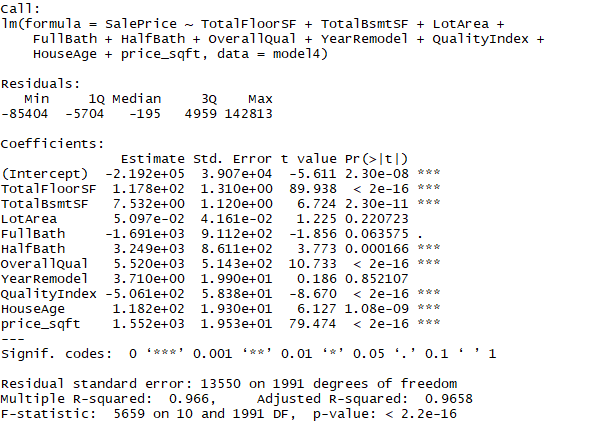
b) the Omnibus Overall F-test

F-Statistic: 1305; p-value is very small

When there is a high F-statitics, there will be very small p-value. We can see that the F-statistics is high which we reject null hypotheses and conclude that this model is a significant and there are at least one predictor correlates to response variable Y.

**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. In other words, Model 3 should be nested within Model 4. . Run a multiple regression model using the explanatory variables from this set to predict SALEPRICE(Y). Conduct and interpret the following hypothesis tests, being sure you clearly state the null and alternative hypotheses in each case:



1. all model coefficients individually

* Ho: B5 = B6 = B7 = B8 =B9 = B10 = 0
* Ha: there is at least 1 inequality

1. the Omnibus Overall F-test

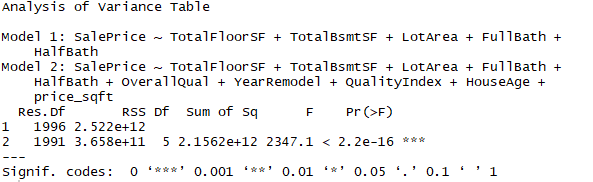
F-Statistic: 5959; p-value is very small

When there is a high F-statitics, there will be very small p-value. We can see that the F-statistics is high which we reject null hypotheses and conclude that this model is a significant and there are at least one predictor correlates to response variable Y.

**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 Model 4 variables, as a set, are useful for predicting SALEPRICE or not. Compute the F-statistic for this nested F-test and interpret the results.

* Ho: B5 = B6 = B7 = B8 =B9 = B10 = 0
* Ha: there is at least 1 inequality



According to F- statistics 2347 and very small p-value, we reject null hypothesis and conclude that the reduced model is not a good enough; Full model is significant.

**Assignment Document:**

Results should be presented and discussed in the numerical order of the questions given. The report should not contain unnecessary results or information. Tables are highly effective for summarizing data across multiple models. The document MUST be submitted in pdf format. Please use the naming convention: CompAssign2\_YourLastName.pdf.