Reed Ballesteros
MSDS-410-DL, Summer 2022
Dr. Mickelson
7/10/2022

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

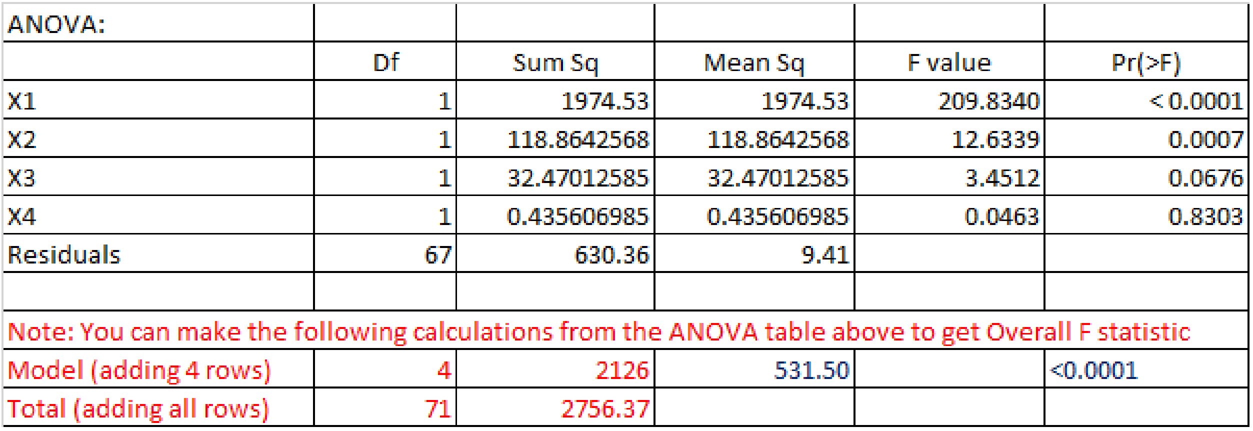
Results should be presented, labeled, and discussed in the numerical order of the questions given. Please use MS-WORD or some other text processing software to record and present your answers and results. The report should not contain unnecessary results or information. Tables are highly effective for 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)

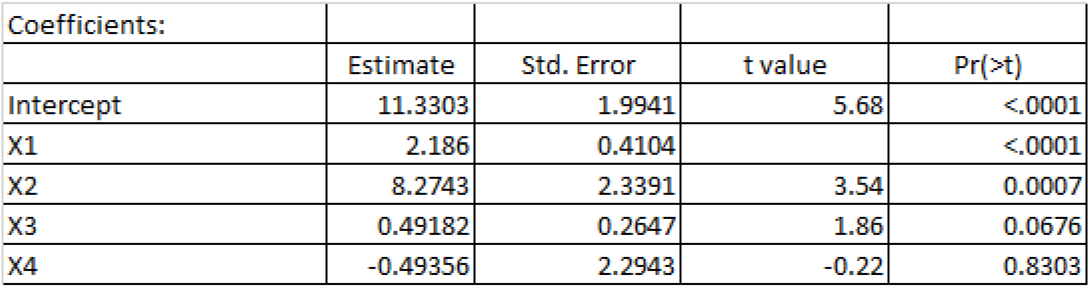
Model 1

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;

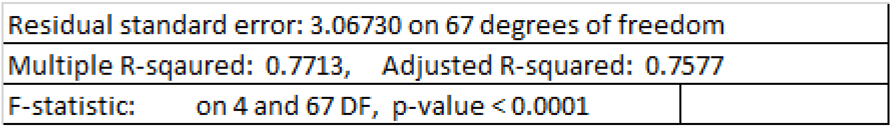
Note 2: The F test corresponding to the Model denotes the overall significance test. In R output, you will see that at the bottom of the Coefficients table)



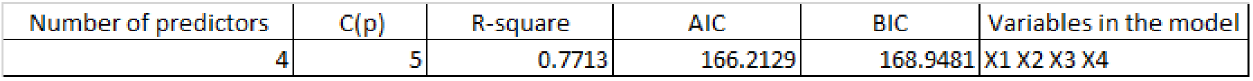
ANOVA:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1974.53	1974.53	209.8340	< 0.0001
X2	1	118.8642568	118.8642568	12.6339	0.0007
X3	1	32.47012585	32.47012585	3.4512	0.0676
X4	1	0.435606985	0.435606985	0.0463	0.8303
Residuals	67	630.36	9.41		
Note: You can make the following calculations from the ANOVA 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?
N = Model Df + Residuals Df + 1 = 4 + 67 + 1 = 72

(2) (3 points) Write out the null and alternate hypotheses for the t-test for Beta1.
H_0: Beta1 = 0
H_A: Beta1 != 0

(3) (3 points) Compute the t- statistic for Beta1. Conduct the hypothesis test and interpret the result.
X1 Estimate/X1 Std. Error = 2.186/0.4104 = 5.3625
p-value = T-Stat 5.3625 with 67 degrees of freedom = 0.000001
Since p-value = 0.000001, Beta1 with coefficient 2.186 is statistically significant. Thus, we reject the null hypothesis H_0.

(4) (3 points) Compute the R-Squared value for Model 1, using information from the ANOVA table. Interpret this statistic.
R-Squared = Sum Sq Model/Sum Sq Total = 2126/2756.37 = 0.7713
With R-Squared = 0.7713, the coefficients in Model 1 represents 77.13% of the variation of the model overall.

(5) (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 = 1 - ((Sum Sq Residual/df Residuals)/(Sum Sq Total/df Total))
= 1 – ((630.36/67)/(2756.37/71)) = 1 – (9.4084/38.8221)
= 1 - 0.2423 = 0.7577
The R-Squared value increases in multiple linear regression regardless if insignificant independent variables are added to a model (even if just a very small amount), while the adjusted R-Squared value only increases if significant independent variables are added. Statistically insignificant independent variables can actually lower the adjusted R-Squared value if added to a model.

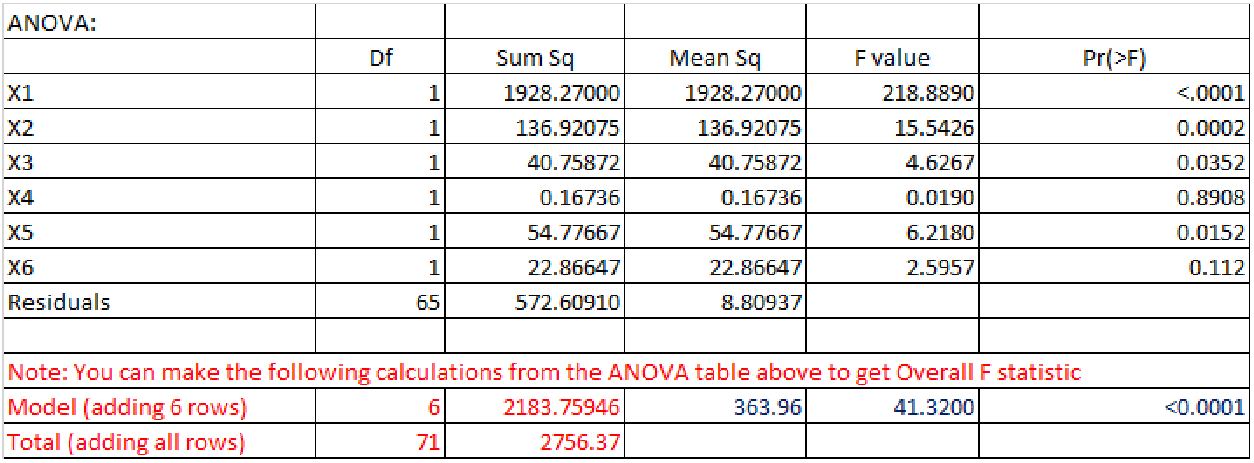
(6) (3 points) Write out the null and alternate hypotheses for the Overall F-test.
H_0: Beta1 = Beta2 = Beta3 = Beta4 = 0
H_A: At least one out of Beta1, Beta2, Beta3, Beta4 is not 0

(7) (3 points) Compute the F-statistic for the Overall F-test. Conduct the hypothesis test and interpret the result.
F-Statistic = Mean Sq Total (Regression)/Mean Sq Residuals (Error)
= 531.50/9.41 = 56.4825
Also, **F-Statistic** = (SSR/k)/(SSE/(n-(k+1)), N = 72, k = 4
= (2126/4)/(630.36/(72-(4+1))
= (2126/4)/(630.36/(72-5)
= (2126/4)/(630.36/67)
= 531.50/9.41
= 56.4825

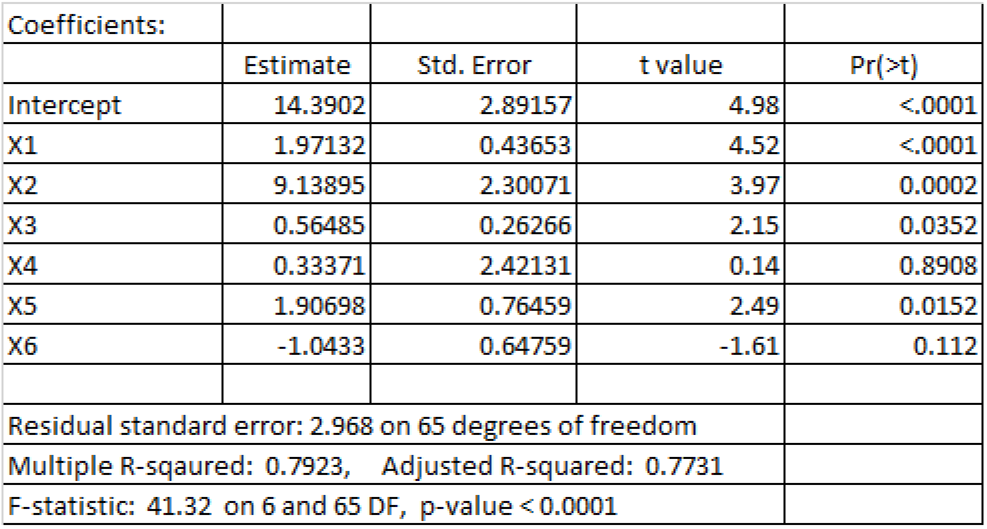
Since the F-Statistic much greater than 1 (1 is the reference value under the null hypothesis where all Betas = 0), we can reject the null hypothesis such that the F-Statistic is high that the sample statistics are unlikely to happen by chance alone while assuming the null hypothesis is true.

Model 2

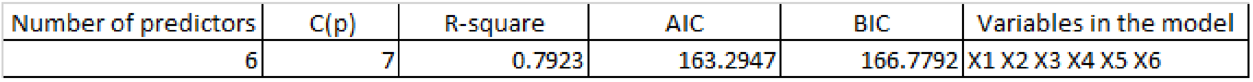
Now let's consider the following R output for an alternative regression model which we will refer to as Model 2.



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.

Model 2 nests Model 1, as Model 2 is the full model (FM) with six variables X1 to X6, while Model 1 is the reduced model (RM) with variables X1 to X4.

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

H_0: Beta5 = Beta6 = 0 (i.e., Model 1, the reduced model, is just as good, or better, than Model 2, the full model)

H_A: At least one of Beta5 or Beta6 is not zero (i.e., Model 2, the full model, is better than Model 1)

(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.

$$\mathbf{F_0} = \frac{(SSE(RM)-SSE(FM))/(dim(FM) - dim(RM)}{SSE(FM)/(n - dim(FM))}$$

$$= \frac{(630.36 - 572.6091)/(6-4)}{572.6091/(72-6)}$$

$$= \frac{(57.7509/2)}{(572.6091/66)}$$

$$= 28.87545/8.6759 = \mathbf{3.3282}$$

With an F-Statistic of 3.3282, if we use a 95% confidence level, we would barely edge out the critical F-value of 3.1359 with df1=2 (df(FM)-df(RM)) and df2=66 (df(RM)), in which the F-Statistic is GREATER THAN the F-critical value and can (barely!) reject the null hypothesis H_0.

If we use a 97.5% confidence level, the critical F-value would be 3.7965 in which the F-statistic would be LESS THAN the F-critical value, and we cannot reject the null hypothesis H_0 at a 97.5% confidence level.

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.

Before we can organize our sets of quantitative variables, we need to perform both data cleanup and a waterfall dropdown.

Cleanup:

- Correct GarageCars with <NA> values to 0
- Correct MasVnrArea with <NA> values to 0
- Correct TotalBsmtSF with <NA> values to 0
- Correct TotRmsAbvGrd with <NA> values to 0
- Correct FullBath with <NA> values to 0

Waterfall dropdown:

- Narrow population to only single-family homes (BldgType = '1Fam')
- Remove GarageCars outlier with GarageCars = 5 (doesn't match with SalePrice)
- Remove LotArea outliers (3) with LotArea > 100000 sq ft (doesn't match with SalePrice)
- Remove TotRmsAbvGrd outlier with TotRmsAbvGrd = 15 sq ft (doesn't match with SalePrice)
- Remove FullBath outliers (7) with FullBath = 0 (extremely rare to find 0 bath in a single family home)

With the above performed, we will make the following sets of variables to SalePrice:

Set 1: strictly based on highly correlated sets of areas (in square feet):

- TotalFloorSF
- TotalBsmtSF
- LotArea
- MaxVnrArea

Set 2: strictly based on highly correlated sets of discrete quantitative values:

- TotRmsAbvGrd
- FullBath
- GarageCars

- YearBuilt
- OverallQual
- OverallCond

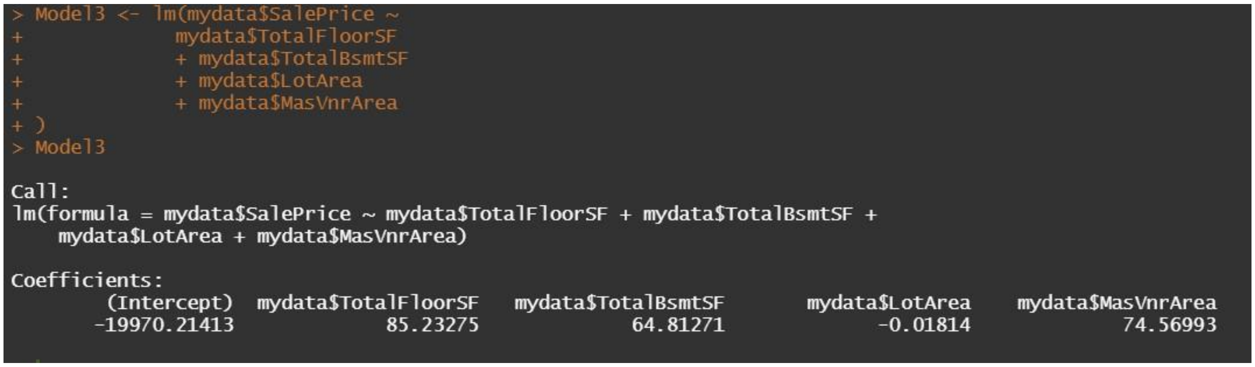
To note, with the exception of YearBuilt, the range of this selection of discrete variables are very small.

(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

The coefficients for Model 3 are the following:



```
> Model3 <- lm(mydata$SalePrice ~
+             mydata$TotalFloorSF
+             + mydata$TotalBsmtSF
+             + mydata$LotArea
+             + mydata$MasVnrArea
+ )
> Model3

Call:
lm(formula = mydata$SalePrice ~ mydata$TotalFloorSF + mydata$TotalBsmtSF +
    mydata$LotArea + mydata$MasVnrArea)

Coefficients:
        (Intercept)  mydata$TotalFloorSF   mydata$TotalBsmtSF       mydata$LotArea    mydata$MasVnrArea
       -19970.21413             85.23275             64.81271             -0.01814             74.56993
```

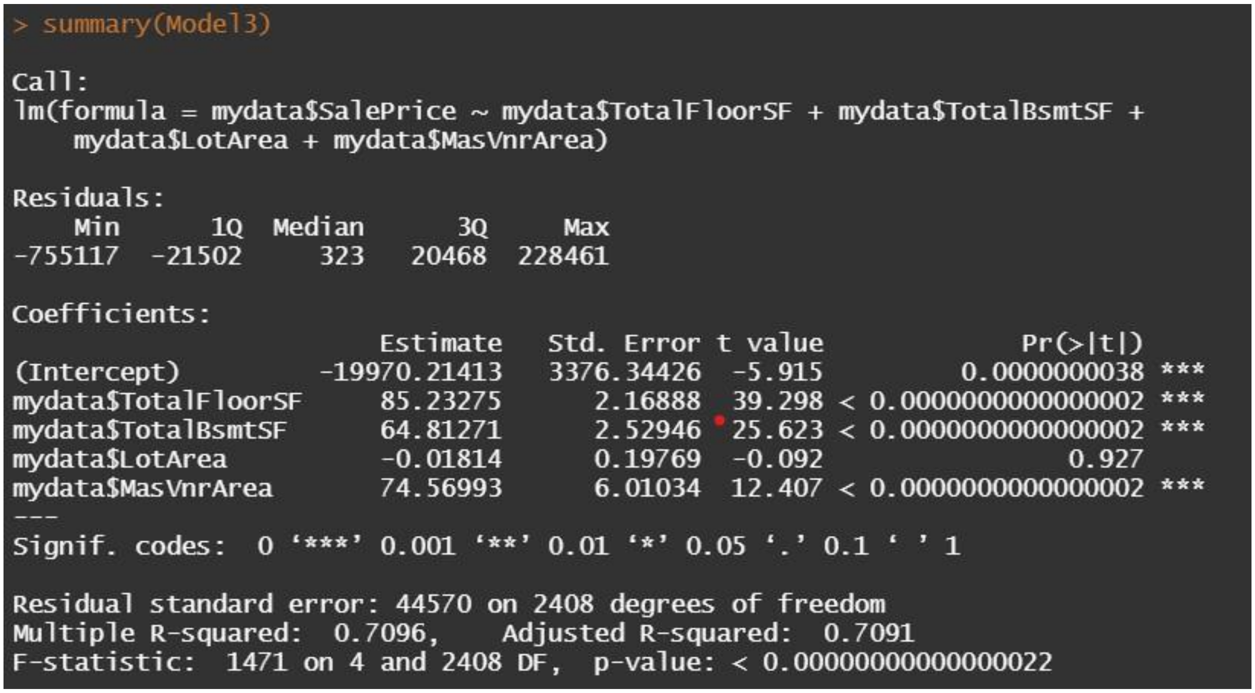
The multiple linear regression equation is:

- y = -19970.21413 + 85.2328*x1 + 64.8127*x2 – 0.01814*x3 + 74.56993*x4

Where:

- y = SalePrice
- x1 = TotalFloorSF (Beta1)
- x2 = TotalBsmtSF (Beta2)
- x3 = LotArea (Beta3)
- x4 = MasVnrArea (Beta4)

Summary for Model 3:



```
> summary(Model3)

Call:
lm(formula = mydata$SalePrice ~ mydata$TotalFloorSF + mydata$TotalBsmtSF +
    mydata$LotArea + mydata$MasVnrArea)

Residuals:
    Min      1Q  Median      3Q     Max
-755117  -21502     323   20468  228461

Coefficients:
                        Estimate   Std. Error t value            Pr(>|t|)
(Intercept)          -19970.21413   3376.34426  -5.915        0.0000000038 ***
mydata$TotalFloorSF      85.23275      2.16888  39.298 < 0.0000000000000002 ***
mydata$TotalBsmtSF       64.81271      2.52946  25.623 < 0.0000000000000002 ***
mydata$LotArea           -0.01814      0.19769  -0.092               0.927
mydata$MasVnrArea        74.56993      6.01034  12.407 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44570 on 2408 degrees of freedom
Multiple R-squared:  0.7096,    Adjusted R-squared:  0.7091
F-statistic:  1471 on 4 and 2408 DF,  p-value: < 0.00000000000000022
```

Null Hypothesis (H_0):

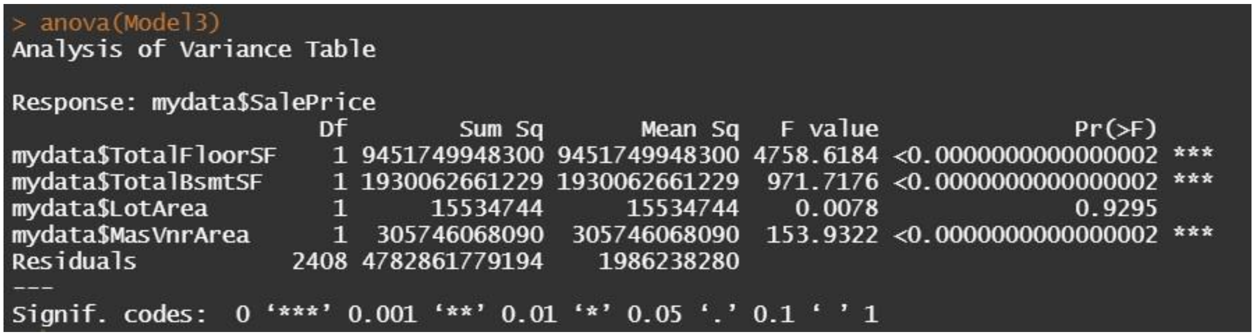
- **Beta1 = Beta2 = Beta3 = Beta4 = 0**

Alternate Hypothesis (H_A):

- **At least one of the Betas above is not zero.**

The R-Squared value is 0.7096, in which these above set of variables together represents 70.1% of the variability towards the dependent variable SalePrice after data cleanup and waterfall dropdown. The overall **F-statistic is 1471 based on 4 and 2408 degrees of freedom**, with a p-value of less than 0.0000000000000002, which overwhelmingly supports that we can reject the Model 3 null hypothesis H_0.

The analysis of variance (ANOVA) table presents the following:



```
> anova(Model3)
Analysis of Variance Table

Response: mydata$SalePrice
                      Df        Sum Sq       Mean Sq   F value              Pr(>F)
mydata$TotalFloorSF    1 9451749948300 9451749948300 4758.6184 <0.0000000000000002 ***
mydata$TotalBsmtSF     1 1930062661229 1930062661229  971.7176 <0.0000000000000002 ***
mydata$LotArea         1      15534744      15534744    0.0078              0.9295
mydata$MasVnrArea      1  305746068090  305746068090  153.9322 <0.0000000000000002 ***
Residuals           2408 4782861779194    1986238280
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Statistical analysis downplays LotArea as statistically insignificant with a very small F-Statistic of 0.0078 and a p-value of 0.9295 compared to the other variables in the set, mostly due to the wide range of values available in LotArea compared to the other variables. As we continue to work with the data, we can attempt to standardize or normalize the data to put them more in in line or relative to each other. As for the other kinds of square footage (TotalFloofSF, TotalBsmtSF, MasVnrArea), an extra square footage in each respective variable influences the SalePrice due to their higher slope values, as their range of values are much smaller compared to LotArea.

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

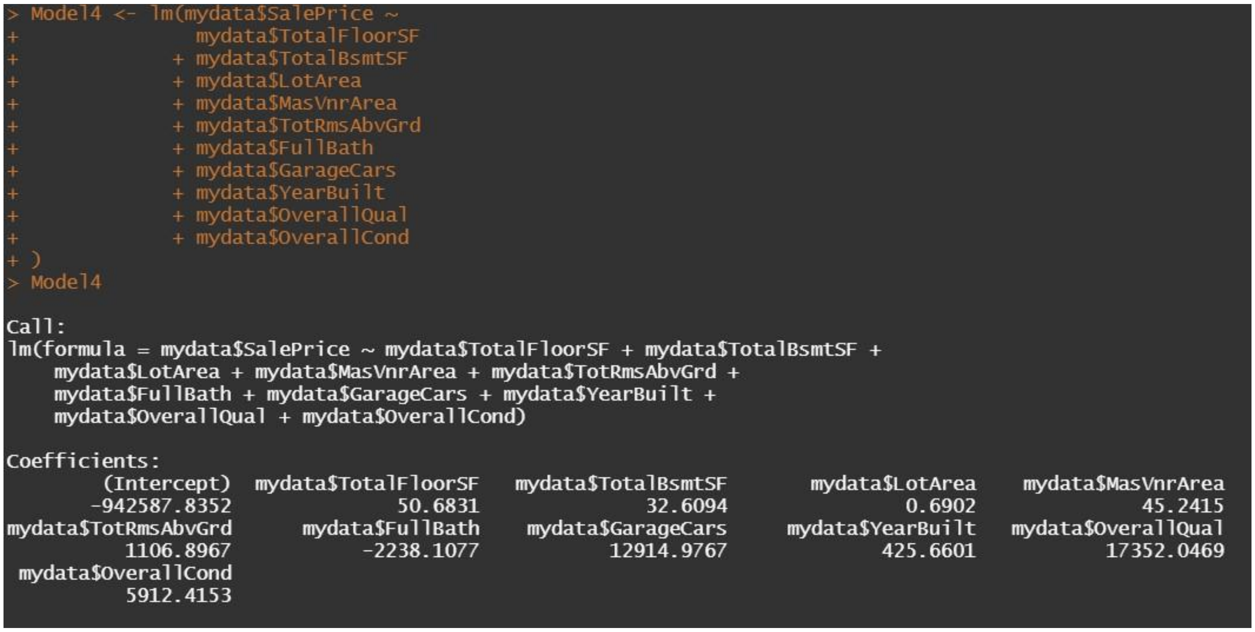
b) the Omnibus Overall F-test

Model 4 adds the following variables from Set 2 described in section (11):

- TotRmsAbvGrd
- FullBath
- GarageCars
- YearBuilt
- OverallQual
- OverallCond

Due to the additions, Model 3 is nested within Model 4.

The coefficients for Model 4 are the following:



```
> Model4 <- lm(mydata$SalePrice ~
+                mydata$TotalFloorSF
+              + mydata$TotalBsmtSF
+              + mydata$LotArea
+              + mydata$MasVnrArea
+              + mydata$TotRmsAbvGrd
+              + mydata$FullBath
+              + mydata$GarageCars
+              + mydata$YearBuilt
+              + mydata$OverallQual
+              + mydata$OverallCond
+ )
> Model4

Call:
lm(formula = mydata$SalePrice ~ mydata$TotalFloorSF + mydata$TotalBsmtSF +
    mydata$LotArea + mydata$MasVnrArea + mydata$TotRmsAbvGrd +
    mydata$FullBath + mydata$GarageCars + mydata$YearBuilt +
    mydata$OverallQual + mydata$OverallCond)

Coefficients:
        (Intercept)  mydata$TotalFloorSF   mydata$TotalBsmtSF       mydata$LotArea    mydata$MasVnrArea
       -942587.8352              50.6831              32.6094               0.6902              45.2415
mydata$TotRmsAbvGrd      mydata$FullBath    mydata$GarageCars     mydata$YearBuilt   mydata$OverallQual
          1106.8967           -2238.1077           12914.9767             425.6601           17352.0469
 mydata$OverallCond
          5912.4153
```

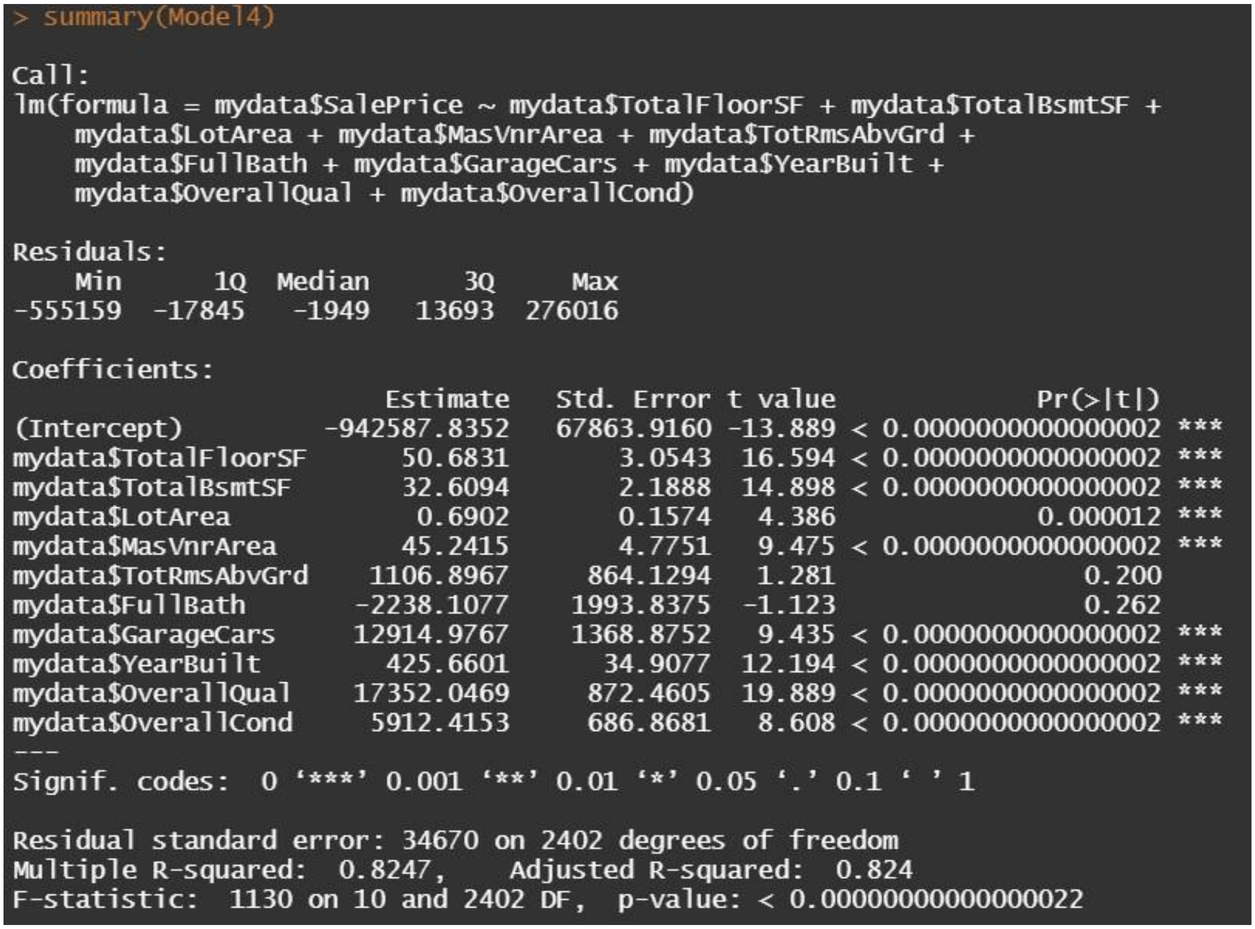
The multiple linear regression equation for Model 4 is:

- y = -942587.8352 + 50.6831*x1 + 32.6094*x2 + 0.6902*x3 + 45.2415*x4 + 1106.8967*x5 - 2238.1077*x6 + 12914.9767*x7 + 425.6601*x8 17352.0469*x9 + 5912.4153*x10

Where:

- y = SalePrice
- x1 = TotalFloorSF (Beta1)
- x2 = TotalBsmtSF (Beta2)
- x3 = LotArea (Beta3)
- x4 = MasVnrArea (Beta4)
- x5 = TotRmsAbvGrd (Beta5)
- x6 = FullBath (Beta6)
- x7 = GarageCars (Beta7)
- x8 = YearBuilt (Beta8)
- x9 = OverallQual (Beta9)
- x10 = OverallQual (Beta10)

Summary for Model 4:



```
> summary(Model4)

Call:
lm(formula = mydata$SalePrice ~ mydata$TotalFloorSF + mydata$TotalBsmtSF +
    mydata$LotArea + mydata$MasVnrArea + mydata$TotRmsAbvGrd +
    mydata$FullBath + mydata$GarageCars + mydata$YearBuilt +
    mydata$OverallQual + mydata$OverallCond)

Residuals:
    Min      1Q  Median      3Q     Max
-555159  -17845   -1949   13693  276016

Coefficients:
                       Estimate   Std. Error t value            Pr(>|t|)
(Intercept)         -942587.8352   67863.9160 -13.889 < 0.0000000000000002 ***
mydata$TotalFloorSF      50.6831       3.0543  16.594 < 0.0000000000000002 ***
mydata$TotalBsmtSF       32.6094       2.1888  14.898 < 0.0000000000000002 ***
mydata$LotArea            0.6902       0.1574   4.386             0.000012 ***
mydata$MasVnrArea        45.2415       4.7751   9.475 < 0.0000000000000002 ***
mydata$TotRmsAbvGrd    1106.8967     864.1294   1.281                0.200
mydata$FullBath       -2238.1077    1993.8375  -1.123                0.262
mydata$GarageCars     12914.9767    1368.8752   9.435 < 0.0000000000000002 ***
mydata$YearBuilt        425.6601      34.9077  12.194 < 0.0000000000000002 ***
mydata$OverallQual    17352.0469     872.4605  19.889 < 0.0000000000000002 ***
mydata$OverallCond     5912.4153     686.8681   8.608 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34670 on 2402 degrees of freedom
Multiple R-squared:  0.8247,    Adjusted R-squared:  0.824
F-statistic:  1130 on 10 and 2402 DF,  p-value: < 0.00000000000000022
```

Null Hypothesis (H_0):

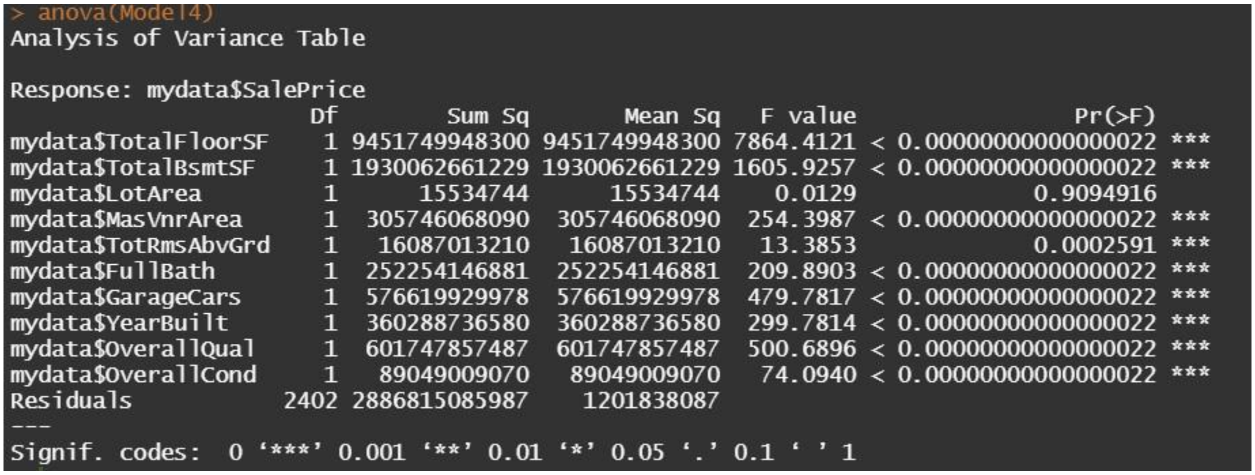
- **Beta1 = Beta2 = Beta3 = Beta4 = Beta5 = Beta6 = Beta7 = Beta8 = Beta9 = Beta10 = 0**

Alternate Hypothesis (H_A):

- **At least one the above Betas above is not zero.**

The R-Squared value is 0.8247, in which these above set of variables together represents a very high 82.5% of the variability towards the dependent variable SalePrice after data cleanup and waterfall dropdown. The overall **F-statistic is 1130 based on 10 and 2402 degrees of freedom**, with a p-value of less than 0.00000000000000022, which supports that we can also reject this Model 4 null hypothesis H_0.

The analysis of variance (ANOVA) table for Mode 4 presents the following:



```
> anova(Model4)
Analysis of Variance Table

Response: mydata$SalePrice
                    Df        Sum Sq       Mean Sq   F value                Pr(>F)    
mydata$TotalFloorSF    1 9451749948300 9451749948300 7864.4121 < 0.0000000000000022 ***
mydata$TotalBsmtSF     1 1930062661229 1930062661229 1605.9257 < 0.0000000000000022 ***
mydata$LotArea         1      15534744      15534744    0.0129             0.9094916    
mydata$MasVnrArea      1  305746068090  305746068090  254.3987 < 0.0000000000000022 ***
mydata$TotRmsAbvGrd    1   16087013210   16087013210   13.3853             0.0002591 ***
mydata$FullBath        1  252254146881  252254146881  209.8903 < 0.0000000000000022 ***
mydata$GarageCars      1  576619929978  576619929978  479.7817 < 0.0000000000000022 ***
mydata$YearBuilt       1  360288736580  360288736580  299.7814 < 0.0000000000000022 ***
mydata$OverallQual     1  601747857487  601747857487  500.6896 < 0.0000000000000022 ***
mydata$OverallCond     1   89049009070   89049009070   74.0940 < 0.0000000000000022 ***
Residuals           2402 2886815085987    1201838087                                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Similar to what we observed in Model 3, LotArea is also downplayed in Model 4 as statistically insignificant with a low F-Value of 0.0129 and a P-Value of 0.9095 and looks worse due to that we added several discrete quantitative variables with very low values and very small ranges. Going forward, we can attempt to normalize all these variables together into relative values and see if we can improve LotArea’s statistical significance to the multiple regression model.

The addition of discrete quantitative variables (TotRmsAbvGrnd, FullBath, GarageCars, YearBuilt, OverallQual, OverallCond) to Model 4 have an even greater dramatic effect to SalePrice as shown by their respective slopes due to their very small and limited range of values. An extra room, bathroom, or an extra car in the garage can heavily influence the sale price. A home that is just one year older than another home can negatively affect its sale price. A home that is just one rating above another home in overall quality or condition can have a significant impact on sale price.

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.

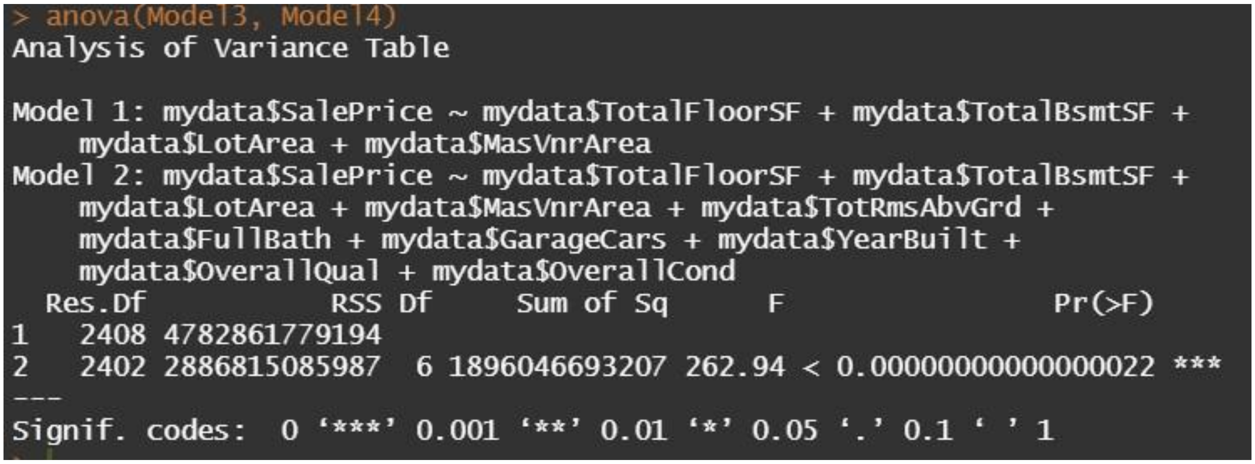
Null Hypothesis (H_0):

- **Beta5 = Beta6 = Beta7 = Beta8 = Beta9 = Beta10 = 0** (i.e., the added variables in Model 4)

Alternate Hypothesis (H_A):

- **At least one of the Betas above is not zero.**

The analysis of variance (ANOVA) table comparing Model 3 and Model 4 presents the following:



```
> anova(Model3, Model4)
Analysis of Variance Table

Model 1: mydata$SalePrice ~ mydata$TotalFloorSF + mydata$TotalBsmtSF +
    mydata$LotArea + mydata$MasVnrArea
Model 2: mydata$SalePrice ~ mydata$TotalFloorSF + mydata$TotalBsmtSF +
    mydata$LotArea + mydata$MasVnrArea + mydata$TotRmsAbvGrd +
    mydata$FullBath + mydata$GarageCars + mydata$YearBuilt +
    mydata$OverallQual + mydata$OverallCond
  Res.Df           RSS Df     Sum of Sq      F                Pr(>F)
1   2408 4782861779194
2   2402 2886815085987  6 1896046693207 262.94 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-Statistic of Model 3 nested within Model 4 is:

$$\mathbf{F_0} = \frac{SSE(RM)-SSE(FM)/(dim(FM)-dim(RM))}{SSE(FM)/(N-dim(FM))}$$, where dim(FM) = 10, dim(RM) = 4, N = 2413

$$= \frac{(4782861779194 - 2886815085987)/(10 - 4)}{2886815085987/(2413 - 10)}$$

$$= \frac{1896046693207/6}{2886815085987/2403}$$

$$= 316007782201/1201337947 = \mathbf{263.0465}$$

with a degrees of freedom difference of 6, and a P-Value of 0.00000000000000022.

Along with that we earlier reported that Model 3 represents 70.1% of the variability to Sale Price and Model 4 represents 82.5% of the variability, Model 4’s additions are statistically significant over Model 3 such that we can reject the null hypothesis H_0 with F_0 = 263.0465 and a P-Value very close to zero. The added complexity of six additional highly correlated variables makes Model 4 a much more robust regression model over Model 3.