# MAT 303 Project One Summary Report

Christopher Williams

Christopher.Williams39@SNHU.edu

Southern New Hampshire University

## 1. Introduction

The data set we will be exploring contains many variables related to the size, value, and composition of houses. I will be analyzing the relationships between different attributes of the houses and their selling prices. This will be used to predict how much the houses will be sold for. My analyses will include Linear Regression Models, scatterplots, F-Tests, Complete Second-Order Models, and Q-Q plots. These tests will determine things like correlation and significance.

## 2. Data Preparation

The important variables in this data set are the price (price), living area (sqft\_living), age of the home(age), the age of appliances (appliance\_age), the crime rate per 100,000 people (crime), grade of construction quality (grade), bathrooms, and view. The following code shows that there are 8 columns and 2692 rows. This equates to 8 variables that will be used and 2692 data points.

"# of columns"

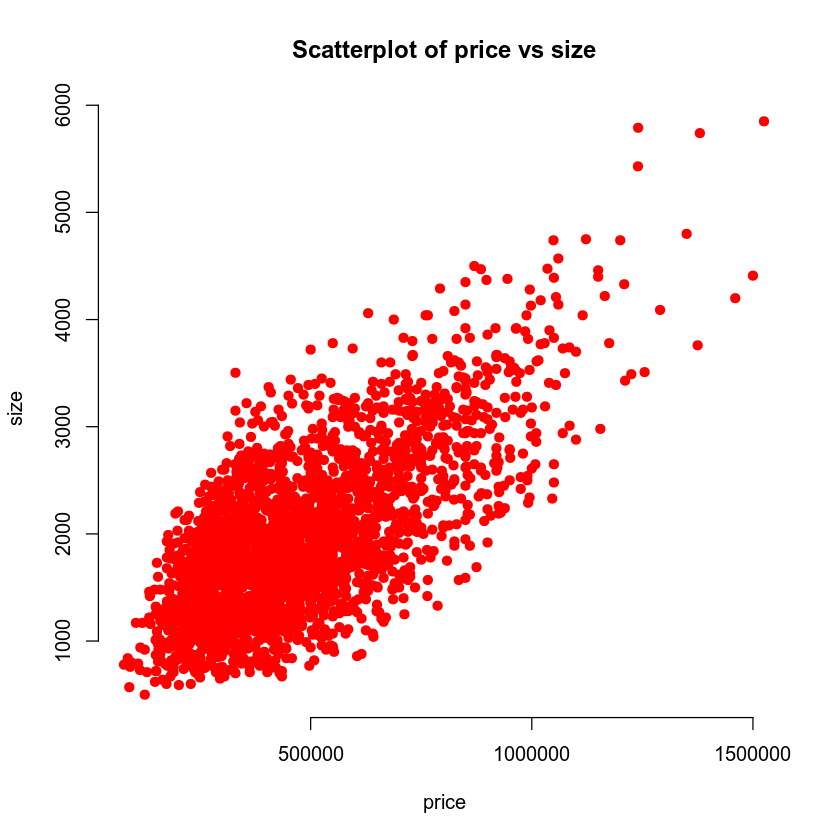
8

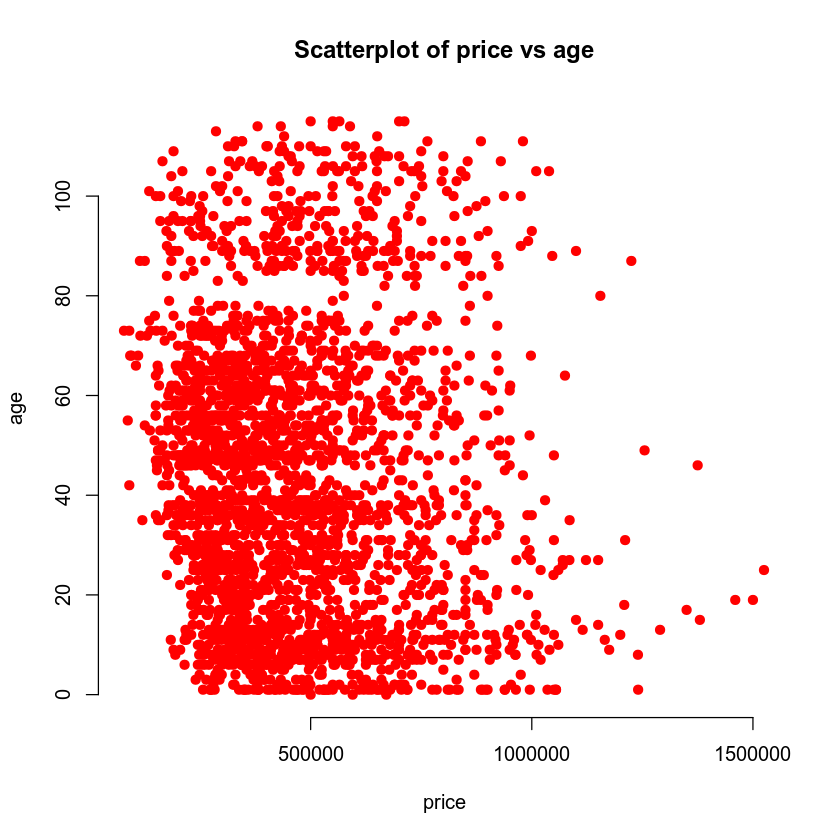
"# of rows"

2692

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

### Correlation Analysis



**

|  |  |  |  |
| --- | --- | --- | --- |
| price | 1.0000 | 0.6895 | -0.0746 |
| sqft\_living | 0.6895 | 1.0000 | -0.3547 |
| age | -0.0746 | -0.3547 | 1.0000 |

The scatterplots above show a strong positive, linear relationship between size and price. The relationship between age and price is much less clear, though it seems the most expensive houses are mostly new. The correlation coefficients show a strong positive relationship between price and size with a very weak negative relationship between price and age. This leads me to the conclusion that age is not likely to play a significant role in price, but size plays a very significant role. The following work will either confirm or disprove this theory.

### Reporting Results

*"Create a multiple regression model of price using size, grade, bathrooms, and view as predictors"*

Call:

lm(formula = price ~ sqft\_living + grade + bathrooms + view,

data = housing)

Residuals:

Min 1Q Median 3Q Max

-378025 -91106 -8356 83217 383148

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.982e+05 2.118e+04 -14.083 < 2e-16 \*\*\*

sqft\_living 9.384e+01 5.524e+00 16.986 < 2e-16 \*\*\*

grade 8.058e+04 3.574e+03 22.545 < 2e-16 \*\*\*

bathrooms -2.182e+04 5.148e+03 -4.238 2.33e-05 \*\*\*

view1 1.657e+05 1.000e+04 16.559 < 2e-16 \*\*\*

view2 2.287e+05 1.134e+04 20.170 < 2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

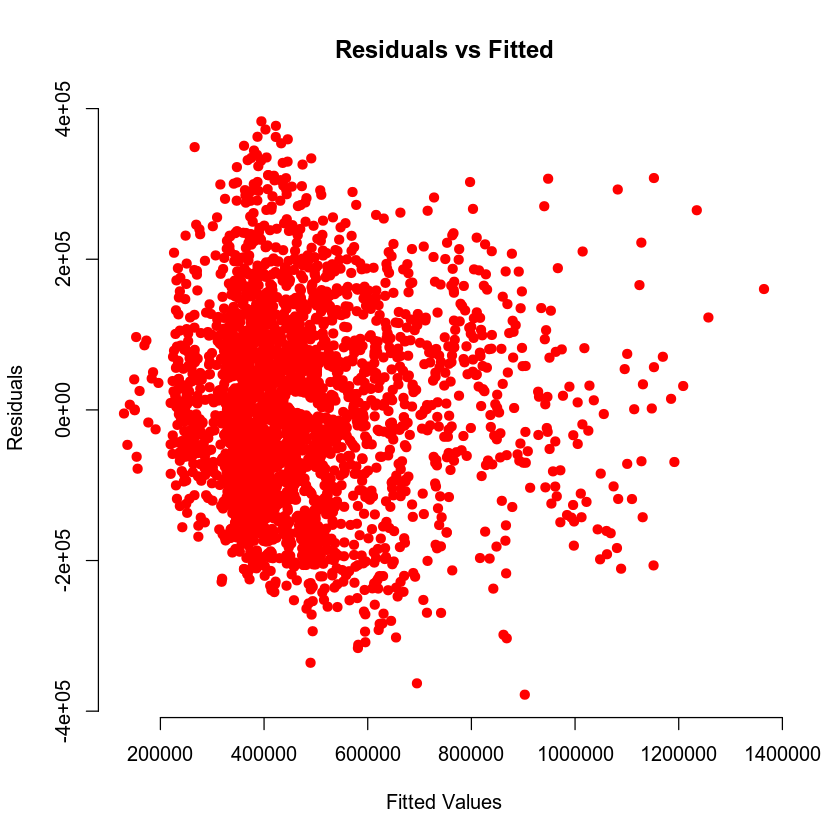
Residual standard error: 125800 on 2686 degrees of freedom

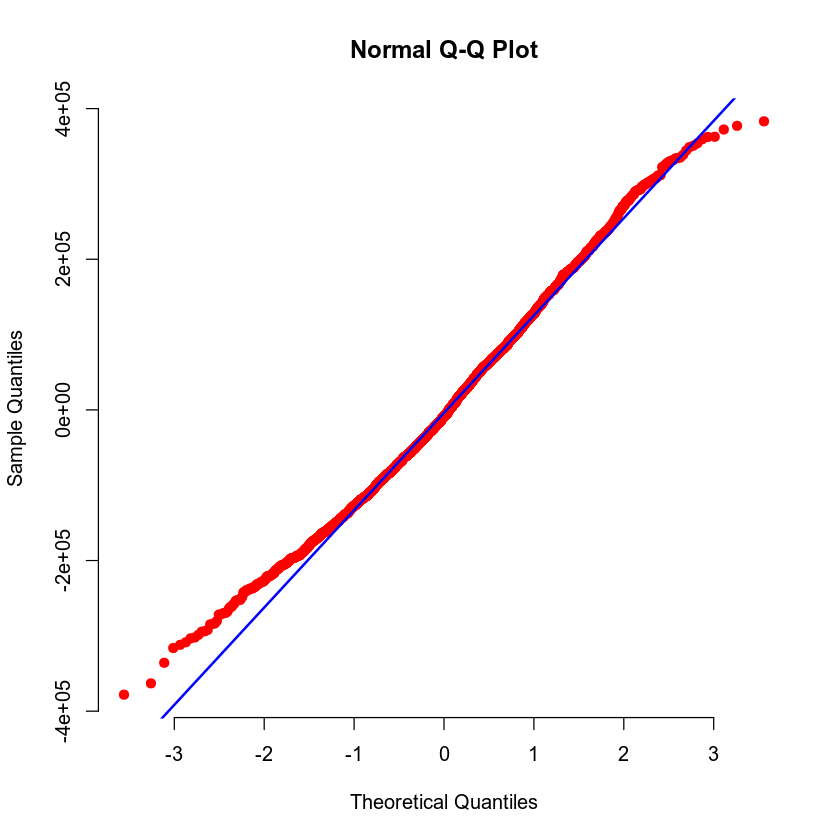
Multiple R-squared: 0.6475, Adjusted R-squared: 0.6469

F-statistic: 986.9 on 5 and 2686 DF, p-value: < 2.2e-16

[1] "Residuals vs fitted values for model1"

[1] "Q-Q plot for model1"





The general form of the multiple regression model with price(Y) as the response variable and living area(X1), grade of the home(X2), number of bathrooms(X3), and view(X4/X5) as predictor variables is E(Y) = B0 + B1X1 + B2X2 + B3X3 + B4X4. Using the linear regression model, we can express this as the equation E(Y) = -2.982e+05 + 0.9384X1 + 0.0008X2 - 0.000218X3 + 1.657e+05X4 + 2.287e+05X5. The value of R^2 is 0.6475, which shows that the data points, on average, are not very far from the fitted line. This indicates a somewhat high level of goddness-of-fit. The beta estimates for size indicate a much stronger influence than any other variable. This shows that one unit of size produces a much greater effect on price than one unit of grade or bathrooms. The beta estimate for view is a very small positive number, meaning that the view likely has a small amount of influence over the price.

One may notice that the "residuals vs fitted values" graph shows that the residuals take a complex shape, with more expensive houses deviating further upwards from the fitted values. The Q-Q plot shows a non-normal distribution. This is called a "heavy tails", meaning the data exhibits more extreme values than expected when measuring data points that are far from the mean.

### Evaluating Significance of Model

The null hypothesis for a regression model is that the independent values do not have an effect on the dependent variable. The alternative hypothesis would then be that the variables do have an effect. A 5% significance would require a p-value of less than 0.05 to show that the model is significant. The p-value is < 2.2e-16, so it is at a 5% significance. Thus, the null hypothesis is rejected. The other terms have p-values of Every term has a p-value of less than 0.05, so they are all significant. This means that they all have a less than five percent chance of having no actual effect on the price.

"90% prediction interval "

A matrix: 1 × 3 of type dbl fit lwr upr 630785.7 422684.5 838887

"90% Confidence Interval"

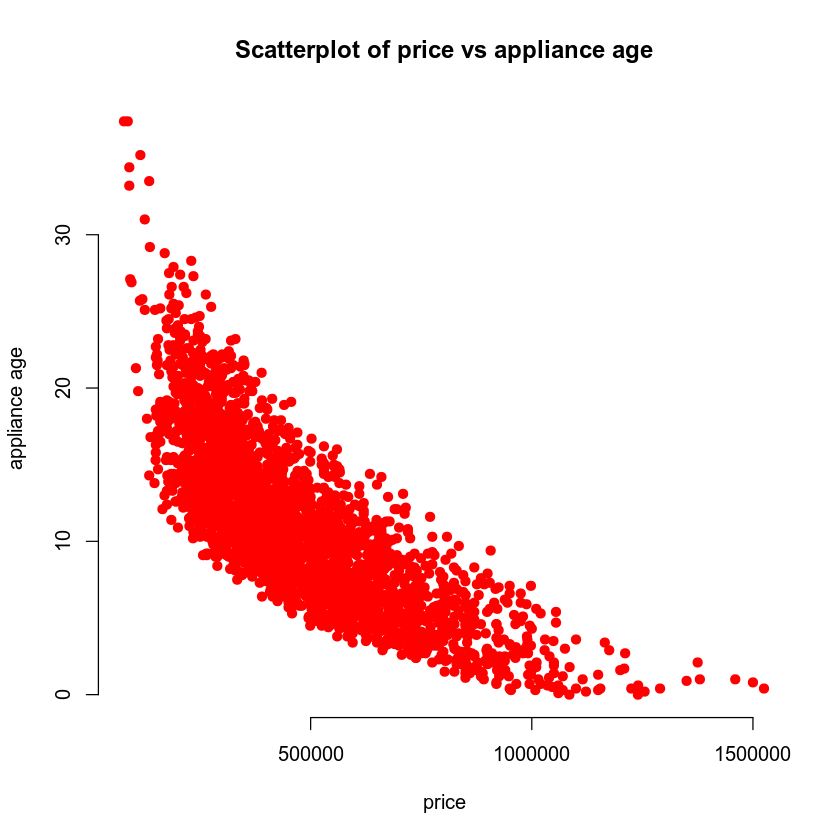
A matrix: 1 × 3 of type dbl fit lwr upr 630785.7 610013.7 651557.7

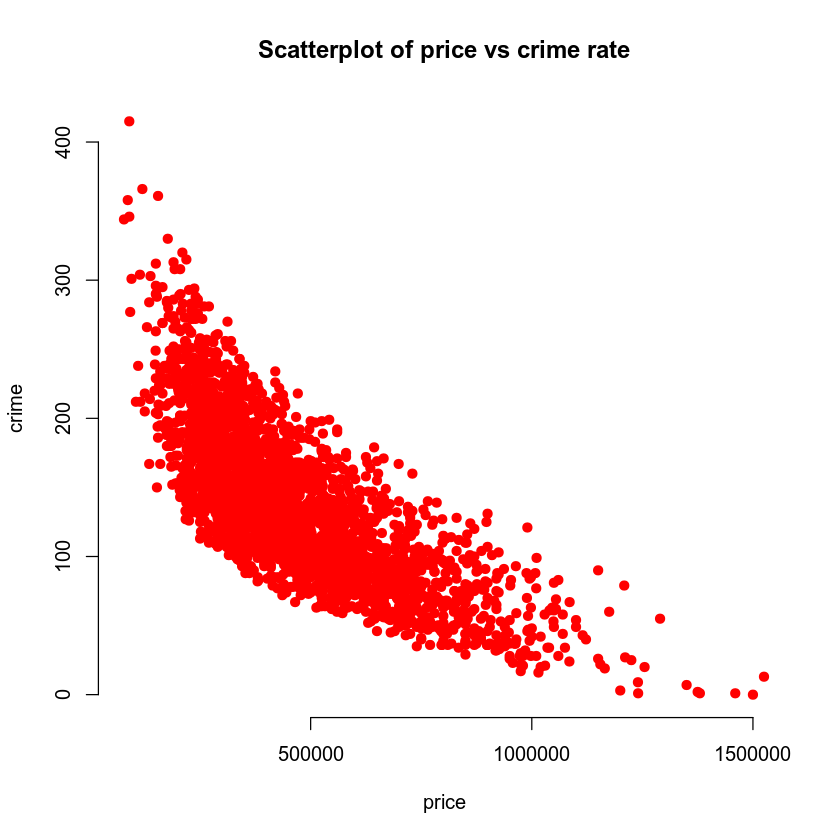
### Making Predictions Using Model

The predicted price for a home that backs out to a lake and has a 2,150 sq ft living area, 7 grade, and three bathrooms is 630,785.7 dollars. The 90% prediction and confidence intervals for the price of this home would be, respectively, 422,685 - 838,887 and 610,014 - 651,558. The prediction interval represents the range in which a future data point has a 90% chance of falling. The confidence interval has a wider interval because it represents the range which we can be 90% certain contains the true mean of all of the data. Dealing with the mean of the data tends to cancel out the effects of outliers, so we can create a more narrow window in which we predict the mean will fall.

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

### Correlation Analysis





The above scatterplots display price (price) vs. the age of appliances (appliance\_age) and price (price) vs. the crime rate per 100,000 people (crime). They look nearly identical. They both show a very strong linear negative relationship. The is no need to use a second order model here, because the data seems to be linear.

### Reporting Results

The general form of a complete second order model for price using age of appliances and crime rate per 100,000 people as predictors would be E(Y) = B0 + B1X1 + B2X2 + B3X1X2 + B4X1^2 + B5X2^2 wher crime is X1 and appliance age is X2. Using a linear regression model (shown below), the equation for this model can be shown to be E(Y) = 1.161e+06 - 3.678e+03X1 - 4.256e+04X2 + 13.9X1X2 + 6.380X1^2 + 833X2^2

Residuals:

Min 1Q Median 3Q Max

-340502 -61196 -6241 57054 427449

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.161e+06 9.116e+03 127.376 <2e-16 \*\*\*

crime -3.678e+03 1.481e+02 -24.827 <2e-16 \*\*\*

appliance\_age -4.256e+04 1.375e+03 -30.951 <2e-16 \*\*\*

I(crime^2) 6.380e+00 7.265e-01 8.782 <2e-16 \*\*\*

I(appliance\_age^2) 8.330e+02 7.933e+01 10.501 <2e-16 \*\*\*

crime:appliance\_age 1.390e+01 1.298e+01 1.072 0.284

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

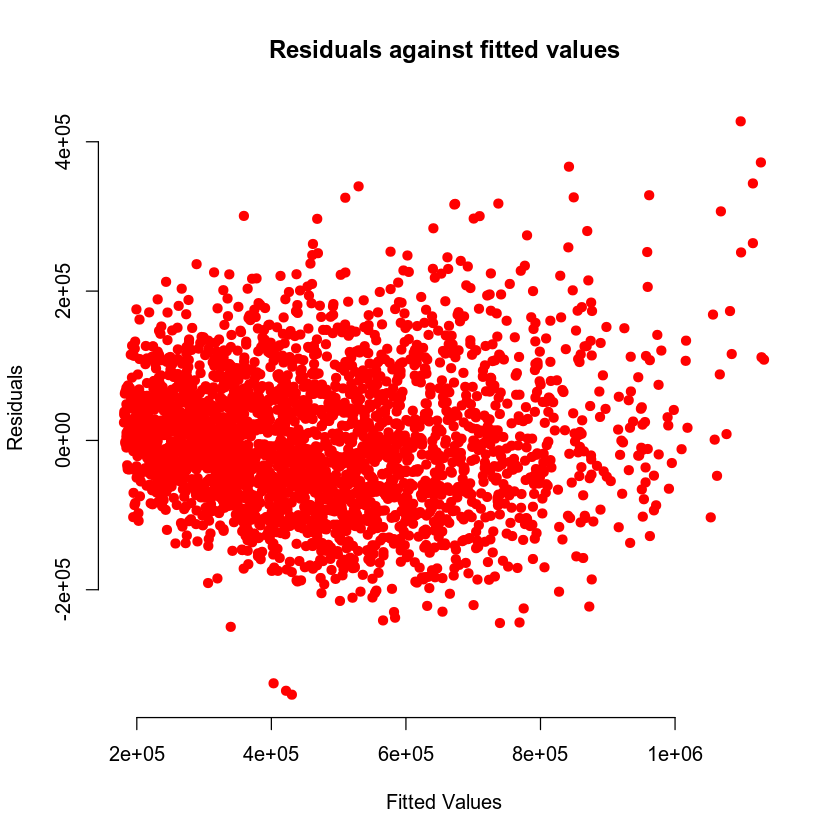
Residual standard error: 92690 on 2686 degrees of freedom

Multiple R-squared: 0.8088, Adjusted R-squared: 0.8084

F-statistic: 2272 on 5 and 2686 DF, p-value: < 2.2e-16

The values of R^2 and R^2 Adjusted for this model are, respectively, 0.8088 and 0.8084. These are quite high, indicating that the model fits the data well. The values are also very close together, indicating that the terms used are appropriate. In other words, the data is not “overfit”.

The residuals and fitted values of this model are displayed in graphs below. The fitted values are the values predicted by the model. The residuals are a measure of how far from those values the actual data points fall. The graphs indicate that the residuals have an upward trend as prices rise. This is a sign that the model is not accounting for some variable that likely has an exponential function.



### 

### Evaluating Significance of Model

The p-value of the overall model is < 2.2e-16, as is the p-value of every individual term other than the coefficient between crime and appliance age. This p-value indicates the likelihood that the null hypothesis is not true. In this case, the null hypothesis for the model is that it does nothing to predict the data points. The null hypothesis for the terms is that they do nothing to predict the price. The alternative hypothesis for the model is then that it does predict the data. The alternative for each term is that it does affect the price. Since < 2.2e-16 is much less than 0.05, these terms and the model have a 5% significance and their null hypotheses are rejected.

The p-value of the coefficient term is greater than 0.05, so its null hypothesis is not rejected and it does not have a 5% significance level.

### Making Predictions Using Model

*"90% prediction interval”*

A matrix: 1 × 3 of type dbl

| fit | lwr | upr |
| --- | --- | --- |
| 864423.4 | 711566.6 | 1017280 |

*"90% confidence interval"*

A matrix: 1 × 3 of type dbl

| fit | lwr | upr |
| --- | --- | --- |
| 864423.4 | 854109.1 | 874737.7 |

The predicted price for a home that has one-year-old appliances and is in an area that has a crime rate of 81.02 per 100,000 individuals is $864,423.4. The 90% prediction interval for the price of this home is $711,567 - $1,017,280. This represents the range in which we can be 90% sure the next data point will fall. This is a wide interval because there are many outliers in in data. The confidence interval is much smaller because it represents the range in which we may be 90% certain the true mean of the population lies. This interval is $854,109 - $874,738.

*[1] "90% prediction interval"*

A matrix: 1 × 3 of type dbl

| fit | lwr | upr |
| --- | --- | --- |
| 271051.6 | 118454.4 | 423648.8 |

[1] "90% confidence interval"

A matrix: 1 × 3 of type dbl

| fit | lwr | upr |
| --- | --- | --- |
| 271051.6 | 265846 | 276257.2 |

The predicted price for a home that has 15-year-old appliances and is in an area that has a crime rate of 200.50 per 100,000 individualsis $271,052. The prediction interval is $118,454 - $423,649. The confidence interval is $265,846 - $276,257. This interval is narrower because it accounts for the mean, which is less affected by the presence of outliers.

## 5. Nested Models F-Test

### Reporting Results

*lm(formula = price ~ crime + appliance\_age + crime:appliance\_age,*

data = housing)

Residuals:

Min 1Q Median 3Q Max

-336702 -63621 -4269 58742 439941

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.145e+06 8.792e+03 130.27 <2e-16 \*\*\*

crime -3.418e+03 7.097e+01 -48.17 <2e-16 \*\*\*

appliance\_age -4.159e+04 8.377e+02 -49.65 <2e-16 \*\*\*

crime:appliance\_age 1.510e+02 4.775e+00 31.63 <2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 94880 on 2688 degrees of freedom

Multiple R-squared: 0.7995, Adjusted R-squared: 0.7993

F-statistic: 3573 on 3 and 2688 DF, p-value: < 2.2e-16

The general form of a first order model for price using age of appliances and crime rate per 100,000 people as predictors can be represented as E(Y) = B0 + B1X1 +B2X2 +B3X1X2, where X1 is crime and X2 is appliance age. As shown in the model above, this can be re-written as the equation E(Y) = 1,145,000 – 3,418X1 – 41,590X2 + 151X1X2.

### Evaluating Significance of Model

The p-value of the model is < 2.2e-16. This shows the chances of the alternative hypothesis being true. In this case the null hypothesis is that the model does not predict the data. That makes the alternative hypothesis that the model is accurate. Since < 2.2e-16 is far below 5%, the null hypothesis is rejected. For the individual terms, their null hypotheses are that they do not affect the price of houses. Their alternatives are that they do affect the prices. Their p-values are all < 2.2e-16, so we can safely reject their null hypotheses at a 5% significance.

### Model Comparison

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| --- | --- | --- | --- | --- | --- |
| <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 2686 | 2.307487e+13 | NA | NA | NA | NA |
| 2688 | 2.419614e+13 | -2 | -1.121271e+12 | 65.26001 | 2.113594e-28 |

The p-value here represents the likelihood that the reduced models null hypothesis is true, which is that the models variables have no real effect on the price. The p-value is many orders of magnitude lower than 0.05, so the null hypothesis can be safely rejected at a 5% significance level. This leads to the alternative hypothesis, which is that the variables of the reduced model so have an effect on price. The interaction terms should then be used to make a more powerful model.

A reduced model is a complete model with some terms taken out. It is useful for when terms are suspected of being useless or overfitted. For reference, the nested model can be expressed E(Y) = B0 + B1X1 +B2X2 +B3X1X2, where X1 is crime and X2 is appliance age. The complete model is expressed E(Y) = B0 + B1X1 + B2X2 + B3X1X2 + B4X1^2 + B5X2^2 wher crime is X1 and appliance age is X2.

## 6. Conclusion

The second model, model2, was found to be the most accurate model for this data. The R^2 and R^2 adjusted values were very high and close together, indicating that the terms chosen for the model were all very appropriate and produced a good fit. This is the model I would use to predict house prices up to $1,000,000. Past that value, the model rapidly loses accuracy by underestimating prices, as is evident in its Q-Q plot. The first model, model1, would perform a little better past $1,000,000, perhaps because of the size and grade terms. If any predictions over this mark are needed, more tests will be required. In conclusion, model2 can be used with great accuracy for predicting house prices up to $1,000,000.