Robin Singh



|  |
| --- |
| Computational Assignment #2: 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)

Table

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated with medium confidence



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

Observations = Total DF + 1 = 71 + 1 = **72**

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

Null: Beta1 = 0

Alt: Beta1 ≠ 0

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

T-stat for Beta1: 2.186/0.4104 = **5.3265**

99% Confidence, 2-tail Test: T stat with alpha=0.01 = 2.65

5.3265>2.65

**Reject null hypothesis with 99% confidence because our calculated t-value is in the rejection region, so there is enough evidence to discard the null hypothesis.**

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

SS Regression= 2126

SS Residual = 630.36

R^2= SS Regresion /(SS Residual+SS Regression)= 2126 / (2126+630.36) = **0.7713**

0.7713, or 77.13%, is the proportion of the variance of the dependent variable predictable from the independent variable. This indicates the model is quite accurate and well at predictions since it is higher than 0.5 or 50%.

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

n=72, k (predictor values or features) = 4

Adjusted R^2 = 1-(((1-R^2)\*(n-1))/(n-k-1)

=1 - (((1-0.7713) \* (72-1)) / (72-4-1)) = **0.7576**

The adjusted R^2 value accounts for the artificially increased R^2 value that occurs when adding more features to the model. This is because the more features you add the more the R^2 will increase, artificially making it more accurate even if the feature did not do much to increase the accuracy of the model.

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

Null: Beta1= Beta2= Beta3= Beta4 =0

Alt: At least one of the Beta’s does not equal the same as others (zero)

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

F = (2126/4) / (630.36/67) = **56.4923**

Hypothesis test with 99% Confidence: p<0.0001

There is not enough evidence to reject the null hypothesis as dictated by the small p-value. We cannot say that one of the Beta’s is not equal to zero.

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

Table

Description automatically generated

Table

Description automatically generated



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 by Model 2 because Model 2 includes all of the features from Model 1 and has more. Hence, Model 1 has a subset of features from Model 2.

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

Null: Beta5=Beta6=0

Alt: Beta5≠Beta6≠0

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 = ((630.36-572.6091)/(6-4))/(572.6091/(72-6))=**3.3282**

95% confidence interval stat=3.1504

Since 3.3282>3.1504, we can reject the null hypothesis that both added features in Model 2 are not very useful in comparison to Model 1.

\*\*\*FOR THIS SECTION I USED MY OWN CLEANED DATA WHICH WE HAD CREATED AFTER THE EDA IN MODELLING ASSIGNMENT 1\*\*\*

**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.

Set1: YearBuilt,YearRemodel

This set tells us what generation this house belonged to and how old it might be.

Set2: QualityIndex, OverallCond, OverallQual, TotalFloorSF, LotArea, LotFrontage, GrLivArea, BedroomAbvGr, YearBuilt, YearRemodel

This set would state how big the house is as well as its quality. It includes the previous Year variables as well as that is an important variable when analyzing quality, due to age having an effect, and remodeling also improving quality as well.

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:

\*\*\*FOR THIS SECTION I USED MY OWN CLEANED DATA WHICH WE HAD CREATED AFTER THE EDA IN MODELLING ASSIGNMENT 1\*\*\*

Text

Description automatically generated

1. all model coefficients individually

YearBuilt:

Null: Beta1=0

Alt: Beta1≠0

t-stat: Est./Std. Err. = 8.613e+02/ 4.910e+01= 17.54

95% conf: t=1.96

17.54>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that YearBuilt does have a relationship to SalePrice and can explain its variance.

YearRemodel:

Null: Beta2=0

Alt: Beta2≠0

t-stat: Est./Std. Err. = 1.058e+03/ 7.113e+01= 14.88

95% conf: t=1.96

14.88>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that YearRemodel does have a relationship to SalePrice and can explain its variance.

1. the Omnibus Overall F-test

Null: Beta1=Beta2 =0

Alt: At least one of the betas does not equal 0

ssr: 3871025802166.74

sse: 7168387346813.78

sst: 11039413148980.7

F-statistic: 612.1 on 2 and 2267 DF, p-value: < 2.2e-16

There is not enough evidence to reject the null hypothesis as dictated by the small p-value. We cannot say that one of the Beta’s is not equal to zero.

Table

Description automatically generated

**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:

Graphical user interface, text

Description automatically generated

1. all model coefficients individually

QualityIndex:

Null: Beta1=0

Alt: Beta1≠0

t-stat = Est./Std. Err. = -2.003e+03/ 4.682e+02= -4.278or abs=4.278

95% Confidence: t(0.025)=1.96

4.278>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that QualityIndex does have a relationship to SalePrice and can explain its variance.

OverallCond:

Null: Beta2=0

Alt: Beta2≠0

t-stat: Est./Std. Err. = 1.696e+042.735e+03 = 6.200

95% conf: t=1.96

6.200>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that OverallCond does have a relationship to SalePrice and can explain its variance.

OverallQual:

Null: Beta3=0

Alt: Beta3≠0

t-stat: Est./Std. Err. = 2.969e+04/ 2.684e+03= 11.059

95% conf: t=1.96

11.059>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that OverallQual does have a relationship to SalePrice and can explain its variance.

GrLivArea:

Null: Beta4=0

Alt: Beta4≠0

t-stat = Est./Std. Err. = 5.427e+01/ 1.710e+01= 3.173

95% Confidence: t(0.025)=1.96

3.173>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that GrLivArea does have a relationship to SalePrice and can explain its variance.

LotArea:

Null: Beta5=0

Alt: Beta5≠0

t-stat: Est./Std. Err. = 2.438e+00/ 1.398e-01= 17.448

95% conf: t=1.96

17.448>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that LotArea does have a relationship to SalePrice and can explain its variance.

TotalFloorSF:

Null: Beta6=0

Alt: Beta6≠0

t-stat: Est./Std. Err. = 2.274e+01/ 1.715e+01= 1.326

95% conf: t=1.96

1.326<1.96

We cannot reject the null hypothesis as our calculate t-stat is less than the threshold, meaning there is potential that TotalFloorSF does not have a relationship to SalePrice and can explain its variance.

BedroomAbvGr:

Null: Beta7=0

Alt: Beta7≠0

t-stat: Est./Std. Err. = -1.406e+04/ 9.561e+02= -14.706 or abs= 14.706

95% conf: t=1.96

14.706>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that BedroomAbvGr does have a relationship to SalePrice and can explain its variance.

LotFrontage:

Null: Beta8=0

Alt: Beta8≠0

t-stat: Est./Std. Err. = 1.428e+02/1.843e+01=7.747

95% conf: t=1.96

7.747>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that LotFrontage does have a relationship to SalePrice and can explain its variance.

YearBuilt:

Null: Beta9=0

Alt: Beta9≠0

t-stat: Est./Std. Err. = 5.367e+02/ 3.304e+01= 16.242

95% conf: t=1.96

16.242>1.96

We can reject the null hypothesis as our calculate t-stat is greater than the threshold, meaning there is potential that YearBuilt does have a relationship to SalePrice and can explain its variance.

YearRemodel:

Null: Beta10=0

Alt: Beta10≠0

t-stat: Est./Std. Err. = 7.238e+01/4.140e+01= 1.748

95% conf: t=1.96

1.748<1.96

We cannot reject the Null hypothesis because the calculated t-statistic is smaller than the 95% confidence t-value. It may potentially have no effect on SalePrice and hence can be removed.

1. the Omnibus Overall F-test

Null: Beta1=Beta2=Beta3=Beta4=Beta5=Beta6=Beta7=Beta8=Beta9=Beta10=0

Alt: At least one of the betas does not equal 0

F-statistic: 1152 on 10 and 2259 DF, p-value: < 2.2e-16

There is not enough evidence to reject the null hypothesis as dictated by the small p-value. We cannot say that one of the Beta’s is not equal to zero.

**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.

Null: Beta1=Beta2=Beta3=Beta4=Beta5=Beta6=Beta7=Beta8=0 (Of second model, or Model 4, for reference)

Alt: Neither Beta9 nor Beta10 equal 0

sse: 7168387346813.78

sse2: 1810018013415.84

F=((7168387346813.78-1810018013415.84)/(10-2))/( 1810018013415.84/(2270-10))

F=836.311752405

99%Confidence interval = For df1=5, df2=2290: 2.321

We reject the null hypothesis because the calculated F value is significantly greater than the table F value, meaning it is in the rejection range. This means that the other variables do have an impact on the SalesPrice Variable.

**Conclusion:**

This assignment allowed me to become familiar with the ANOVA tables and its aspects, as well as learn how the various calculations are done, and what they mean. It is important to determine whether or not a feature is beneficial to your model of not, as only then can it be improved in terms of accuracy. I also learned that with more features, the R^2 value rises artificially and that the adjusted R^2 value is what matters to take into account the added features and not over inflate the value. Through hypothesis testing with t and F statistics we can verify is a feature was actually useful to the model or not.

**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.