Kay Quiballo | MSDS 410 | 01.17.2023

Modeling Assignment 3: Statistical Inference in Multiple Linear Regression

Assignment Overview

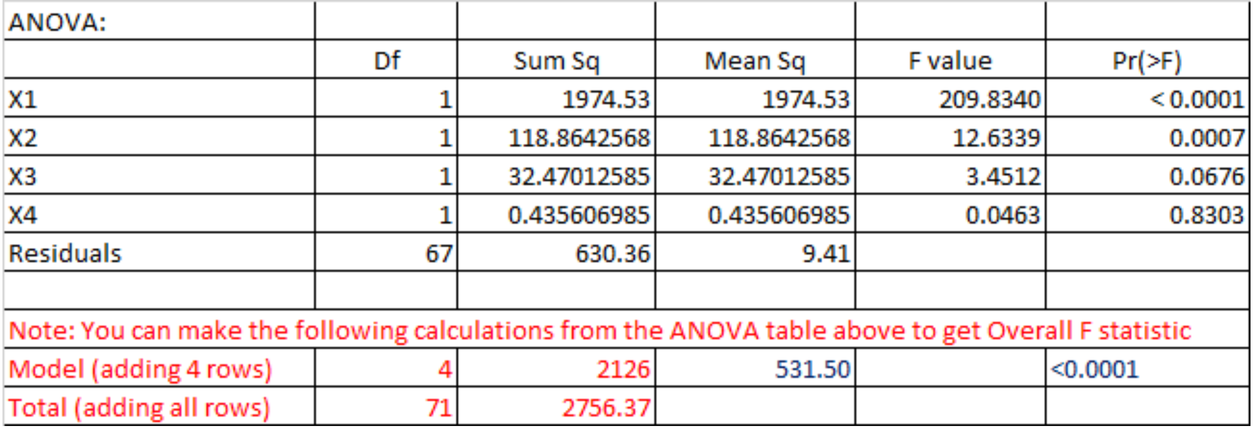
Write down the generic formula for any computation and then fill in the values need for the computation from the problem statement. Keep all decimals to four places, i.e. X.xxxx.

PART 1: MECHANICS AND COMPUTATIONS (30 points)

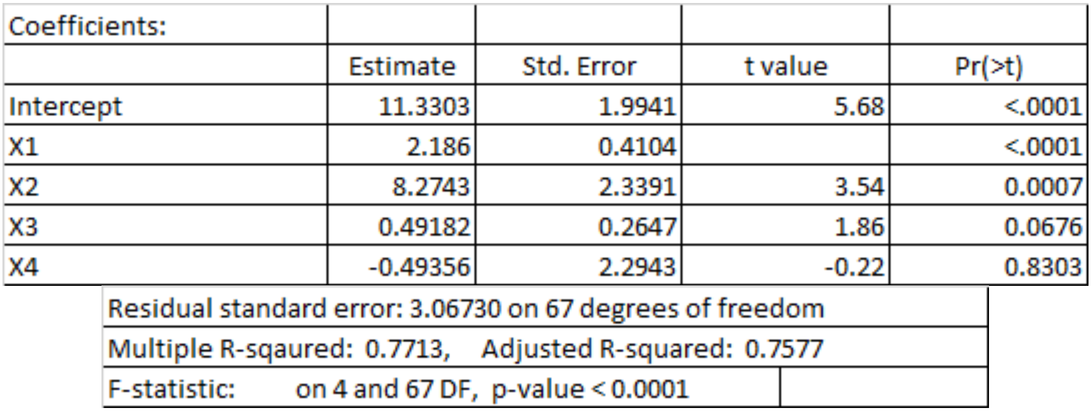
Model 1

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			

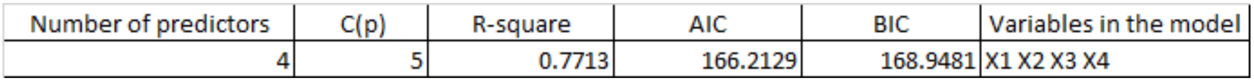


Coefficients:				
	Estimate	Std. Error	t value	Pr(>t)
Intercept	11.3303	1.9941	5.68	<.0001
X1	2.186	0.4104		<.0001
X2	8.2743	2.3391	3.54	0.0007
X3	0.49182	0.2647	1.86	0.0676
X4	-0.49356	2.2943	-0.22	0.8303

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?

DF_model = p = 4

DF_error = N – p – 1 = 72 – 4 – 1 = 67

DF_total = N-1 → 71 = N-1 → **N = 72**

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

H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$

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

T = (m – m0) / SE
T = (2.186 – 0) / 0.4104
T = 5.3265 → the p-val < 0.0001 so we reject H_0 and conclude that $\beta_1 \neq 0$

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

R2 = SSTR / SST
R2 = 2126 / 2756.37
R2 = 0.7713

77.13% of the variation in the response can be explained by the predictors of Model 1.

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

R2_adj = 1 - MSE / MST
R2_adj = 1 – (SSE/(n-p-1)) / (SST/(n-1))
R2_adj = 1 – (630.36/(67)) / (2756.37/(71))
R2_adj = 0.7577

R2_adj is a modified version of R2 that takes into account the number of predictors in a model. If the predictor added improves the model more than by mere chance, R2_adj increases. R2_adj will be less than or equal to R2 because it penalizes the model for including additional predictors.

(6) (3 points) Write out the null and alternate hypotheses for the Overall F-test.
H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
H_A: $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$ or $\beta_4 \neq 0$

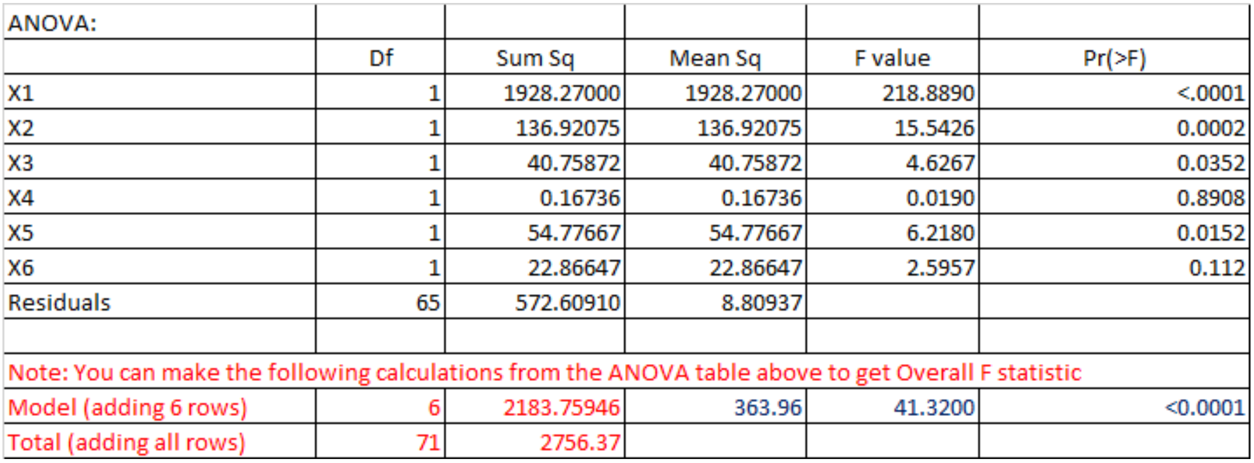
(7) (3 points) Compute the F-statistic for the Overall F-test. Conduct the hypothesis test and interpret the result.
F = MSR / MSE
F = (SSR / p) / (SSE / (n-p-1))
F = (2126 / 4) / (630.36 / 67)
F = 56.4923
p-val = 9.2e-21

We reject the null hypothesis and conclude that at least one of the coefficients in Model 1 is not equal to 0.

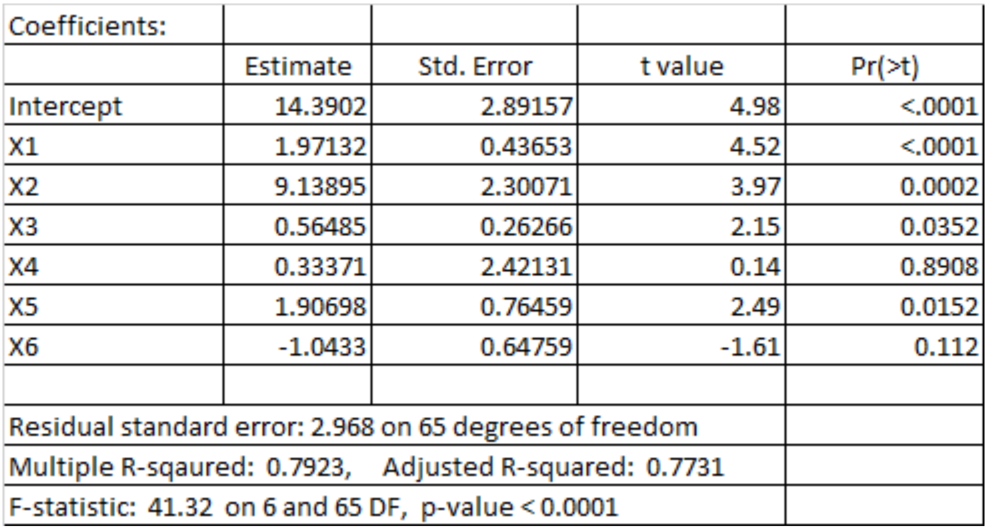
Kay Quiballo | MSDS 410 | 01.17.2023

Model 2

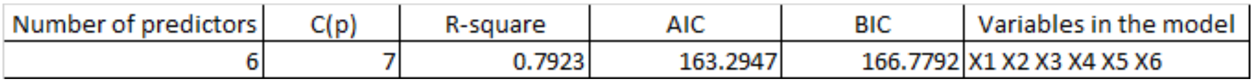
Now let’s consider the following R output for an alternative regression model which we will refer to as Model 2.



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 1 is nested in Model 2 because the predictors of Model 1 {X1, X2, X3, X4} are a subset of the predictors of Model 2 {X1, X2, X3, X4, X5, X6}.

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

H_0: $\beta_4 = \beta_5 = 0$
H_A: $\beta_4 \neq 0$ or $\beta_5 \neq 0$

(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.

DF_model = p = 6
DF_error = N – p – 1 = 72 – 6 – 1 = 65
DF_total = N-1
71 = N-1
N = 72

Note: SS1 is the nested model (M1)
F = [(SS1 – SS2) / (df1 – df2)] / [SS2 / df2]
F = [(630.36 - 572.60910) / ((72 – 4) – (72 – 6))] / [572.60910 / (72 – 6)]
F = (57.7509 / 2) / (572.60910 / 66)
F = 3.3282

p-val(df = (df1-df2, df2) = (2, 66), F=3.3282) = 0.0420

We reject the null hypothesis and conclude that β4 ≠ 0 or β5 ≠ 0

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.

The 10 continuous variables chosen were one with the highest correlation with SalePrice.

Qualitity variables – These variables indicate the quality of the house's components.
1) OverallQual: Rates the overall material and finish of the house
2) QualityIndex: OverallQual * OverallCond

Area/Space Variables – These variables indicate how large the house is or how many rooms.
3) GrLivArea: Above grade (ground) living area square feet
4) GarageCars: Size of garage in car capacity
5) GarageArea: Size of garage in square feet
6) FirstFlrSF: First Floor square feet
7) FullBath: Basement full bathrooms
8) TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Time Variables – These variables take into account the year the property was built or sold.
9) YearBuilt: Original construction date
10) HouseAge: YrSold - YearBuilt

(11) 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:

Model 3: SalePrice ~ OverallQual + QualityIndex
predicted(SalePrice) = β_0 + β_1*OverallQual + β_2*QualityIndex

a) all model coefficients individually

Model	**Null Hyp**	**Alt Hyp**	**P-val**
Model 3	$\beta_1 = 0$	$\beta_1 \neq 0$	<2e-16
Model 3	$\beta_2 = 0$	$\beta_2 \neq 0$	0.0066

With P-vals < 0.05 for all T-tests, we reject the null hypothesis and conclude that both coefficients from Model 3 are not 0.

b) the Omnibus Overall F-test

Null: $\beta_1 = \beta_2 = 0$
Alternative: $\beta_1 \neq 0$ or $\beta_2 \neq 0$
F-statistic: 1344 on 1 and 2928 DF, p-value: < 2.2e-16 (*code in appendix*)

With P-val < 0.05 for the F-test, we reject the null hypothesis and conclude that at least one of the coefficients of Model 3 is not 0.

Model 4

(12) 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:

Model 4: SalePrice ~ OverallQual + QualityIndex
predicted(SalePrice) = β_0 + β_1*OverallQual + β_2*QualityIndex
+ β_3* YearBuilt – β_4* HouseAge

a) all model coefficients individually

Model	**Null Hyp**	**Alt Hyp**	**P-val**
Model 4	$\beta_1 = 0$	$\beta_1 \neq 0$	<2e-16
Model 4	$\beta_2 = 0$	$\beta_2 \neq 0$	0.145
Model 4	$\beta_3 = 0$	$\beta_3 \neq 0$	0.434
Model 4	$\beta_4 = 0$	$\beta_4 \neq 0$	0.186

Based on the P-vals, only the coefficient with a P-val < 0.05 is for OverallQual. We reject the null hypothesis and conclude that the coefficient for OverallQual is not 0. We fail to reject the null hypothesis for all other coefficients and conclude that they are 0.

c) the Omnibus Overall F-test

Null: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
Alternative: $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$ or $\beta_4 \neq 0$
F-statistic: 1355 on 4 and 2925 DF, p-value: < 2.2e-16 (*code in appendix*)

With P-val < 0.05 for the F-test, we reject the null hypothesis and conclude that at least one of the coefficients of Model 4 is not 0.

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.

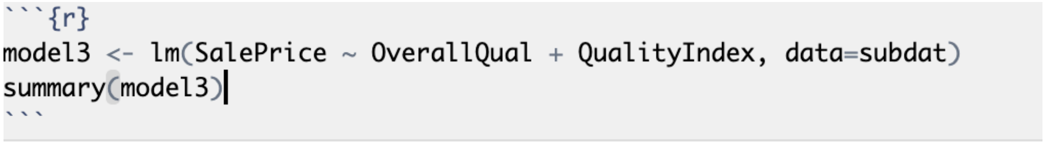
Null: $\beta_3 = \beta_4 = 0$
Alternative: $\beta_3 \neq 0$ or $\beta_4 \neq 0$

F-statistic: 40.876, p-value: < 2.2e-16 (*code in appendix*)

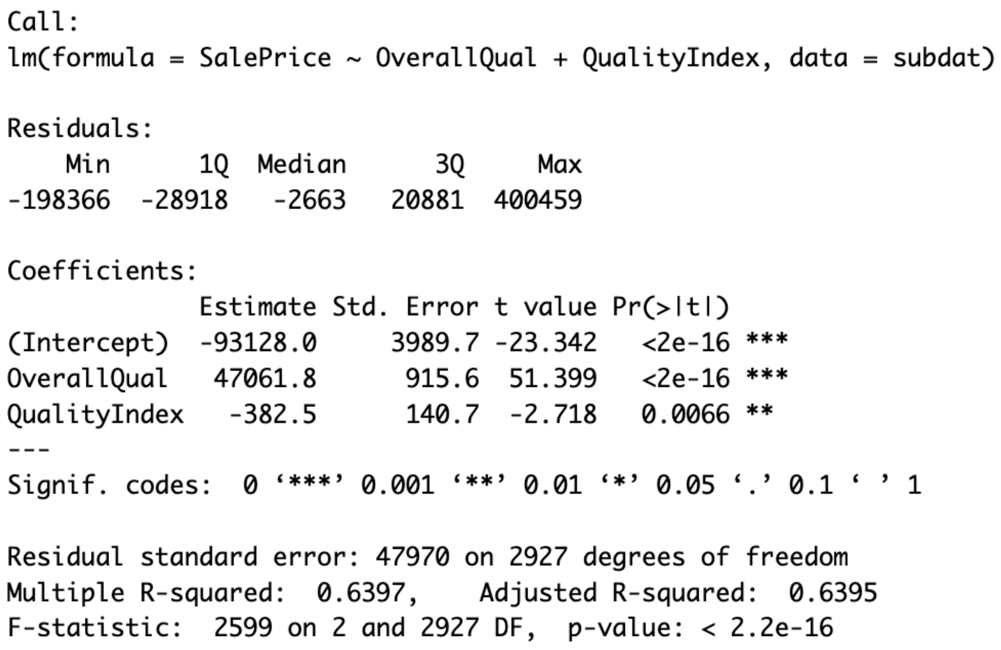
With P-val < 0.05 for the F-test, we reject the null hypothesis and conclude that at least one of the coefficients for YearBuilt or HouseAge is not 0.

Appendix

Code for Part II: Application



~~~
```{r}
model3 <- lm(SalePrice ~ OverallQual + QualityIndex, data=subdat)
summary(model3)
```
~~~

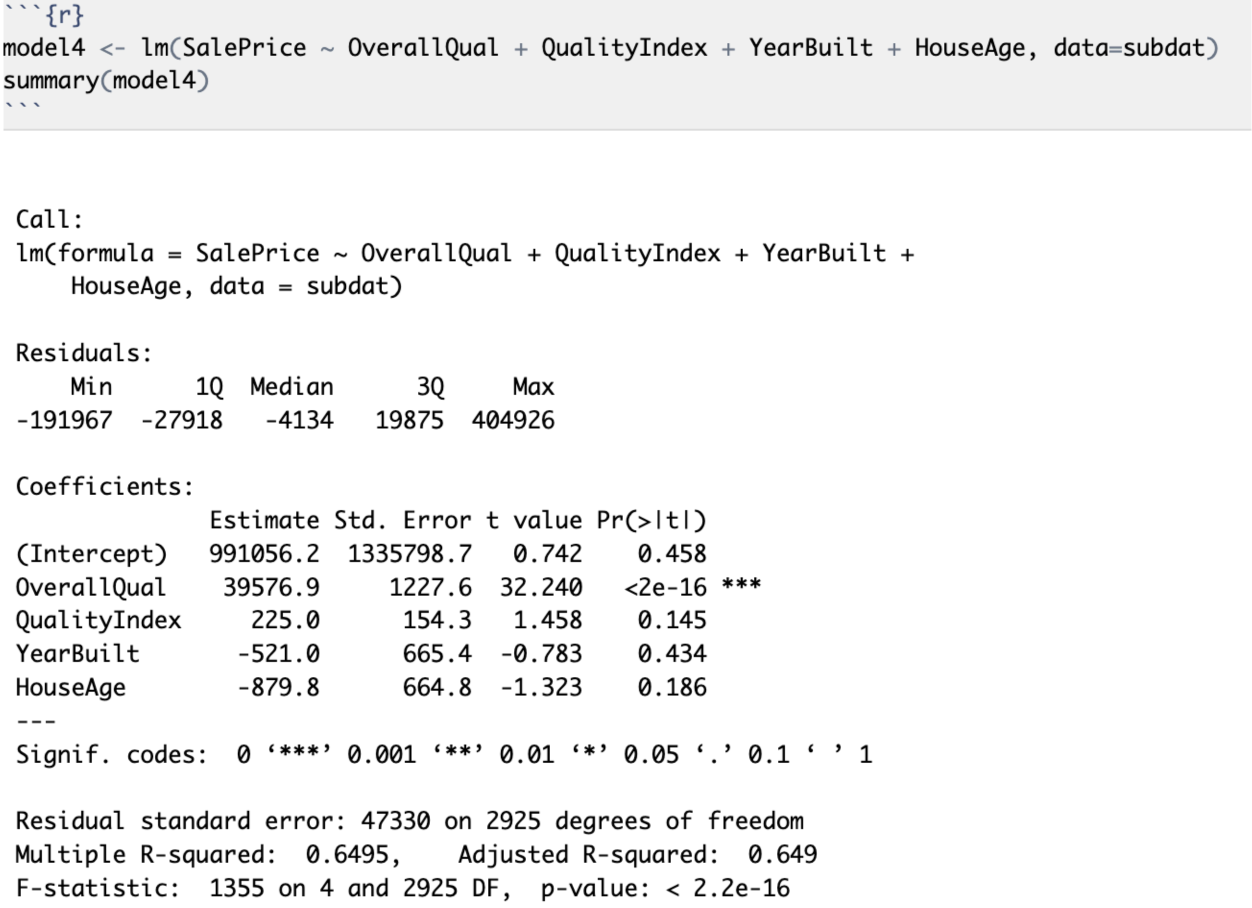


```
Call:
lm(formula = SalePrice ~ OverallQual + QualityIndex, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max 
-198366  -28918   -2663   20881  400459 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -93128.0     3989.7 -23.342   <2e-16 ***
OverallQual   47061.8      915.6  51.399   <2e-16 ***
QualityIndex   -382.5      140.7  -2.718   0.0066 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47970 on 2927 degrees of freedom
Multiple R-squared:  0.6397,	Adjusted R-squared:  0.6395 
F-statistic:  2599 on 2 and 2927 DF,  p-value: < 2.2e-16
```



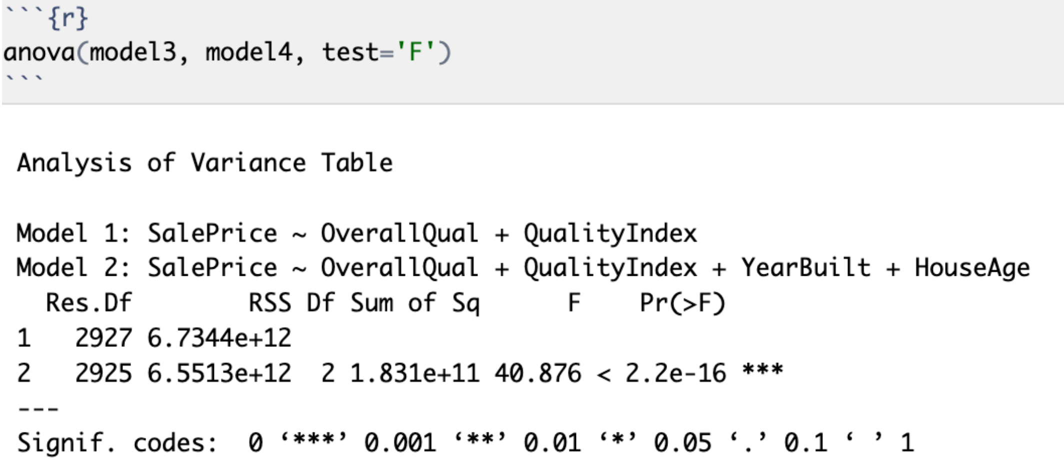
~~~
```{r}
model4 <- lm(SalePrice ~ OverallQual + QualityIndex + YearBuilt + HouseAge, data=subdat)
summary(model4)
```
~~~

```
Call:
lm(formula = SalePrice ~ OverallQual + QualityIndex + YearBuilt + 
    HouseAge, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max 
-191967  -27918   -4134   19875  404926 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   991056.2  1335798.7   0.742    0.458    
OverallQual    39576.9     1227.6  32.240   <2e-16 ***
QualityIndex     225.0      154.3   1.458    0.145    
YearBuilt       -521.0      665.4  -0.783    0.434    
HouseAge        -879.8      664.8  -1.323    0.186    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47330 on 2925 degrees of freedom
Multiple R-squared:  0.6495,	Adjusted R-squared:  0.649 
F-statistic:  1355 on 4 and 2925 DF,  p-value: < 2.2e-16
```



~~~
```{r}
anova(model3, model4, test='F')
```
~~~

```
Analysis of Variance Table

Model 1: SalePrice ~ OverallQual + QualityIndex
Model 2: SalePrice ~ OverallQual + QualityIndex + YearBuilt + HouseAge
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)    
1   2927 6.7344e+12                                  
2   2925 6.5513e+12  2 1.831e+11 40.876 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```