# MAT 303 Project One Summary Report

Justin Farquhar

Justin.farquhar@snhu.edu

Southern New Hampshire University

## 1. Introduction

This data set contains a housing data set containing multiple variables that are likely to affect the overall price of a house. This contains the price, the number of bedrooms, the number of bathrooms, the size of the living area, the size of the upper level, the size of the lot, the age of the home, the measure of craftsmanship and quality of materials used, the age of all appliances, the crime rate, if the home has a backyard, the school rating, and the type of view the house has. These results can be used to set better prices to help ensure that the buyer is getting their money worth and so that the seller is getting the most that they are able to. This could help houses sell faster so they are not sitting on the market as long, getting people into their new home sooner. Through this project I will be making multiple models, including first and second order, as well as a nested model to determine the necessity of complexity within the model. This includes providing general and predictive equations for the models, as well as finding the residuals and fitted values and plotting those against one another.

## 2. Data Preparation

The most important variable in my opinion is price, as it is what we are looking at within all of these models. Outside of that, bedrooms, bathrooms, the various area variables, crime, and school rating are what I believe to be important. Bedrooms and bathrooms because that is one of the main things that someone would be looking for specifically within a new house, whether that be upgrading their current place or trying to find a nice balanced starter home. The multiple square foot variables are important as again they are typically factors that people specifically look for within the house they are trying to get. Even if a house has the number of bedrooms and bathrooms that they are looking for, if they are extremely small it mostly defeats the point. For crime rate, someone is going to be much less likely to purchase a home with a substantially increased rate of crime, which could greatly affect the overall price. Finally for school rating, if the potential buyer has kids, or is planning for kids, having the best possible options for their children to succeed and get a better education. Within this dataset there are 23 columns and 2692 rows.

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

*A red dot diagram with numbers

Description automatically generated*

*A graph of blue dots

Description automatically generated*

Looking at the first scatterplot, we can see living area plotted with price. The information is unfortunately very condensed, but there is a very clear positive linear relationship between the two. Indicating that as living area of price increases, the other also increases. The second scatter plot has the age of the home and price plotted with each other. Within this plot, there is much more variability than the first, indicating that there is not really a clear relationship between the age of the home and the price.

A screenshot of a table

Description automatically generated

This next image shows the correlation coefficients between price, size of the living area, and age of the home. To make any analysis off of these values it is important to know that the closer to 1.0 that the number is, the more related that the values are. Taking this into account, price and the size of the living area has a value of 0.6895, indicating a fairly strong positive relationship. When looking at price and age, the value is only -0.0746, indicating that there is a very miniscule negative relationship between the two.

### Reporting Results

General Form

Prediction equation

Looking at the overall model, we can analyze the R values of the data set to see how well these variables account for the changes in price. According to the model, the R-squared value is 0.6029 and the adjusted R-squared value is 0.602. This means that these variables account for approximately 60.29% of the variability of price. With the R-squared value being nearly the same as the adjusted R-squared value indicates that the model is not overfitting the data and that the predictors in the model are meaningful and contribute to explaining the variation in the dependent variable. To then look at the beta estimates for the living area and lake view predictors, living area’s beta value is 129.3 and lake view’s is 249000. The value for living area is likely on the lower side as the square footage can end up being fairly large, leading to higher increases as this value grows. Then looking at lake view, it makes sense that a house having a direct lake view provides a substantial increase to the price, though I did not expect it to be as large as it is.

A diagram of red dots

Description automatically generated

A graph with a red line

Description automatically generated

Looking at the first scatterplot which had the residuals plotted against the fitted values, we can see that the values appear as a fairly random scatter around the horizontal 0 line. There is some grouping towards the left of the plot, but I believe this is fine and the values are still spread around enough to still fall within homoscedasticity. To then look at the Normal Q-Q plot, we can see that the values stay along the line with a small tail towards the left side. Having the values close to the line like this helps validate the normality of residuals, which is the case despite the little tail to the left.

### Evaluating Significance of Model

To then look through the significance of the model at a 5% level of significance, the overall model has a significance of less than 2.2e-16, indicating that it is significant. To then look at each variable, the p values for sqft living, age, view1, and view2 are all less than 2e-16, showing that each of these predictors are significant. For sqft\_above, the p value is 0.00894, indicating that it is significant. For bathrooms, the p value is 9.13e-13, indicating that it is significant. To look at the null hypothesis, we would have one of these values being 0, which is not the case, meaning we do not accept the null hypothesis and in the same sense, we can then accept the alternative hypothesis.

### Making Predictions Using Model

A screenshot of a computer

Description automatically generated

The above image includes the prediction interval and confidence interval for a home 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 predicted value is 459828.2, with a lower limit of 197338.3 and an upper limit of 722318.1. For the confidence interval there is a lower limit of 443453.9 and an upper limit of 476202.6. The confidence interval is much more narrow as it is accounting for the uncertainty in the estimated mean, meanwhile the prediction interval is wider because it accounts for uncertainty in the model coefficients and random variation.

A screenshot of a computer

Description automatically generated

T The above image includes the prediction interval and confidence interval for a home 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 predicted value is 1074285, with a lower limit of 810010.7 and an upper limit of 1338560. For the confidence interval there is a lower limit of 1039525 and an upper limit of 1109045.

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

A diagram of a scatter plot

Description automatically generated

A blue dotted graph with white text

Description automatically generated

Looking at the first scatterplot that plot price against school rating we can see a clear positive linear relationship. This shows that as school rating increases, so does price. To then look at the second scatter plot that plots price against crime rate per 100000 people, we can see a negative linear relationship. This shows that as crime rate decreases, price increases. Seeing that each scatterplot has a curve to the theoretical line going through the data, a second order model is very appropriate for using these variables.

### Reporting Results

General Form

Prediction equation

For this model we get a R-squared value of 0.8088 and an adjusted R-squared value of 0.8084. This means that these variables account for approximately 80.88% of the variability of price. With the R-squared value being nearly the same as the adjusted R-squared value indicates that the model is not overfitting the data and that the predictors in the model are meaningful and contribute to explaining the variation in the dependent variable.

A diagram of red dots

Description automatically generated

A graph of a normal q-q plot

Description automatically generated

Looking at the first scatterplot which had the residuals plotted against the fitted values, we can see that the values appear as a fairly random scatter around the horizontal 0 line. There is some grouping towards the left of the plot, but I believe this is fine and the values are still spread around enough to still fall within homoscedasticity. To then look at the Normal Q-Q plot, we can see that the values stay along the line with a small tail towards the right side. Having the values close to the line like this helps validate the normality of residuals, which is the case despite the little tail to the right.

### Evaluating Significance of Model

To determine if the model is significant at a 5% level of significance we again look through the p values of the model. To start, the overall model p value is less than 2.2e-16, which indicates that it is significant. For school\_rating, the p value is 0.000406, indicating that it is significant. For crime, the p value is 1.90e-09, indicating that it is significant. For school\_rating^2 the p value is less than 2e-16, indicating that it is significant. For crime^2 the p value is less than 2e-16, indicating that it is significant. For school\_rating with crime the p value is 0.281513, indicating that it is not significant. As there are no values that are 0, we can reject the null hypothesis, which then allows us to accept the alternative hypothesis.

### Making Predictions Using Model

A screenshot of a computer

Description automatically generated

The above image includes the prediction interval and confidence interval for a home that has an average school rating of 9.80 and a crime rate of 81.02. The predicted value is 874497, with a lower limit of 721606.2 and an upper limit of 1027388. For the confidence interval there is a lower limit of 863681.4 and an upper limit of 885312.7.

A screenshot of a computer

Description automatically generated

The above image includes the prediction interval and confidence interval for a home that has an average school rating of 4.28 and a crime rate of 215.50. The predicted value is 199706.7, with a lower limit of 46991.65 and an upper limit of 352421.7. For the confidence interval there is a lower limit of 191753.5 and an upper limit of 207659.9.

## 5. Nested Models F-Test

### Reporting Results

General Form

Prediction equation

For this model we get a R-squared value of 0.7995 and an adjusted R-squared value of 0.7993. This means that these variables account for approximately 79.95% of the variability of price. With the R-squared value being nearly the same as the adjusted R-squared value indicates that the model is not overfitting the data and that the predictors in the model are meaningful and contribute to explaining the variation in the dependent variable.

### Evaluating Significance of Model

To determine if the model is significant at the 5% level of significance, we again look at the p values. The p value of the overall model is less than 2.2e-16, indicating that it is significant. Each of the other predictor variables school\_rating, crime, and school\_rating to crime all have a p value of less than 2e-16, indicating that they are all significant. As none of these values are 0, we can reject the null hypothesis again and accept the alternative hypothesis.

### Model Comparison

A reduced model is a simpler version of the regression model, containing fewer predictors. Meanwhile, the complete model is a much more complex version of the regression model containing all predictors of interest.

Reduced Model

Complete Model

To compare these two models we run a different test directly comparing them and giving us another p value, which returned 2.22716e-28, which strongly indicates that the additional terms in the complete model significantly improve the fit compared to the reduced model. As the value is not 0, we can reject the null hypothesis, and as such can accept the alternative hypothesis.

## 6. Conclusion

After going through these analyses, I would opt to use the second or the complete model. This model provided the highest R-squared value, indicating that the predictor variables were actually accounting for the variability within price. The reduced model was only slightly behind in R-squared value, but based on our last tests the complete model’s additional variables significantly improved the model.

The practical importance of these analyses allows us to better understand which of these variables how an overall effect on the price of houses. As a prospective homebuyer, this is very interesting to see, and allows me to see how far off the current market is to the actual value of the house that I am looking at.