# MAT 303 Project One Summary Report

Kayla Sacks

kayla.sacks@snhu.edu

Southern New Hampshire University

## 1. Introduction

For this statistical analysis I will be using multiple variables from a historical housing data set. I will be using this data set to analyze how different variables can affect housing pricing. I will start by creating a multiple regression model using both quantitative and qualitative variables. I will then create a complete second order regression model using two quantitative variables and finish with a nested model of the second order regression model. Realtors can use the results of this analysis to estimate the house prices in different areas or to provide an estimate to customers that have specific needs or requests for the home they want.

## 2. Data Preparation

The data set that is being used has 23 columns and 2692 rows. The important variables I will be using for this analysis are:

* price – Sale price of the home, the response variable for all models.
* sqft\_living – Size of the living area in sqft, a predictor variable used in the first model.
* sqft\_above – Size of the upper level in sqft, a predictor variable used in the first model.
* age – Age of the home, a predictor variable used in the first model.
* bathrooms – Number of bathrooms in the house, a predictor variable used in the first model.
* view – If the house backs out to a lake, trees, or a road, a predictor variable used in the first model. This is a qualitative variable with road as the base.
* school\_rating – Average school rating in the area, a predictor variable used in the complete and reduced second model.
* crime – Crime rate per 100,000 people, a predictor variable used in the complete and reduced second model.

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

### Correlation Analysis

To start the analysis, two scatterplots were created to visualize the relation of price vs living area and price vs age of home. These scatterplots can be seen below.

*A graph of red dots

Description automatically generated*

The scatterplot of price against living area shows a clear trend. We can see that as living area increases, so does price. The scatterplot of price against home age does not show any trends.

A screenshot of a graph

Description automatically generated

From the correlation matrix above, we can see that:

* The correlation coefficient between price and living area (sqft\_living) is 0.6895. This shows a moderate positive correlation.
* The correlation coefficient between price and age is -0.0746. This shows a very weak negative correlation.

### Reporting Results

The general form for this multiple regression model is:

The prediction equation for the regression model is:

Where is the predicted value for price, is living area (sqft\_living), is the side of the upper level (sqft\_above), is the age of the home, is the number of bathrooms, is view1 and is view2. View1 and view2 represent the view that the home backs out to. These variables were used to create the following regression model:

*A screenshot of a computer

Description automatically generated*

Using the output of the regression model, the updated prediction equation is:

The (R-squared) value for this regression model is 0.6029 and the (adjusted R-squared) is 0.602. This R-squared value tells us that the regression model explains 60% of the variation in house pricing using the predictor variables.

The beta estimate for living area is 1.293e+02 which means for each 1 increase in sqft, the house price will increase by . The beta estimate for a lake view is 2.490e+05 which means that if the house has a lake view, the price will increase by .

The Residuals against Fitted Values plot below does not show any pattern which supports an assumption of homoscedasticity and linearity.

*A red dot diagram with numbers and a white background

Description automatically generated*

Although there is slight deviation from the line on the Normal Q-Q plot below, it is not a significant amount.

*A graph with red line

Description automatically generated*

### Evaluating Significance of Model

To evaluate if the model is significant to a 5% level of significance, the overall F-test will be performed. For this test we will start by testing if the model rejects the null hypothesis () in favor of the alternate hypothesis (). The equations for these are:

The overall p-value for this model is < 2.2e-16 which is less than the 0.05 level of significance so we will reject the null hypothesis. The rejection of the null hypothesis for the F-test tells us that at least one of the predictor variables is statistically significant to the dependent variable. Now we will move onto the individual T-test to see which beta values will meet the 5% level of significance by using the null hypothesis and alternate hypothesis. The hypotheses for the beta values with i = 1, 2, 3, 4, 5, 6 is:

= Living area(sqft\_living) p-value is < 2e-16.

= Upper level size (sqft\_above) p-value is 0.00894.

= Age of house (age) p-value is < 2e-16.

= Number of bathrooms(bathrooms) p-value is < 9.13e-13.

= Home backs out to trees(view1) p-value is < 2e-16.

= Home backs out to a lake(view2) p-value is < 2e-16.

With the conclusion of the individual T-test, we can see that all beta variables meet the 0.05 level of significance and reject the null hypothesis in favor of the alternate hypothesis.

### Making Predictions Using Model

**Prediction 1**

The following equation is used to predict the price for a house that has 2150 sqft living area, 1050 sqft upper level living area, is 15 years old, has 3 bathrooms, and backs out to road:

The calculated price from the equation is close to the value from the model prediction which is 459,828.2.

The 90% prediction interval for an individual response tells us that we can be 90% certain that a house price will fall within the interval of (239563, 680093.4) using the prediction values above.

The 90% confidence interval for the mean for house pricing tells us that we can be 90% certain that the average house price will fall within the interval of (446087.9, 473568.5) using the same prediction values above.

**Prediction 2**

The following equation is used to predict the price for a house that has 4250 sqft living area, 2100 sqft upper level living area, is 5 years old, has 5 bathrooms, and backs out to a lake:

The calculated price from the equation is close to the value from the model prediction which is 1,074,285.

The 90% prediction interval for an individual response tells us that we can be 90% certain that a house price will fall within the interval of (852522.6, 12196048) using the prediction values above.

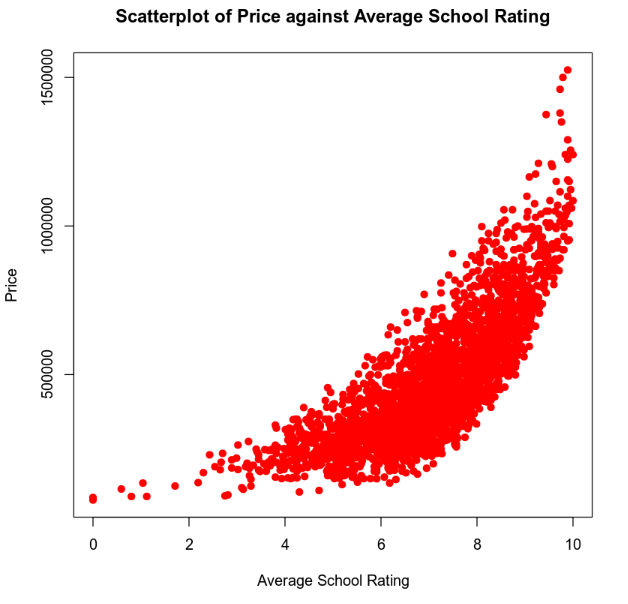
The 90% confidence interval for the mean for house pricing tells us that we can be 90% certain that the average house price will fall within the interval of (1045117, 1103454) using the same prediction values above.

Since the prediction interval takes uncertainty into consideration, it will be wider than the confidence interval.

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

### Correlation Analysis

For the next part of the analysis, two new scatterplots were created. These scatterplots for price vs average school rating in the area and for price vs the crime rate per 100,000 people is below.

*A graph of a crime rate

Description automatically generated*

*Since nether scatterplot above follows a linear pattern as they show a noticeable curve, a second order model would be appropriate using these variables.*

### Reporting Results

The general form for this complete second order model which includes two predictor variables:

The prediction equation for this regression model is:

The second order regression model for price with school rating and crime rate was created with the summary shown below.

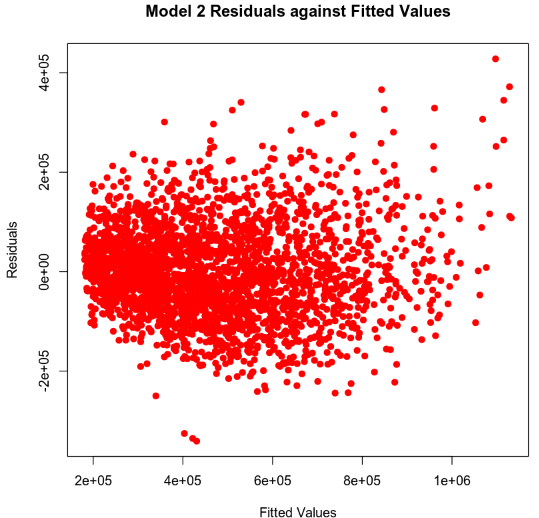
*A screenshot of a computer

Description automatically generated*

Using the outputs from the regression model, the prediction model is updated.

The (R-squared) value for this model is 0.8088 with a (Adjusted R-squared) at 0.8084. The tells us that this complete second order model explains 81% of the variation in price using the school rating and crime rate.

The Residuals against Fitted Values plot below does not show any pattern which supports an assumption of homoscedasticity and linearity.

**

Although there is a slightly larger deviation from the line on the Normal Q-Q plot below then with model1, it is still not a significant amount.

*A graph with red and blue lines

Description automatically generated*

### Evaluating Significance of Model

To evaluate if the model is significant to a 5% level of significance, the overall F-test will be performed. For this test we will start by testing if the model rejects the null hypothesis () in favor of the alternate hypothesis (). The equations for these are:

The overall p-value for this model is < 2.2e-16 which is less than the 0.05 level of significance so we will reject the null hypothesis. The rejection of the null hypothesis for the F-test tells us that at least one of the predictor variables is statistically significant to the dependent variable. Now we will move onto the individual T-test to see which beta values will meet the 5% level of significance by using the null hypothesis and alternate hypothesis. The hypotheses for the beta values with i = 1, 2, 3, 4, 5 is:

= average school rating(school\_rating) p-value is 0.000406.

= crime rate per 100,000 (crime) p-value is 1.90e-09.

= Interaction term between school\_rating and crime p-value is 0.281513.

= school\_rating² p-value is < 2e-16.

= crime² p-value is < 2e-16.

With the conclusion of the individual T-test, we can see that the only beta value that does not reject the null hypothesis is the interaction term with a p-value of 0.281513 which does not meet the 0.05 level of significance. Since the interaction term does not reject the null hypothesis, we can conclude that it is not statistically significant but the other predictor variables are statistically significant to home price.

### Making Predictions Using Model

**Prediction 1**

The following equation is used to predict the price for a house in an area with an average school rating of 9.80 and a crime rate of 81.02 per 100,000 individuals:

The calculated price from the equation is close to the value from the model prediction which is 874,497.

The 90% prediction interval for an individual response tells us that we can be 90% certain that a house price will fall within the interval of (721606.2, 1027388) using the prediction values above.

The 90% confidence interval for the mean for house pricing tells us that we can be 90% certain that the average house price will fall within the interval of(863681.4, 885312.7) using the same prediction values above.

**Prediction 2**

The following equation is used to predict the price for a house in an area with an average school rating of 4.28 and a crime rate of 215.50 per 100,000 individuals:

The calculated price from the equation is close to the value from the model prediction which is 199706.7.

The 90% prediction interval for an individual response tells us that we can be 90% certain that a house price will fall within the interval of (46991.65,352421.7) using the prediction values above.

The 90% confidence interval for the mean for house pricing tells us that we can be 90% certain that the average house price will fall within the interval of (191753.5, 207659.9) using the same prediction values above.

## 5. Nested Models F-Test

### Reporting Results

The general form for this first order model which includes two predictor variables:

The prediction equation for this regression model is:

The first order regression model for price with school rating and crime rate was created with the summary shown below.

*A screenshot of a computer

Description automatically generated*

Using the outputs from the regression model, the updated prediction model is:

### Evaluating Significance of Model

To evaluate if the model is significant to a 5% level of significance, the overall F-test will be performed. For this test we will start by testing if the model rejects the null hypothesis () in favor of the alternate hypothesis (). The equations for these are:

The overall p-value for this reduced model is < 2.2e-16 which is less than the 0.05 level of significance so we will reject the null hypothesis. The rejection of the null hypothesis for the F-test tells us that at least one of the predictor variables is statistically significant to the dependent variable. Now we will move onto the individual T-test to see which beta values will meet the 5% level of significance by using the null hypothesis and alternate hypothesis. The hypotheses for the beta values with i = 1, 2, 3 is:

With the conclusion of the individual T-test, we can see that all three beta variables have a p-value < 2e-16 so they will all reject the null hypothesis. We can conclude that all three predictor variables are statistically significant to house pricing.

### Model Comparison

We can now compare this model with the complete second order model for price using the average school rating in the area and crime rate per 100,000 people as predictors to test is the quadratic terms are statistically significant in predicting the prices of homes.

In general, a complete model will include all terms to be evaluated while a reduced model only includes some of the possible terms. The nested F-test will be used to assess if we can exclude the variables removed from the reduced model, which in this case are the quadratic terms.

The general form for the complete and reduced model is:

Complete model:

Reduced model:

The prediction regression equation for the complete and reduced model is:

Complete model:

Reduced model:

For both the complete and reduced models, is the predicted price value, is the average school rating and is the crime rate per 100,000 people. with as the slope parameter, is the beta estimate for school rating, is the beta estimate for crime rate, is the beta estimate for the interaction term, is the beta estimate for school rating squared and is the beta estimate for crime squared.

Now we will run the nested model F-test at a 5% level of significance to evaluate if the quadratic (squared) terms are needed. The output of the ANOVA F-test is shown below.

*A white background with black text

Description automatically generated*

We will start with the null hypothesis and alternate hypothesis which are:

For this model, the null hypothesis is testing if the coefficient for the two quadratic (squared) terms are 0, and if true it means that the reduced model is sufficient otherwise the alternate hypothesis, which tells that at least one of the coefficients is statistically significant, will be used. We will be testing to a 5% level of significance. The p-value from the nested model F-test (ANOVA) is 2.22716e-28 which is much lower than the 0.05 level of significance meaning we will reject the null hypothesis and use the alternate hypothesis. The conclusion of this test tells us that the reduced model is not sufficient and the complete second order model should be used.

## 6. Conclusion

This analysis includes a first order regression model, complete second order regression model, a reduced model, and a nested F-test to evaluate if the complete model is needed or if the reduced model is sufficient. After completing each test, I would use the second model to predict house price in leu of the first model or the reduced model. Although all variables in the first model rejected the null hypothesis and are statistically significant, the five predictor variables (living area, upper level size, age of the home, number of bathrooms, and view) only explains 60% of the variation in the dependent variable. The second model, which uses the two predictor variables of average school rating and crime per 100,000 people, an interaction term between the two variables as well as their squared values is 81%. This tells us that 81% of the variation in the dependent variable can be explained the two variables.

The complete second order model was chosen over the reduced model since when the nested F-test was done, it was concluded that the p-value was much lower than the 0.05 level of significance so the null hypothesis would be rejected in favor of the alternate hypothesis. Using the alternate hypothesis shows that the coefficients for the squared terms are non-zero and the complete model should be used.

The practical importance of the analysis that was performed is that it can be used to predict what a home is worth based on several factors such as the average school rating and the crime rate per 100,000 people. This data can be used when appraising the value of a house or evaluating if a house is priced reasonably.