# MAT 303 Project One Summary Report

Camilo Hoyos

camilo.hoyos@snhu.edu

Southern New Hampshire University

## 1. Introduction

The data set that we are exploring contains real estate sales information. There are 2692 records of historical sales with varying attributes about the home, crime, and view. The response variable is going to be the price of a home. The results of this analysis would allow one to predict the price of a home being sold based on a handful of predictor variables. In this report we will run a first order regression model, a complete second order regression model, and a nested F-test to determine if the quadratic variables provide a more accurate model.

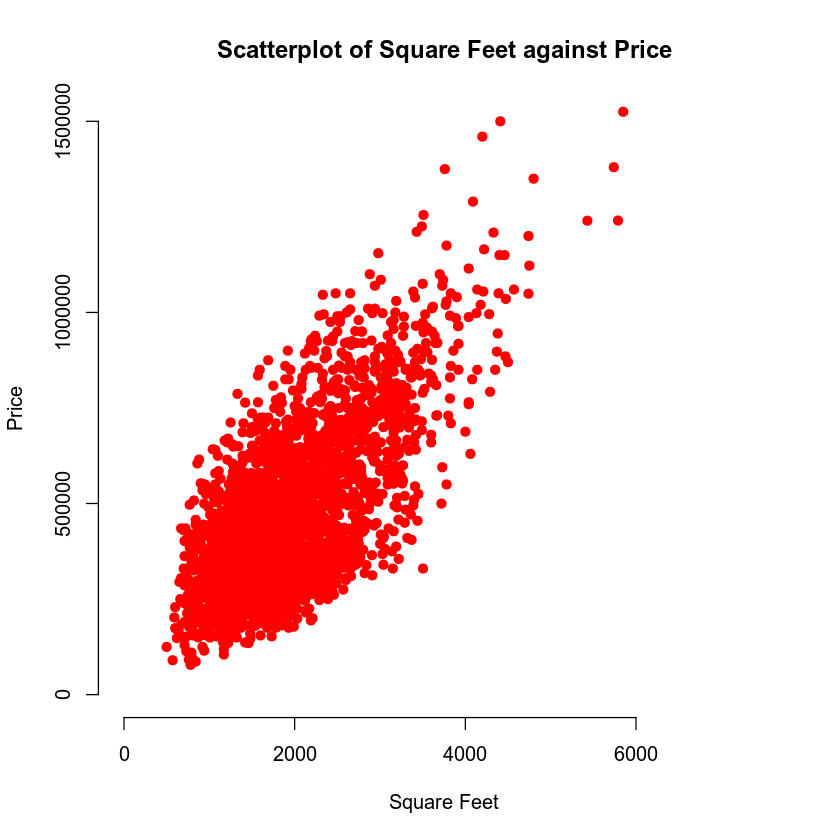
## 2. Data Preparation

The important variables in this data set are price, age, sqft\_living, grade, bathrooms, view, crime, and appliance\_age. Price will be the response variable and all other listed variables will be used as predictor variables. The only variable that will function as a qualitative variable is the view attribute. There are a total of 22 columns and 2692 rows.

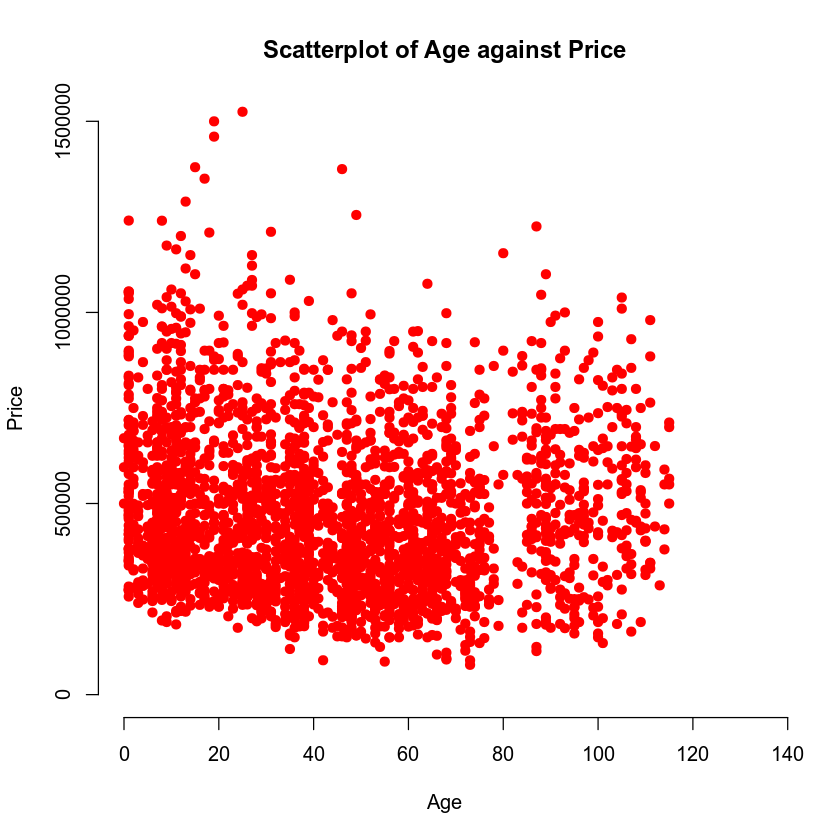
## 3. Model #1 - First Order Regression Model with Quantitative and Qualitative Variables

### Correlation Analysis

Price vs Living Area

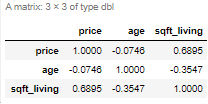


Price vs Age of Home



Price and livable square footage have a positive and strong correlation. The more livable square footage available, the more the house will cost. That cannot be said for the price against the age of a home. Varying ages achieve the same, lower, or higher price with a slight downwards trend. However, it is not a very strong relationship.

Correlation coefficient between price, age, and square feet.

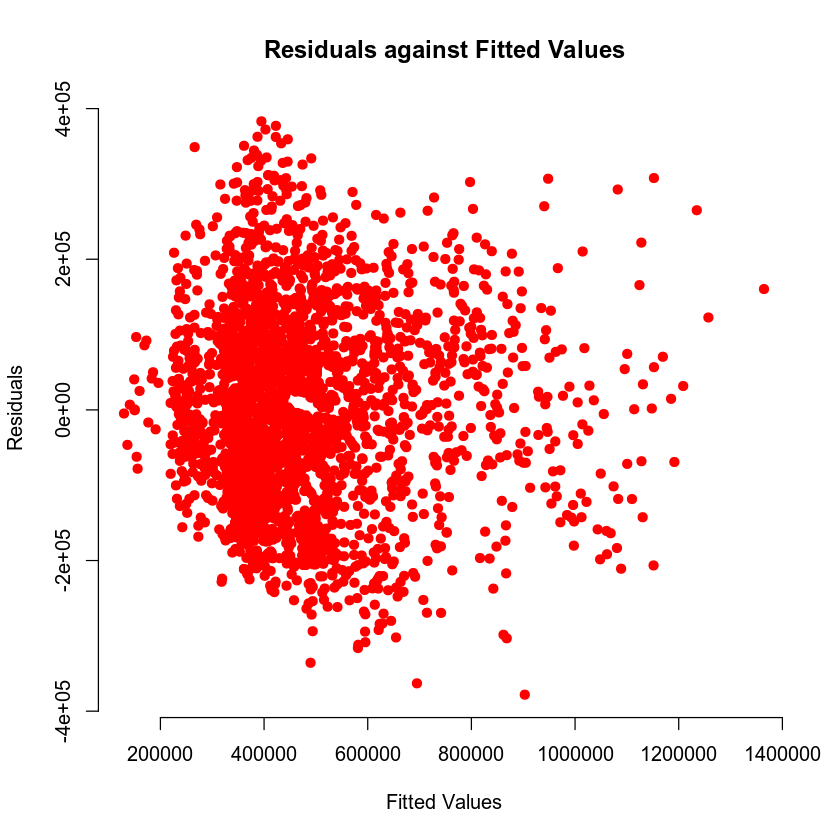


What we can see from the correlation coefficient is that there is a strong correlation between price and sqft\_living in a positive trend. There is a weaker correlation between age and price that is a negative trend.

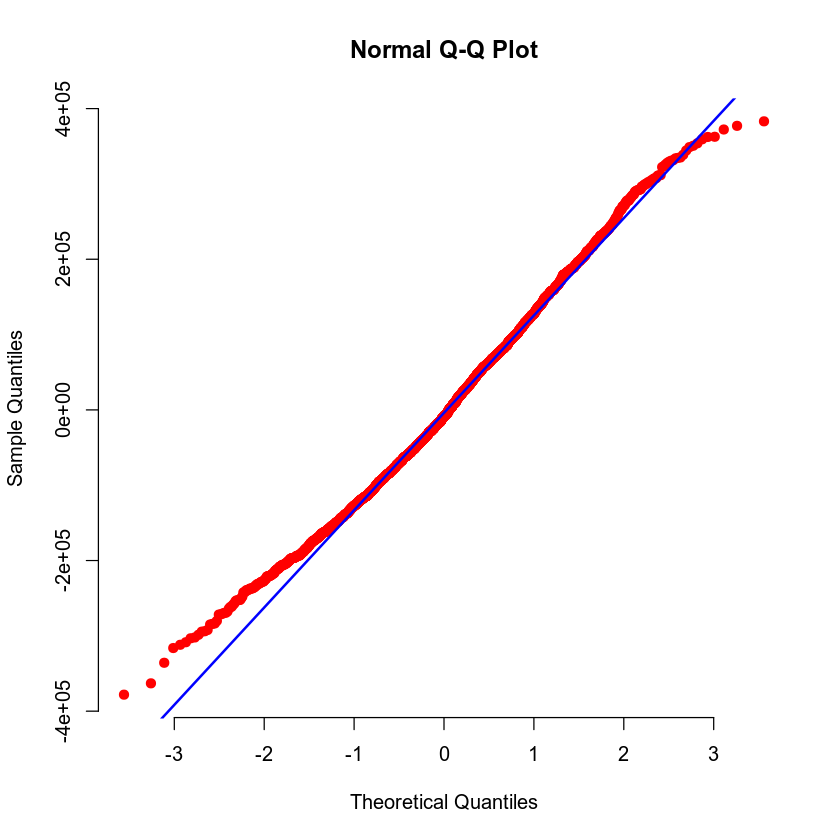
### Reporting Results

The general form of the first order regression model is: , where Y is price, X1 is livable square feet, X2 is grade, X3 is bathrooms, X4 is the view out to trees, and X5 is the view out to a lake. The first order regression model is: . The coefficient of multiple determination is 0.6475 or 64.75%. This states that 64.75% of the variance in price can be related to livable square feet, the grade of the house, the number of bathrooms, and the view of the house. The adjusted coefficient of multiple determination is 0.6468 or 64.68%. This is used to gauge how well the model generated fits the sample data used. The beta estimate for living area is 92.83, what this relates to is that for every unit of livable square footage it will increase the price by 92.83 units. The beta estimate for lake view is 228663.79, and this states that if the house has a lake view, it will increase the price by 228663.79 units.

Residuals against fitted values



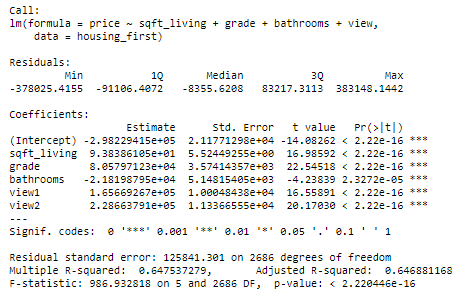
Normal Q-Q plot



From the residuals against fitted values we can see that there is no nonlinear pattern so we do not violate the constant variance assumption. We can then confirm that this is homoscedastic. The normal Q-Q plot assists with the normality assumption. The plotted points line up closely against the line provided and therefore does not violate the normality assumption.

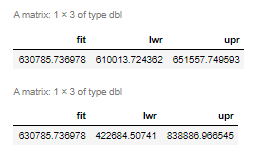
### Evaluating Significance of Model

The null hypothesis (H0) is that there is no relationship between price, livable square footage, grade, number of bathrooms, and the view. The alternative hypothesis (Ha) is that there is a relationship between price, livable square footage, grade, number of bathrooms, and the view such that it allows for a prediction of price. The p-value of the F-test is 2.22e-16 which is lower than the 5% level of significance needed to reject the null hypothesis. Therefore, we can reject the null hypothesis and accept the alternative hypothesis. All terms pass a 5% level of significance; sqft\_living, grade, view1, and view2 have a p-value of 2.22e-16 and bathrooms have a p-value of 2.32e-5. This means that all terms can be accepted and do not cause a deviation in the model.

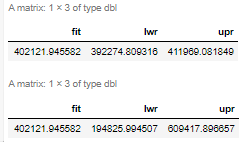
**

### Making Predictions Using Model

The predicted price of a house with a view that backs out to a lake, has a 2150 sq ft living area, a grade of 7, and three bathrooms would be $630,785.74. The 90% prediction interval has a fit of $630,785.74, a lower limit of $422,684.51 and an upper limit of $838,886.97. This states that there is a 90% level of confidence that a particular price of a house with the provided variables will be between $422,684.51 and $838,886.97. The 90% confidence interval has a fit of $630,785.74, a lower limit of $610,013.72, and an upper limit of $651,557.75. This states that there is a 90% level of confidence that the average price of a house with the provided variables will be between $610,013.72 and $651,557.75.



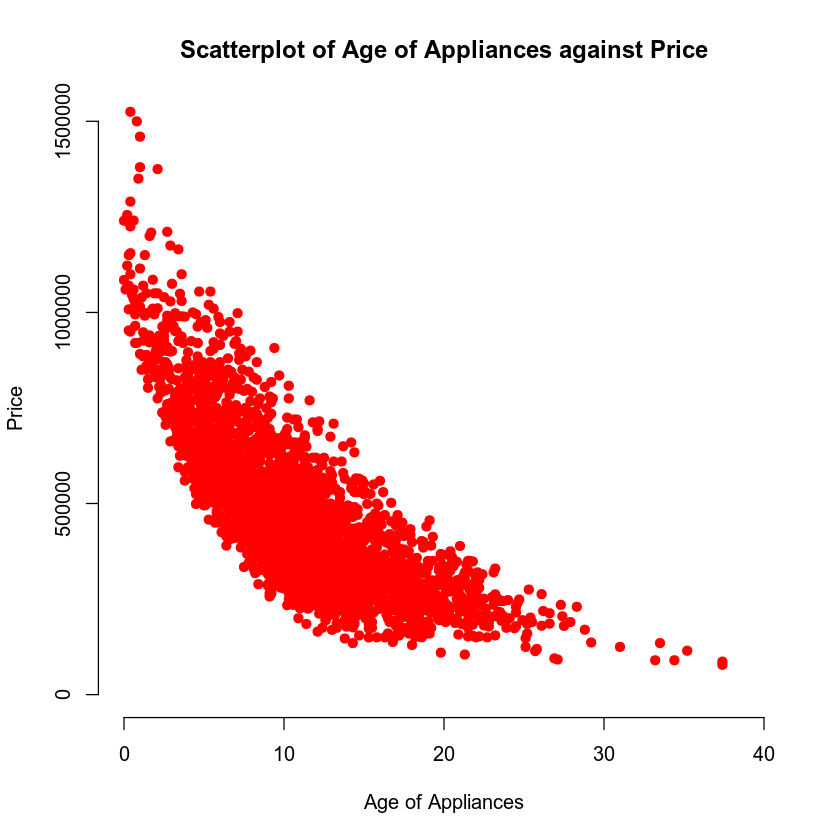
The predicted price of a house with a view that backs out to a road, has a 2150 sq ft living area, a grade of 7, and three bathrooms would be $402,121.95. The 90% prediction interval has a fit of $402,121.95, a lower limit of $194,825.99, and an upper limit of $609,417.90. This states that there is a 90% level of confidence that a particular price of a house with the provided variables will be between $194,825.99 and $609,417.90. The 90% confidence interval has a fit of $402,121.95, a lower limit of $392,274.81, and an upper limit of $411,969.08. This states that there is a 90% level of confidence that the average price of a house with the provided variables will be between $392,274.81 and $411,969.08. The prediction interval has a wider range because it takes regression error () into account from the population model.



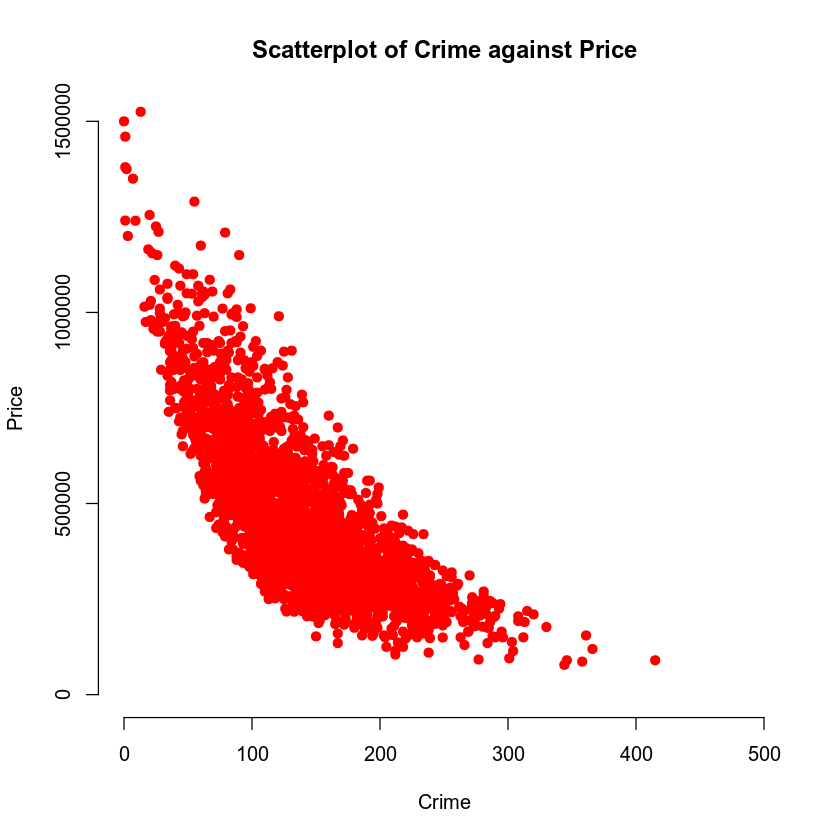
## 4. Model #2 - Complete Second Order Regression Model with Quantitative Variables

### Correlation Analysis

Price against age of appliances



Price against crime rate per 100,000 people

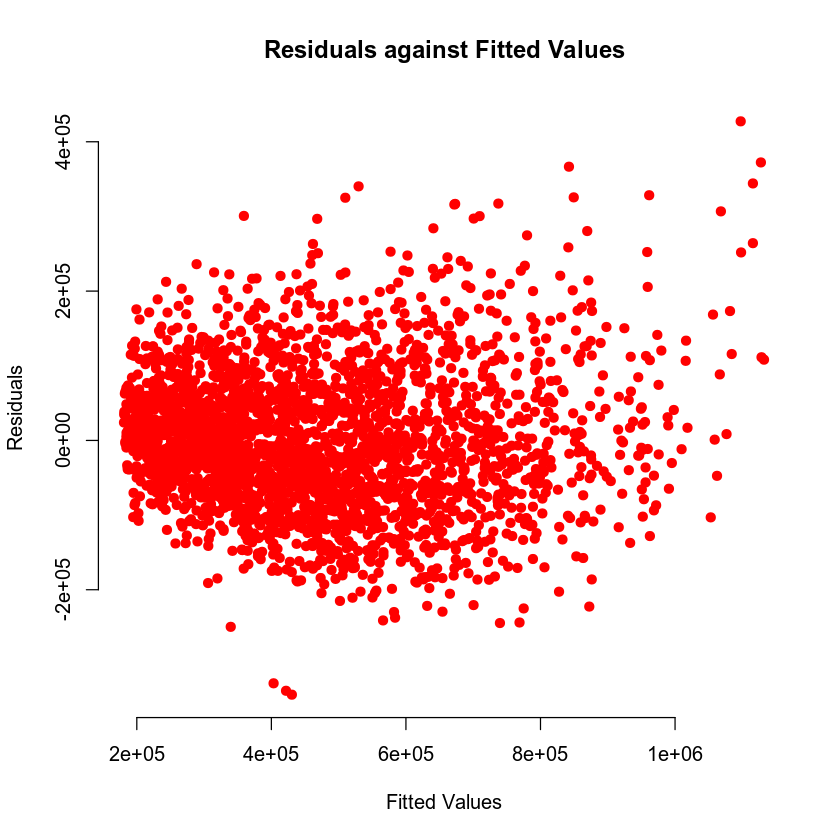
****

A second order model is appropriate because we can see how there is a curvature in the plotted points. A linear model would not fit knowing the curvature shown in both scatterplots so having quadratic variables would have a better fit.

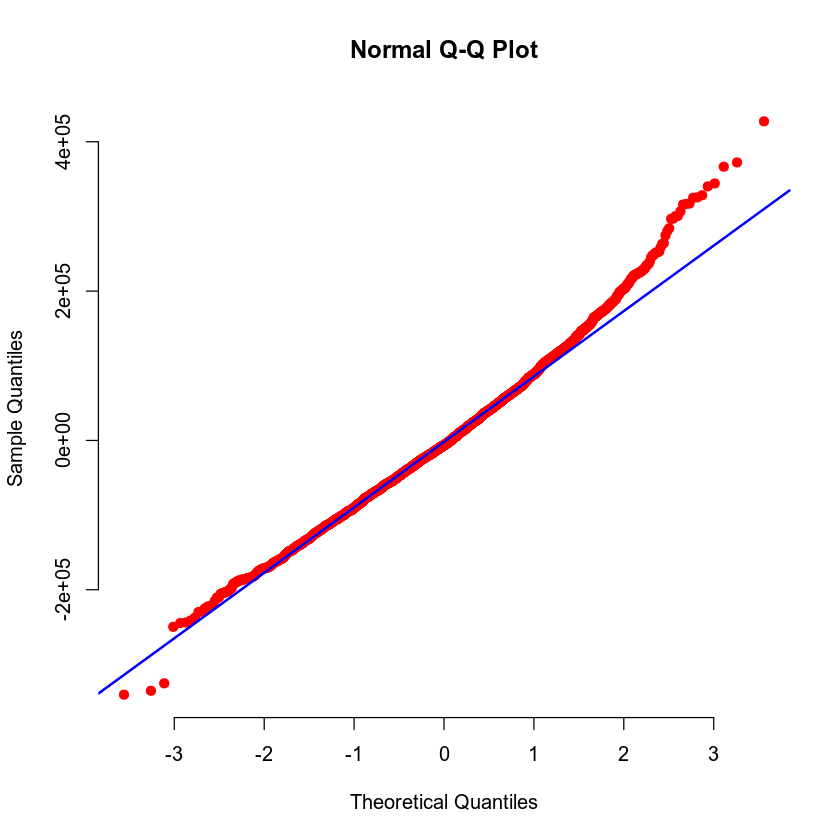
### Reporting Results

The general form of a complete second order model for price is: where Y is the price of the house, X1 is the age of appliances, X2 is the crime per 100,000 people, X1X2 is the interaction term between age of appliances and crime per 100,00 people. The complete second order regression model for price is: . The coefficient of multiple determination is 0.8088 or 80.88%. This describes that 80.88% of the variability of price can be explained by crime and the age of appliances. The adjusted coefficient of multiple determination is 0.8084 or 80.84%. This describes how well the model fits the provided data.

Residuals against fitted values



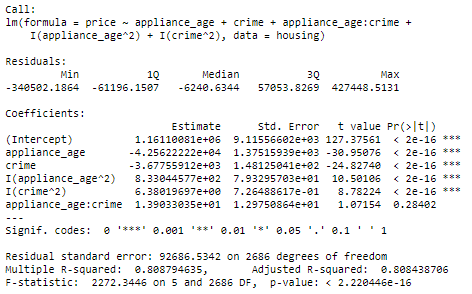
Normal Q-Q plot



From the residuals against fitted values we can see that there is no nonlinear pattern so we do not violate the constant variance assumption. We can then confirm that this is homoscedastic. The normal Q-Q plot assists with the normality assumption. The plotted points line up closely against the line provided and therefore does not violate the normality assumption.

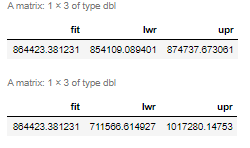
### Evaluating Significance of Model

The null hypothesis (H0) is that there is no relationship between price, crime per 100,00 people, the age of appliances, and the interaction between the two terms. The alternative hypothesis (Ha) is that there is a relationship between price, crime per 100,00 people, the age of appliances, and the interaction between the two terms such that a price could be predicted. The F-test p-value is 2.22e-16 which is well below the 5% level of significance. We can safely reject the null hypothesis and accept the alternative hypothesis. All terms pass the T-test as they have a p-value of 2e-16 with one exception, the interaction term age of appliance and crime per 100,00 people. The interaction term fails the test and therefore does not provide an accurate level of significance and it may skew results.

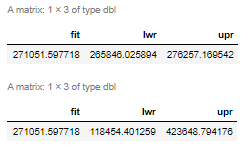


### Making Predictions Using Model

The predicted price of a house with one year old appliances and a crime rate of 81.02 per 100,00 individuals is $864,423.38. The 90% prediction interval has a fit of $864,423.38, a lower limit of $711,566.61, and an upper limit of $1,017,280.15. This provides that with a 90% level of confidence a particular house will be priced between $711,566.61 and $1,017,280.15. The 90% confidence interval has a fit of $864,423.38, a lower limit of $854,109.09, and an upper limit of $874,737.67. This describes with a 90% level of confidence the average house price will be between $854,109.09 and $874,737.67.

****

The predicted price of a house with fifteen year old appliances and a crime rate of 200.5 per 100,00 individuals is $271,051.60. The 90% prediction interval has a fit of $271,051.60, a lower limit of $118,454.40, and an upper limit of $423,648.79. This provides with a 90% level of confidence that the price of a particular house will be between $118,454.40 and $423,648.79. The 90% confidence interval has a fit of $271,051.60, a lower limit of $265,846.03, and an upper limit of $276,257.17. This states with a 90% level of confidence the average price of a house will be between $265,846.03 and $276,257.17.

****

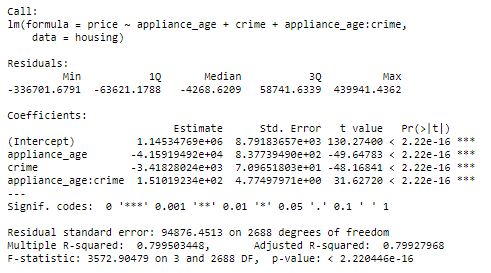
## 5. Nested Models F-Test

### Reporting Results

The general form of a first order model for price is: where Y is price, X1 is the age of appliances, X2 is crime rate per 100,00 people, and X1X2 is the interaction between the two terms. The first order regression model is: .

### Evaluating Significance of Model

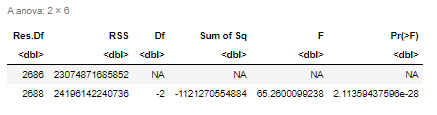
The null hypothesis (H0) is that there is no relationship between price, age of appliances, crime, and the interaction term between the two predictor variables. The alternative hypothesis (Ha) is that there is a relationship between price, age of appliances, crime, and the interaction term between the two predictor variables such that a price can be predicted. The F-test provides a p-value of 2.22e-16 which is lower than the level of significance threshold of 5%. This allows us to reject the null hypothesis and accept the alternative hypothesis. The T-test provides the p-value for all terms and shows that they all have a p-value of 2.22e-16 which is below the 5% level of significance. We can accept these terms as they are significant to the model.

**

### Model Comparison

A complete model is the base model with which you want to compare, this should have all terms to evaluate. A reduced model is a subset of the complete model to compare against. The reduced model in this comparison is: . The complete model in this comparison is: . You can see that the complete model contains quadratic terms whereas the reduced model does not have the quadratic terms.

The null hypothesis (H0) is that dropping the quadratic terms creates a more accurate model. The alternative hypothesis (Ha) is that keeping the quadratic terms is more accurate than removing them. The nested F-Test provides a p-value of 2.11e-28 which is below a 5% level of significance. This allows us to reject the null hypothesis and accept the alternative hypothesis that keeping the quadratic terms is more accurate than dropping them.



## 6. Conclusion

I would use the complete second order model to predict housing prices . It contains a more accurate model statistically such as the coefficient of multiple determination which has a value of 80.88%. My only concern with the model is that the interaction term between appliance age and crime does not pass a 5% level of significance test so i would like to evaluate dropping it using a nested F-test analysis.

The practical importance of the analyses performed helps compare a first order model with both qualitative and quantitative variables, a complete second order model, and a subset of the second order model utilizing a nested F-test. By comparing the three models, we deliberate which model best fits the data we have and which model can provide the best possible prediction. Using statistical values such as F-test p-values, ANOVA, residual and fitted values, normal QQ plots, and nested F-test, we can identify the best model.