# MAT 303 Project One Summary Report

Mason Cabirac

Mason.cabirac@snhu.edu

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## 1. Introduction

The dataset that I am exploring focuses on the price of homes as a dependent variable, and factors such as size of the living area, age of the home, crime rate, etc. as independent variables. The results may be used to predict the prices of homes. For this project, I have prepared a first order regression model with quantitative and qualitative variables, a complete second order model, and a simplified, first order form of this model. Of these last two, this document contains results of a nested F-test, so as to determine the significance of one model over the other.

## 2. Data Preparation

The important variables in the dataset include price, size of the living area (square feet), size of the upper level (square feet), age, crime rate (per 100,000), average rating of schools in the area, and “view” as a qualitative variable of three possible values, where the home backs out to either a road, to trees, or to a lake.

Number of columns in the dataset: 23

Number of rows: 2692

So it seems that there are 23 variables in total, each represented by a different column, but only those listed above will be used in this report.

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

### Correlation Analysis

A diagram of a red dotted diagram

Description automatically generated with medium confidence

A graph of red dots

Description automatically generated

Though results vary considerably, there is a reasonable suggestion by the first scatterplot that higher price homes will have a greater living area.

Price against age in the scatterplot shows no strong correlation, though it is interesting to note the odd gap in availability of houses in the 80 year mark.

Correlation Coefficients:

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

The Pearson Correlation Coefficients show that price has a moderate to strong correlation with the living area (sq ft). A larger living space will often have a higher price.

Price has a negative, very small correlation with age. That is, it doesn’t matter much how old the house is, if we need to predict the price of it.

### Reporting Results

The general form and prediction equation of the multiple regression model, using price as the response variable and living area, upper level area, age of the home, number of bathrooms, and view as predictor variables:

Y = b0 + b1X0 + b2X1 + b3X2 + b4X3 + b5X4X5 + b6X4X5

Where for bi, i = 0 for the intercept, 1 for the square feet of the living area, 2 for the square feet of the upper level area, 3 for the age, 4 for the bathrooms, 5 and 6 for the view, two coefficients, and thus two terms being necessary.

And for Xi, i = 0 for the square feet of the living area, 1 for the square feet of the upper level area, 2 for the age, 3 for the bathrooms, 4 and 5 for the view (where the back leads to either a lake, to trees, or a road).

Model 1:

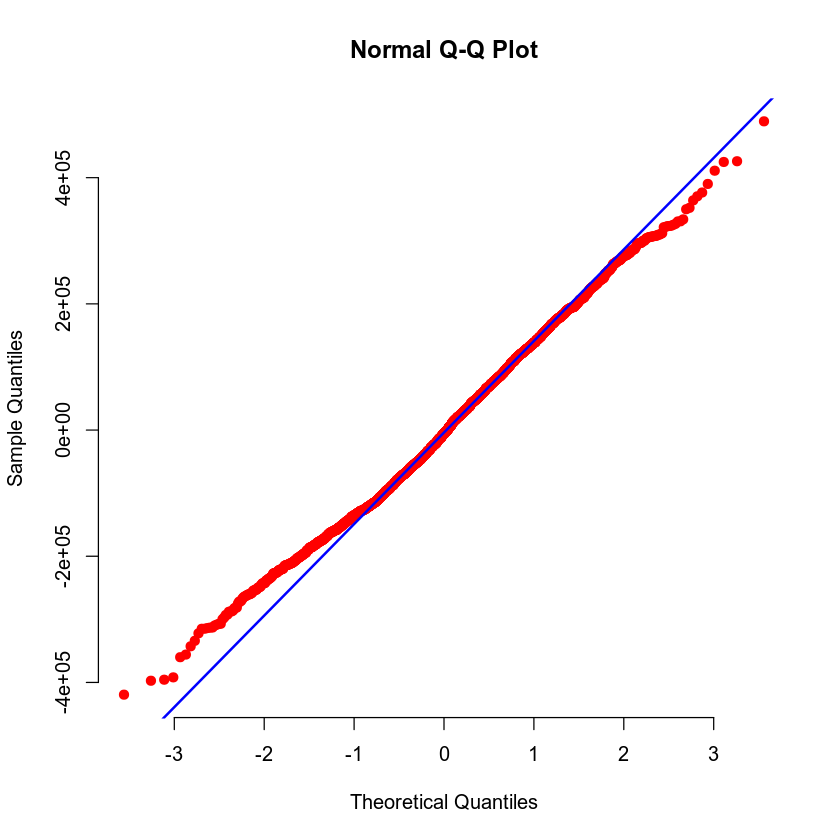
7.709e+03 + 1.293e+02X1 + 1.951e+01X2 + 1.451e+03X3 + 4.397e+04X4X5 + 1.675e+05X4X5

The R2 value (0.6029) and adjusted R2 value (0.602) are nearly the same and pose Model 1 to have weak reliability and consistency. (for which the highest value possible is 1).

The beta estimate for living area shows that as that area increases, there is a rise in price. A lake view will result in an increase in price as well.

A red dot diagram with white text

Description automatically generated



The Residuals against Fitted Values plot shows a clear violation of homoscedasticity. Prices higher than $600,000 are represented on the chart by points further away than before that mark. The densest parts of either side of this appear to be heart-shaped, where the pricier half of the heart seems to be dissolving. There are many outliers which show violate further constance of variance.

The Normal Q-Q plot shows a loyalty to the line of data points between about -1 and 1.5 on the Theoretical Quantiles axis. Serious deviation appears at points less than -1 and greater than 2. This is especially problematic for the lower of those values, which appear to (mis)represent a greater number of points.

### Evaluating Significance of Model

|  |  |
| --- | --- |
| **For this beta-coefficient:** | **P-value** |
| **Intercept** | **0.58495** |
| **Sqft\_living** | **< 2e-16** |
| **Sqft\_above** | **0.00894** |
| **Age** | **< 2e-16** |
| **Bathrooms** | **9.13e-13** |
| **View1** | **< 2e-16** |
| **View2** | **< 2e-16** |

For an overall F-test for Model 1, the null hypothesis is b0 = b1 = b2 = b3 = b4 = b5 = 0. The alternative hypothesis is that any one of these bi does not equal zero. Most of the p-values for each beta-coefficient are less than 0.05, so for a significance factor of 5%, the alternative hypothesis is true and Model 1 is significant in predicting the prices of homes.

For individual t-tests for Model 1, the null hypothesis is that for a particular bi, i = 0. The alternative is that i ≠ 0. Only the p-value for the intercept is more than 0.05, so the null hypothesis is true for the intercept. The intercept, therefore, is not a predictor. All other variables in Model 1 as shown above are predictors, having p-values of less than 0.05.

### Making Predictions Using Model

The predicted price for a home that has 2150 sq ft living area, 1050 sq ft upper level living area, is 15 years old, has 3 bathrooms and backs out to the road has a predicted price of $459,828.2. For such a home, prediction and confidence intervals are as follows:

Prediction Interval:

| **fit** | **lwr** | **upr** |
| --- | --- | --- |
| 459828.2 | 239563 | 680093.4 |
| **Confidence Interval:**  **fit** | **lwr** | **upr** |
| 459828.2 | 446087.9 | 473568.5 |

The predicted price for a home that has 4250 sq ft living area, 2100 sq ft upper level living area, is 5 years old, has 5 bathrooms, and backs out to a lake is $1,074,285. For such a home, prediction and confidence intervals are as follows:

| Prediction Interval: | | |
| --- | --- | --- |
| ***fit*** | ***lwr*** | ***upr*** |
| *1074285* | *852522.6* | *1296048* |

| Confidence Interval: | | |
| --- | --- | --- |
| ***fit*** | ***lwr*** | ***upr*** |
| *1074285* | *1045117* | *1103454* |

The prediction intervals account for the uncertainty of future predictions, while the confidence intervals only account for the data that the model has already used. Thus, the prediction intervals are wider.

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

### Correlation Analysis

A graph of a scatter plot

Description automatically generated

A graph of red dots

Description automatically generated

Second order models may be appropriate for the parabolic images of these scatterplots. There is some apparent evidence of quadratic correlations between price and school ratings and between price and crime.

With higher school ratings, there is an increase in price, except that as the price increases, the school ratings level off. With a lower crime rate, there is a steep rise in price that also levels off. But this fails to illustrate that most data points are heavily concentrated to a middle section of each parabola, so it is not clear whether a second order model is truly necessary.

### Reporting Results

The general form and prediction equation of a complete second order model for price using an average school rating of in the area and crime rate per 100,000 people as predictors:

Y = b0 + b1X0 + b2X1 + b3X02+ b4X12+ b5X0X1

For bi, i = 0 for the intercept, 1 for the school rating, 2 for the crime rate, 3 for school\_rating2, 4 for crime2 and 5 for school\_rating:crime.

And where for Xi, i = 0 is for the school rating and 1 for the crime rate.

Model 2:

7.339e+05 - 7.375e+04X0 - 3.155e+03X1 + 1.165e+04X02 + 6.377e+00X12 - 5.227e+01X0X1

The R2 value (0.8088) and adjusted R2 value (0.8084) are nearly the same and pose Model 2 to be very reliable and consistent (for which the highest value possible is 1). However, the significance values will show whether the model’s predictions are true to direct observation.

A diagram of red dots

Description automatically generated

A graph with a red line

Description automatically generated

The Residuals against Fitted Values shows a violation of homoscedasticity. There is a dense oval shape centered about y = 0. That density is somewhat lost at about y = 1e+05 and x = 6e+05, in each case approaching greater values of x and y. Variance increases among outliers.

The Normal Q-Q plot is accurate enough between x = -3 and x = 1.5. There is some deviation of values less than -2, but representation is a great matter of worry for values greater than 1.5 and especially 2.5.

### Evaluating Significance of Model

**The p-values for each beta-coefficient:**

|  |  |
| --- | --- |
| **For this beta-coefficient:** | **P-value** |
| **Intercept** | **1.45e-12** |
| **School\_rating** | **0.000406** |
| **Crime** | **1.90e-09** |
| **School\_rating2** | **< 2e-16** |
| **Crime2** | **< 2e-16** |
| **School\_rating:crime** | **0.281513** |

For an overall F-test for Model 1, the null hypothesis is b0 = b1 = b2 = b3 = b4 = b5 =b6 = 0. The alternative hypothesis is that any one of these bi does not equal zero. Most of the p-values for each beta-coefficient are less than 0.05, so for a significance factor of 5%, the alternative hypothesis is true and Model 2 is significant in predicting the prices of homes.

For individual t-tests for Model 2, the null hypothesis is that for a particular bi, i = 0. The alternative is that i ≠ 0. Only the p-value for school\_rating:crime is more than 0.05, so the null hypothesis is true for the intercept. The school rating and crime interaction, therefore, is not a predictor. All other variables in Model 2 as shown above are predictors, having p-values of less than 0.05.

### Making Predictions Using Model

For 9.80 school rating and 81.02 per 100,000 crime rate, the predicted price is $874,497 –

Prediction Intervals:

|  | | |
| --- | --- | --- |
| **fit** | **lwr** | **upr** |
| **874497** | **721606.2** | **1027388** |

**Confidence Intervals:**

|  | | |
| --- | --- | --- |
| **fit** | **lwr** | **upr** |
| **874497** | **863681.4** | **885312.7** |

For 4.28 school rating and 215.50 per 100,000 crime rate, the predicted price is $199,706.7 --

**Prediction Intervals:**

|  | | |
| --- | --- | --- |
| **fit** | **lwr** | **upr** |
| **199706.7** | **46991.65** | **352421.7** |

**Confidence Intervals:**

|  | | |
| --- | --- | --- |
| **fit** | **lwr** | **upr** |
| **199706.7** | **191753.5** | **207659.9** |

## 5. Nested Models F-Test

### Reporting Results

The general form and prediction equation of a first order model for price using average school rating in the area and crime rate per 100,000 people as predictors is as follows:

Y = b0 + b1X0 + b2X1 + b3X2

Where for bi, i = 0 is for the intercept, 1 for the school rating, 2 for the crime rate and 3 for the interaction term of the school rating and crime together.

And for X0, i = 0 is for the school rating, 1 is for the crime rate.

Reduced Model:

Y = -410233.37 - 155559.97X0 - 2230.07X1 - 564.85X0X1

### Evaluating Significance of Model

P-values for the reduced model:

|  |  |
| --- | --- |
| **For this beta-coefficient:** | **P-value** |
| **Intercept** | **<2e-16** |
| **School\_rating** | **<2e-16** |
| **Crime** | **<2e-16** |
| **School\_rating:crime** | **<2e-16** |

The p-value for each beta-coefficient is less than 0.05. For an overall F-test, the null hypothesis is b0 = b1 = b2 = b3 = b4 = 0. The alternative hypothesis is that for at least one i, bi ≠ 0. Therefore, the null hypothesis is rejected and the reduced model is significant at a 5% level.

The p-value for each beta-coefficient is less than 0.05. For an individual t-test, the null hypothesis is

bi = 0for i= 1, 2, 3 or 4. The alternative hypothesis is bi ≠ 0 for i = 1, 2, 3 or 4. Each individual bi has a p-value less than 0.05. So individually, the null hypothesis is rejected for each individual t-test, rendering each variable a predictor of home prices.

### Model Comparison

In general, a reduced model is a simplified version of a complete model. The reduced model does not have all of the terms of the complete model, but the complete model does have all the terms of the reduced model.

The general form and prediction equation of the reduced model:

Y = b0 + b1X0 + b2X1 + b3X2

Y = -410233.37 - 155559.97X0 - 2230.07X1 - 564.85X0X1

The general form and prediction equation of the complete model (Model 2):

Y = b0 + b1X0 + b2X1 + b3X02+ b4X12+ b5X0X1

7.339e+05 - 7.375e+04X0 - 3.155e+03X1 + 1.165e+04X02 + 6.377e+00X12 - 5.227e+01X0X1

For a nested model F-test the null hypothesis is that the beta-estimates for the squared terms are 0, meaning that the variables for those terms are not predictors. The alternative hypothesis is that they are not equal to 0, meaning that the variables for those terms are predictors. The p-value for the test is 2.22716e-28, which is less than 0.05. Therefore, the null hypothesis is rejected, and the complete model is necessary.

## 6. Conclusion

Of the models compared in the nested F-test, I would choose Model 2 to predict house prices. Model 1 is not very consistent according to its R2 values.