# MAT 303 Project One Summary Report

Joshua Wozny

joshua.wozny@snhu.edu

Southern New Hampshire University

## 1. Introduction

Our employer, a real estate firm, has tasked us with studying selling prices of homes, and model how various home features can be used to predict it. We will be using a provided set of historical data that lists sales prices along with several factors relevant to features of the home. We will be using first and second order regression analysis to model these factors. The models will be tested using F-test statistics, residual and variance analysis, including graphs to visualize. These models will be used by the firm to predict how to price their listings competitively.

## 2. Data Preparation

The provided housing dataset includes our response variable for the sales price (price), and the following possible quantitative predictor variables: number of bedrooms (bedrooms), number of bathrooms (bathrooms), size of living space (sqft\_living), size of upper level (sqft\_above), size of the lot (sqft\_lot), age of home (age), craftsmanship of home (grade), avg age of appliances (appliance\_age), and crime rate (crime). It also includes the following qualitative variables: backyard that indicates if the home has a backyard (1) or not (0) and view which indicates the type of view (lake=2, trees=1, road=0). There are 2,692 rows of data in our set. (See Table 1)

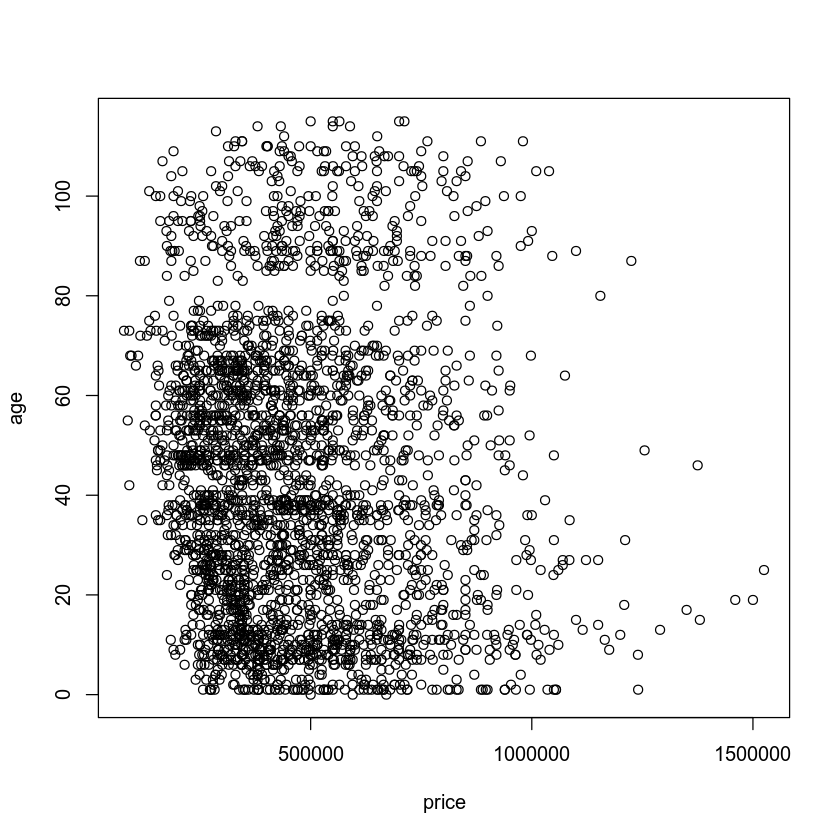
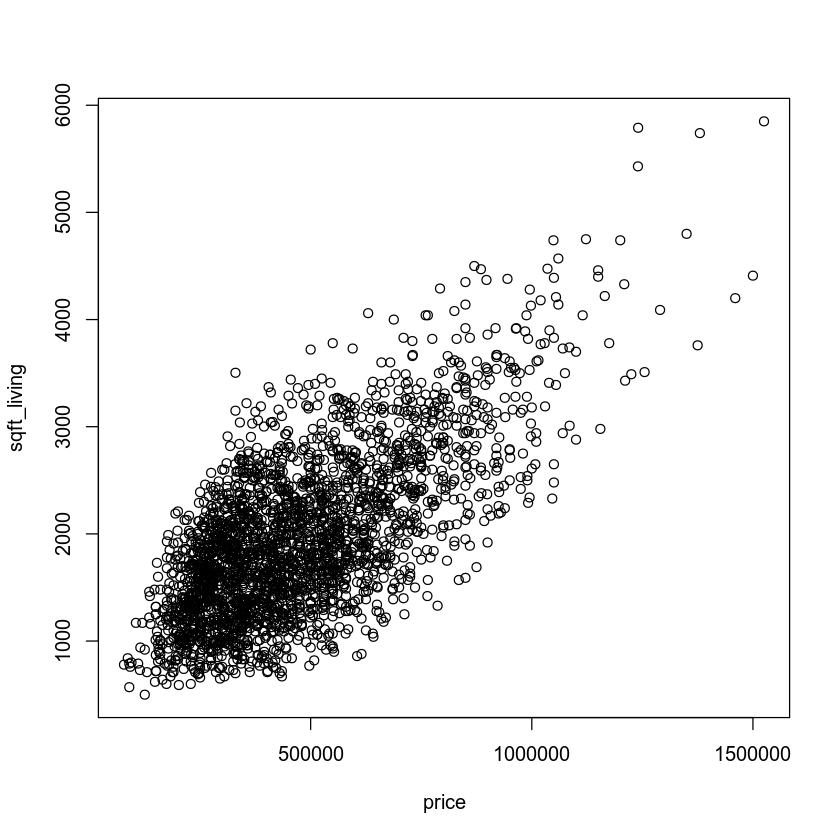
**Table 1**

| **Variable** | **What does it represent?** |
| --- | --- |
| price | Sale price of the home |
| bedrooms | Number of bedrooms |
| bathrooms | Number of bathrooms |
| sqft\_living | Size of the living area in sqft |
| sqft\_above | Size of the upper level in sqft |
| sqft\_lot | Size of the lot in sqft |
| age | Age of the home |
| grade | Measure of craftsmanship and the quality of materials used to build the home |
| appliance\_age | Average age of all appliances in the home |
| crime | Crime rate per 100,000 people |
| backyard | Home has a backyard (backyard=1) or not (backyard=0) |
| view | Home backs out to a lake (view=2), backs out to trees (view=1), or backs out to a road (view=0) |

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

### Correlation Analysis

To determine variables to include in our model we first visually inspect the plots of possible predictors against the response variable to determine if there is a significant relationship.



From the plot of price vs. living space (highlighted in blue), we can see a significant positive correlation between price and square footage of the living space. From the plot of price vs. age (highlighted in red), we see no significant positive correlation between price and age of the building.

We next create a correlation matrix to confirm our visual inspection, and to quantify the amount of correlation that exists. As expected the correlation coefficient of living space is significant and positive with a coefficient of 0.689. Also, as expected, the correlation coefficient of age is very low, showing almost no correlation with a value of -0.075. (See Table 2)

**Table 2**

| **correlation** | **price** | **sqft\_living** | **age** |
| --- | --- | --- | --- |
| **price** | 1 | 0.6894838 | -0.07460764 |
| **sqft\_living** | 0.68948378 | 1 | -0.35473332 |
| **age** | -0.07460764 | -0.3547333 | 1 |

### Reporting Results

The general form of our regression model (model 1) is:

**E(y)=β0 + β1 x1 + β2 x2 + β3 x3 + β4 x4 + β5 x5 + β6 x1 x2 + β7 x1 x3 + β8 x2 x3**

Where y is home price, x1 is living area in sqft., x2 is the grade of the home, x3 is the number of bathrooms, x4 is the view=1(trees), and x5 is the view=2(lake). This includes interactive terms for each quantitative variable. Performing our regression analysis we determine our co-efficients for our model of home prices:

**Y =-254512.9 + 119.8x1 + 77442.6x2 - 80719.3x3 + 163983.8x4 + 226231.9x5 - 5.0x1x2 + 5.7 x1x3 + 6391.6 x2x3**

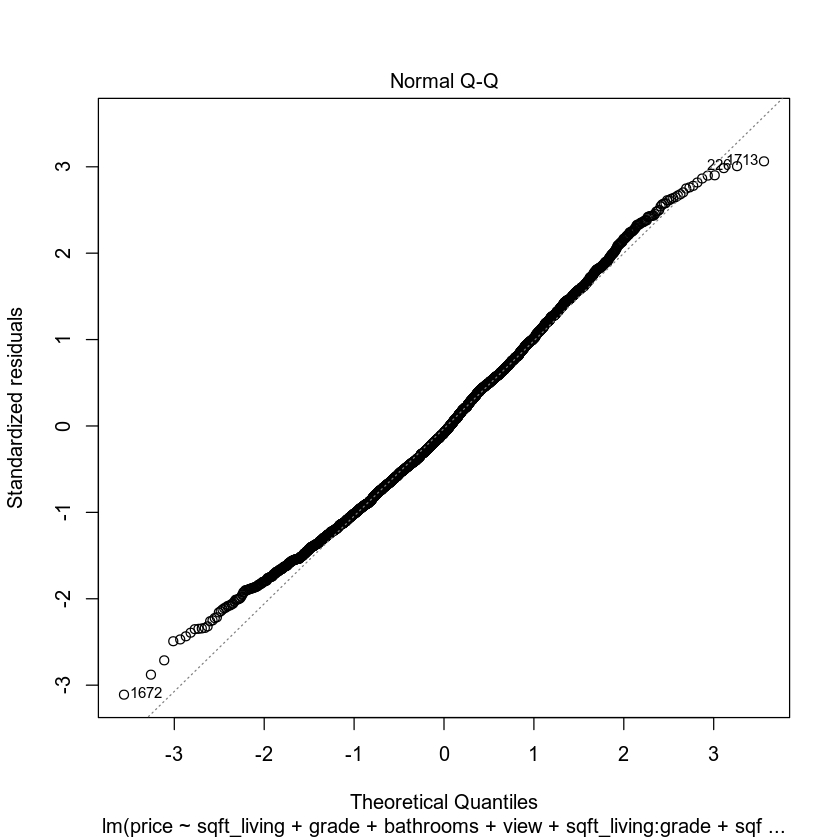
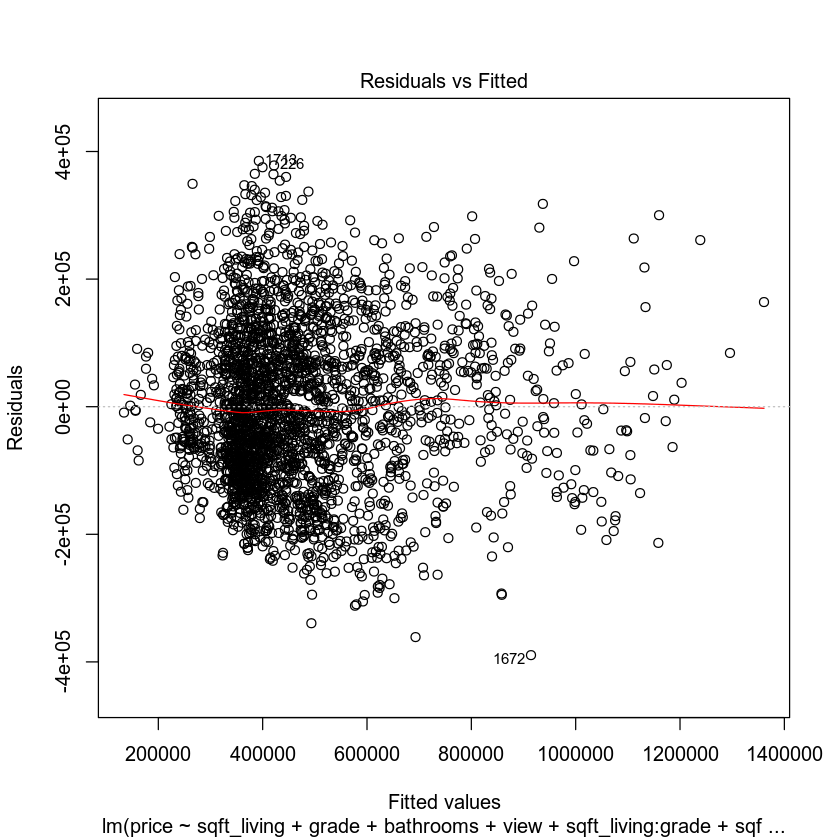
With an R-squared value of 0.6481, the model predicts 64.81% of the variance within our response variable, making the model a fair, although not stellar candidate for our use case. The adjusted R-squared value of 0.647 would be used to compare other models of our response variable to help determine which model was superior.

The coefficient for living area, **β1 ,** of 119.8 indicates that for each additional square foot of living space, all other terms remainingthe same, the home value increases by $119.80 mediated by the interaction terms **β6** and **β7** which add an additional $5.00x grade plu $5.70x number of bathromoms, all other terms remaining the same. The coefficient for a lake view (view=2), **β5 ,** of 226231.9 indicates that a lake view increases the value of a home by $226,231.90, all other terms remaining the same.

Our model 1 has the following fitted values and residuals:

| **fitted** | **residuals** |  | **fitted** | **residuals** |  | **fitted** | **residuals** |  | **fitted** | **residuals** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 586148.8 | -261148.8 |  | 472026.7 | 67973.3 |  | 557260.1 | 137739.9 |  | 360918.9 | 36581.1 |
| 735037.0 | 9963.0 |  | 826289.5 | -71289.5 |  | 347604.9 | -97604.9 |  | 368434.6 | 160665.4 |
| 450975.9 | 120024.1 |  | 958881.2 | -40881.2 |  | 344077.6 | 60922.4 |  | 349423.1 | 66076.9 |
| 404217.3 | -9217.3 |  | 349416.1 | 61583.9 |  | 538420.2 | 134579.8 |  | 624109.3 | -284109.3 |
| 363986.6 | 85963.4 |  | 339750.8 | -64750.8 |  | 569619.9 | 25380.1 |  | 809586.0 | 182114.0 |
| 362090.5 | -119090.5 |  | 362497.9 | 92502.1 |  | 461867.9 | -224867.9 |  | 297445.3 | 247554.7 |
| 251848.1 | -31898.1 |  | 558200.7 | -219200.7 |  | 384011.4 | -159011.4 |  | 381051.5 | 94948.5 |
| 748588.6 | 50411.4 |  | 479422.4 | -34422.4 |  | 516628.1 | 8371.9 |  | 659741.5 | -92241.5 |
| 378238.2 | -171238.2 |  | 235611.0 | 50339.0 |  | 410588.4 | -55588.4 |  | 359084.5 | -109184.5 |
| 453858.4 | 296141.6 |  | 665152.0 | 124848.0 |  | 574983.5 | -84983.5 |  | 391708.9 | -13208.9 |
| 445518.9 | -95518.9 |  | 378396.6 | -68396.6 |  | 340265.8 | -71265.8 |  | 589535.4 | 90464.6 |
| 348410.5 | -76410.5 |  | 432156.0 | 97794.0 |  | 496119.0 | -133119.0 |  | 583486.4 | 46513.6 |
| 514931.7 | -129981.7 |  | 396118.8 | -57118.8 |  | 263319.6 | 116680.4 |  | 351389.6 | 208610.4 |
| 370476.9 | -98476.9 |  | 696264.5 | 93735.5 |  | 805100.8 | 121399.2 |  | 576005.8 | 73944.2 |
| 403590.1 | -125590.1 |  | 359696.2 | 235303.8 |  | 537692.9 | 12307.1 |  | 542553.5 | 132446.5 |

And plotting residuals yields the following plots:



The plot of residuals vs. fitted values (in blue) shows a slight amount of funneling, but overall shows a good spread of values mostly equal distance from zero, meeting the assumption of homoscedasticity sufficiently for our use case. The Q-Q plot (in pink) shows residuals following the “normal” slope line very closely, indicating that our model meets the assumption of normality. Overall the model appears to be valid.

### Evaluating Significance of Model

We perform an F-test to identify model significance. The null hypothesis is that no linear relationship exists between any of our predictor variables and the response variable, and our alternate hypothesis is that at least one of our predictor variables has a linear relationship with our response variable such that:

Hnull: β1 = β2 = β3 = β4 = β5 = β6 = β7 = β8 = 0

Halt: β1 or β2 or β3 or β4 or β5 or β6 or β7 or β8 ≠ 0

With a test statistic of 617.6 and a p-value of < 2.2e-16, approaching zero, we can reject the null hypothesis and conclude that at least one of the predictor variables is significant with 95% confidence.

To evaluate each individual variable within our model we need to perform a t-test for each variable, with i in range (1:8).

Hnull: βi = 0, predictor having no significance

Halt: βi ≠ , predictor having significance

Results of each are summarized below:

|  | t-value | p-value |
| --- | --- | --- |
| sqft\_living | 3.5170 | 0.0004 |
| grade | 6.5750 | 0.0000 |
| bathrooms | -1.9540 | 0.0508 |
| view1 | 16.3070 | < 2e-16 |
| view2 | 19.4160 | < 2e-16 |
| sqft\_living:grade | -1.1510 | 0.2500 |
| sqft\_living:bathrooms | 0.8520 | 0.3940 |
| grade:bathrooms | 1.0650 | 0.2870 |

At the 5% level of significance, all non-interactive terms, with the exception of bathrooms with a p-value of .0508, are less than 0.05 so are significant. None of our interaction terms are significant with p-values substantially greater than 0.05. Given these values we should consider dropping the interactive terms from our final model and retaining all of the non-interactive terms.

### Making Predictions Using Model

Given our model, **Y =-254512.9 + 119.8x1 + 77442.6x2 - 80719.3x3 + 163983.8x4 + 226231.9x5 - 5.0x1x2 + 5.7 x1x3 + 6391.6 x2x3**, we have been asked to make a prediction of the sale price for a home that backs out to a lake and has a 2,150 sq ft living area, 7 grade, and three bathrooms. The model predicts a price of $624,910 for such a home with a 90% prediction interval of (416718, 833102) and a 90% confidence interval of (602886,646934) We can expect future individual observations of price to be between $416,718 and $833,102 given a home with the same view, living area, grade, and number of bathrooms. This accounts for the random variation in individual values. However, based on the population mean, we can expect most observations to fall between $602,886 and $646,934.

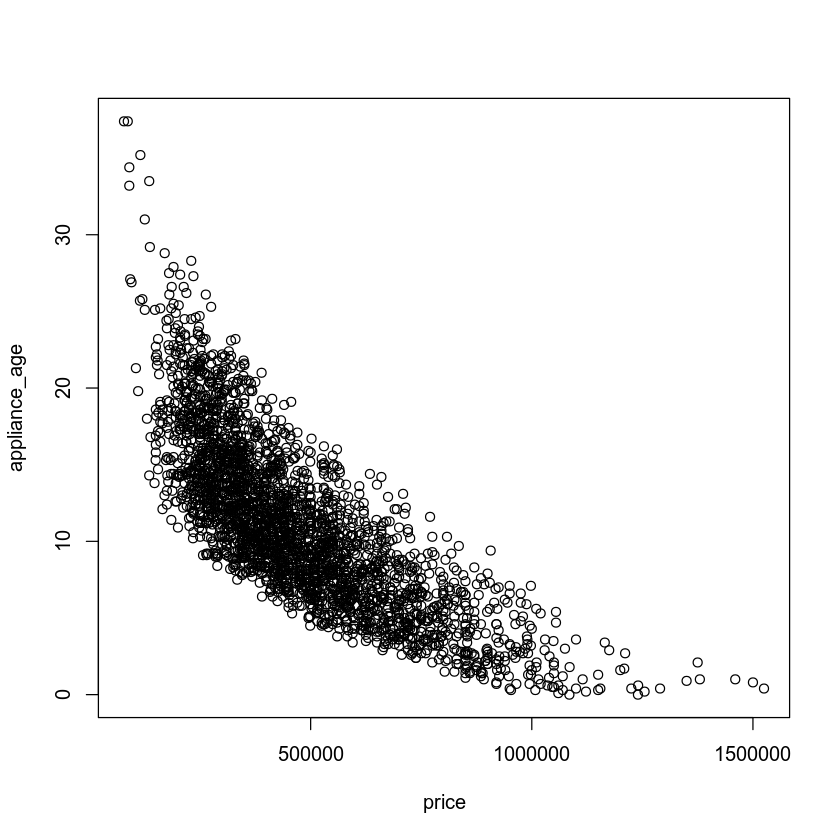
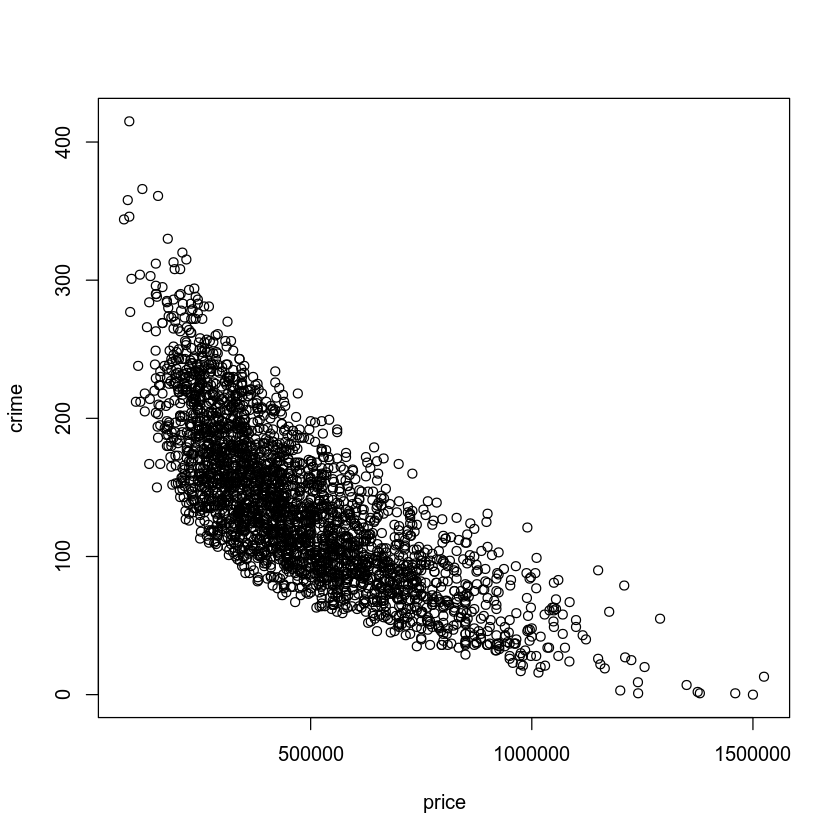
We have also been asked to make a prediction of the sale price for a home that backs out to a road and has a 2,150 sq ft living area, 7 grade, and three bathrooms (all factors are the same as our previous prediction **except** the view. As expected the model predicts a price of $398,678 for such a home with a 90% prediction interval of (191343, 606013) and a 90% confidence interval of (387322,410034). This is exactly what we expected as the difference is $226,232 from the prediction above - the coefficient of the view predictor variable. We can expect future individual observations of price to be between $191,343 and $606,013 given a home with the same view, living area, grade, and number of bathrooms. This accounts for the random variation in individual values. However, based on the population mean, we can expect most observations to fall between $387,322 and $410,034.

The prediction interval will always be wider than the confidence interval as it accounts for random variations in individual observations.

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

### Correlation Analysis

To determine variables to include in our model we first visually inspect the plots of possible predictors against the response variable to determine if there is a significant relationship.

**

From the plot of price vs. crime rate (highlighted in green), we can see a significant negative correlation between price and crime rate. However, it is not a linear relationship, the rate changes over time yielding a curved plot. From the plot of price vs. age (highlighted in purple), we see a very similar significant negative correlation between price and age of appliances, also non-linear and curving. These plots indicate that a second order regression may be appropriate for our model.

We next create a correlation matrix to confirm our visual inspection, and to quantify the amount of correlation that exists. As expected, the correlation coefficient of both appliance age and crime rate are significant and negative with a coefficient of -0.786 and -0.768, respectively. (See Table 3)

**Table 3**

| **correlation** | **price** | **appliance\_age** | **crime** |
| --- | --- | --- | --- |
| **price** | 1 | -0.7859585 | -0.76831 |
| **appliance\_age** | -0.7859585 | 1 | 0.6673604 |
| **crime** | -0.7683115 | 0.66736043 | 1 |

### Reporting Results

The general form of our regression model (model 2) is:

**E(y)=β0 + β1 x1 + β2 x2 + β3 x1 x2 + β4 x12 + β5 x22**

Where y is home price, x1 is avg age of appliances, x2 is the crime rate per 100,000 individuals. This includes interactive terms for each quantitative variable and the second-order (x^2) variable for each as well. Performing our regression analysis we determine our co-efficients for our model of home prices:

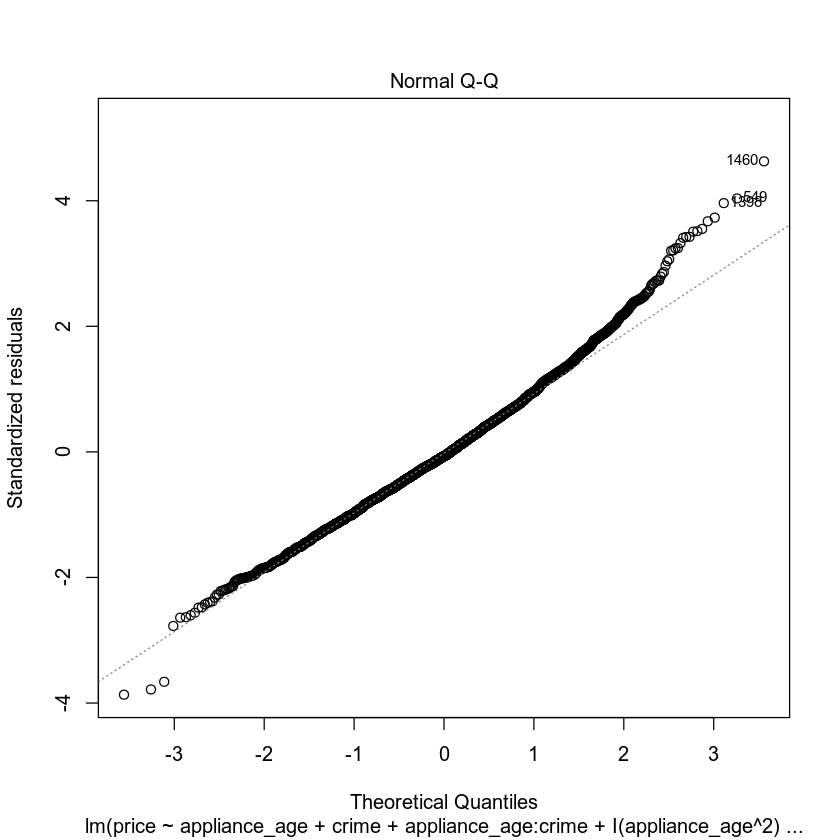
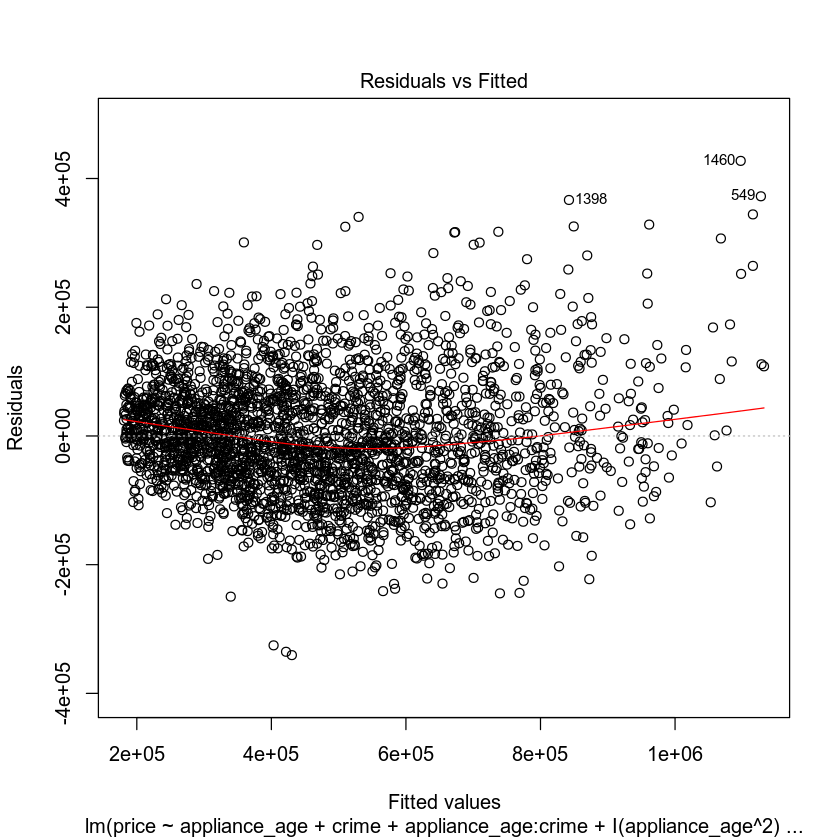
**Y =116100.8 - 22562.2x1 - 3677.6x2 + 13.9x1x2 + 833.0x12 + 6.4x22**

With an R-squared value of 0.8088, the model predicts 80.88% of the variance within our response variable, making the model a good candidate for our use case. The adjusted R-squared value of 0.8084 would be used to compare other models of our response variable to help determine which model was superior.

Our model 2 has the following fitted values and residuals:

| **fitted** | **residual** |  | **fitted** | **residual** |  | **fitted** | **residual** |  | **fitted** | **residual** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 586148.8 | -261148.8 |  | 472026.7 | 67973.3 |  | 557260.1 | 137739.9 |  | 360918.9 | 36581.1 |
| 735037.0 | 9963.0 |  | 826289.5 | -71289.5 |  | 347604.9 | -97604.9 |  | 368434.6 | 160665.4 |
| 450975.9 | 120024.1 |  | 958881.2 | -40881.2 |  | 344077.6 | 60922.4 |  | 349423.1 | 66076.9 |
| 404217.3 | -9217.3 |  | 349416.1 | 61583.9 |  | 538420.2 | 134579.8 |  | 624109.3 | -284109.3 |
| 363986.6 | 85963.4 |  | 339750.8 | -64750.8 |  | 569619.9 | 25380.1 |  | 809586.0 | 182114.0 |
| 362090.5 | -119090.5 |  | 362497.9 | 92502.1 |  | 461867.9 | -224867.9 |  | 297445.3 | 247554.7 |
| 251848.1 | -31898.1 |  | 558200.7 | -219200.7 |  | 384011.4 | -159011.4 |  | 381051.5 | 94948.5 |
| 748588.6 | 50411.4 |  | 479422.4 | -34422.4 |  | 516628.1 | 8371.9 |  | 659741.5 | -92241.5 |
| 378238.2 | -171238.2 |  | 235611.0 | 50339.0 |  | 410588.4 | -55588.4 |  | 359084.5 | -109184.5 |
| 453858.4 | 296141.6 |  | 665152.0 | 124848.0 |  | 574983.5 | -84983.5 |  | 391708.9 | -13208.9 |
| 445518.9 | -95518.9 |  | 378396.6 | -68396.6 |  | 340265.8 | -71265.8 |  | 589535.4 | 90464.6 |
| 348410.5 | -76410.5 |  | 432156.0 | 97794.0 |  | 496119.0 | -133119.0 |  | 583486.4 | 46513.6 |
| 514931.7 | -129981.7 |  | 396118.8 | -57118.8 |  | 263319.6 | 116680.4 |  | 351389.6 | 208610.4 |
| 370476.9 | -98476.9 |  | 696264.5 | 93735.5 |  | 805100.8 | 121399.2 |  | 576005.8 | 73944.2 |
| 403590.1 | -125590.1 |  | 359696.2 | 235303.8 |  | 537692.9 | 12307.1 |  | 542553.5 | 132446.5 |

And plotting residuals yields the following plots:



The plot of residuals vs. fitted values (in green) shows very little funneling, showing a good spread of values mostly equal distance from zero, meeting the assumption of homoscedasticity. The Q-Q plot (in purple) shows residuals following the “normal” slope line very closely, indicating that our model meets the assumption of normality. Overall the model appears to be valid.

### Evaluating Significance of Model

We perform an F-test to identify model significance. The null hypothesis is that no linear relationship exists between any of our predictor variables and the response variable, and our alternate hypothesis is that at least one of our predictor variables has a linear relationship with our response variable such that:

Hnull: β1 = β2 = β3 = β4 = β5 = 0

Halt: β1 or β2 or β3 or β4 or β5 ≠ 0

With a test statistic of 2272 and a p-value of < 2.2e-16, approaching zero, we can reject the null hypothesis and conclude that at least one of the predictor variables is significant with 95% confidence.

To evaluate each individual variable within our model we need to perform a t-test for each variable, with i in range (1:5).

Hnull: βi = 0, predictor having no significance

Halt: βi ≠ , predictor having significance

Results of each are summarized below:

|  | t-value | p-value |
| --- | --- | --- |
| appliance\_age | -30.951 | <2e-16 |
| crime | -24.827 | < 2e-16 |
| appliance\_age^2 | 10.501 | < 2e-16 |
| crime^2 | 8.782 | < 2e-16 |
| appliance\_age:crime | 1.072 | 0.284 |

At the 5% level of significance, all non-interactive terms, are significantly less than 0.05 so are significant. Our interaction term is not significant with p-values substantially greater than 0.05. Given these values we should consider dropping the interactive term from our final model and retaining all of the non-interactive terms.

### Making Predictions Using Model

Given our model, **Y =116100.8 - 22562.2x1 - 3677.6x2 + 13.9x1x2 + 833.0x12 + 6.4x223**, we have been asked to make a prediction of the sale 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. The model predicts a price of $864,423 for such a home with a 90% prediction interval of (711567, 10177280) and a 90% confidence interval of (854109,874738) We can expect future individual observations of price to be between $711,567 and $1,017,280 given a home with the same age of appliances and crime rate. This accounts for the random variation in individual values. However, based on the population mean, we can expect most observations to fall between $854,109 and $874,738.

We have also been asked to make a prediction of the sale price for a home thathas 15-year-old appliances and is in an area that has a crime rate of 200.50 per 100,000 individuals. The model predicts a price of $271,052 for such a home with a 90% prediction interval of (118454, 423649) and a 90% confidence interval of (265866,276257). We can expect future individual observations of price to be between $118,454 and $423,649given a home with the same age of appliances and crime rate. This accounts for the random variation in individual values. However, based on the population mean, we can expect most observations to fall between $265,866 and $276,257.

## 5. Nested Models F-Test

### Reporting Results

The general form of our regression model (model 2, reduced) is:

**E(y)=β0 + β1 x1 + β2 x2 + β3 x1 x2**

Where y is home price, x1 is avg age of appliances, x2 is the crime rate per 100,000 individuals. This includes interactive terms for each quantitative variable. Performing our regression analysis we determine our co-efficients for our model of home prices:

**Y =1145347.7 - 41591.9x1 - 3418.3x2 + 151.0x1x2**

### Evaluating Significance of Model

We perform an F-test to identify model significance. The null hypothesis is that no relationship exists between any of our predictor variables and the response variable, and our alternate hypothesis is that at least one of our predictor variables has a linear relationship with our response variable such that:

Hnull: β1 = β2 = β3 = 0

Halt: β1 ≠ 0 or β2 ≠ 0 or β3 ≠ 0

With a test statistic of 35.73 and a p-value of < 2.2e-16, approaching zero, we can conclude that at least one of the predictor variables is significant with 95% confidence.

To evaluate each individual variable within our model we need to perform a t-test for each variable, with i = (1,2,3,4,5).

Hnull: βi = 0, predictor having no significance

Halt: βi ≠ , predictor having significance

Results of each are summarized below:

|  | t-value | p-value |
| --- | --- | --- |
| appliance\_age | -49.65 | <2e-16 |
| crime | -48.17 | <2e-16 |
| appliance\_age:crime | 31.63 | <2e-16 |

With p-values significantly below 0.05, all of our terms are significant. They should all be retained in our model.

### Model Comparison

A nested model refers to a model that contains all of the elements of the other, plus at least one other term. The model with the additional term or terms is referred to as the complete model, and the other as the reduced model. The F-Test for nested models is used to determine if the additional terms in the complete model are significant, such that they should be included in the final model, or insignificant, such that they should be dropped from the model.

The general form of our reduced regression model is:

**E(y)=β0 + β1 x1 + β2 x2 + β3 x1 x2**

The general form of our complete regression model is:

**E(y)=β0 + β1 x1 + β2 x2 + β3 x1 x2 + β4 x12 + β5 x22**

*y is home price, x1 is avg age of appliances, x2 is the crime rate per 100,000 individuals.*

We perform a nested F-test to determine if the additional quadratic terms in our complete model are significant. The null hypothesis for this test is that the beta estimates for squared terms are zero, meaning that the squared terms are not needed and the reduced model is sufficient. The alternative hypothesis is that at least one of the beta estimates for squared terms is non-zero, meaning that the squared terms are needed and the complete model is necessary such that:

**Hnull: β4 = β5 = 0**

**Halt: β4 ≠ 0 or β5 ≠ 0**

At a 95% confidence level, with a test statistic of 15.95 and a p-value of 1.103452e-06, the test indicates that we can reject the null hypothesis, indicating that the additional terms in the complete model should be retained in our final model.

## 6. Conclusion

Based on our analysis, all three of our models - model 1, model 2, and the reduced model 2 - perform adequately for our use case. However, given the higher alternate R-squared value of 0.8084 for model 2 versus 0.647 for model 1, I would recommend model 2. It accounts for 80.84% of the variance in home prices. We also performeda nested F-test to determine that we should retain the complete model rather than the reduced model. This was confirmed with a p-value of 1.103452e-06.

We have modeled home prices using quantitative predictor variables of the average appliance age and the crime rate per 100,000 individuals. We modeled the interactions of appliance and crime rate and the second order expressions for each predictor variable and conclude the following Second order regression model equation best models our response variable, Sales price of a home:

**Y =116100.8 - 22562.2x1 - 3677.6x2 + 13.9x1x2 + 833.0x12 + 6.4x22,**

*y is home price, x1 is avg age of appliances, x2 is the crime rate per 100,000 individuals.*

All tests performed were designed to test the accuracy and significance of the model as a whole and for each variable within the model. If these tests had failed, it is likely that the model would be accurate only for values within our dataset, and would have no predictive ability whatsoever. Performing these tests provide us confidence that the model is fit for purpose.

All tests performed confirm that our model is fit for purpose, and can be used confidently to predict home prices as a function of the average age of appliances and the crime rate per 100,000 individuals with 80.04% accuracy.