# MAT 303 Project One Summary Report

Sarah Steinbaum

sarah.steinbaum@snhu.edu

Southern New Hampshire University

## 1. Introduction

The following analyses will be used to predict the price of a home based on certain factors such as the size of a house’s living space and the crime rate for the area. The data set is a large set of historical data to assist in predicting house prices around the Seattle area. The first analysis will use first order regression with quantitative and qualitative variables including the number of bathrooms, size of the living space and upper-level space, the age of the home, and the view. The second analysis will use complete second order regression with quantitative variables including the school rating and crime rate per 100,000 people. We will then analyze a nested model of the complete second order regression model and conduct an F-test to determine if a reduced model is sufficient or if a complete model is needed.

## 2. Data Preparation

This data set includes many important variables. The response variable is the price of the home (*price*). The quantitative predictor variables are the number of bathrooms (*bathrooms*), square footage of the living space (*sqft\_living*), square footage of the upper level (*sqft\_upper*), the age of the home (*age*), school rating (*school\_rating*), and the crime rate (*crime*). The qualitative predictor variable is the view which can back out to a road (view=0), trees (view=1), or lake (view=2). The data set is compromised of 23 columns and 2,962 rows.

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

### Correlation Analysis

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

The first scatterplot shows the relationship between the price and a house and the size of the living area in square feet. There is a positive trend that shows as the size of the house increases, generally, so does the price. The second scatterplot portrays the correlation between the price of a house and its age which shows little to no correlation as the data points do not show any particular pattern. The correlation coefficient between price and the living area is 0.6895 which indicates a moderately strong positive correlation. However, the correlation coefficient between price and age of the house is -0.0746 which indicates a very weak negative correlation between the variables as the coefficient is very close to zero.

### Reporting Results

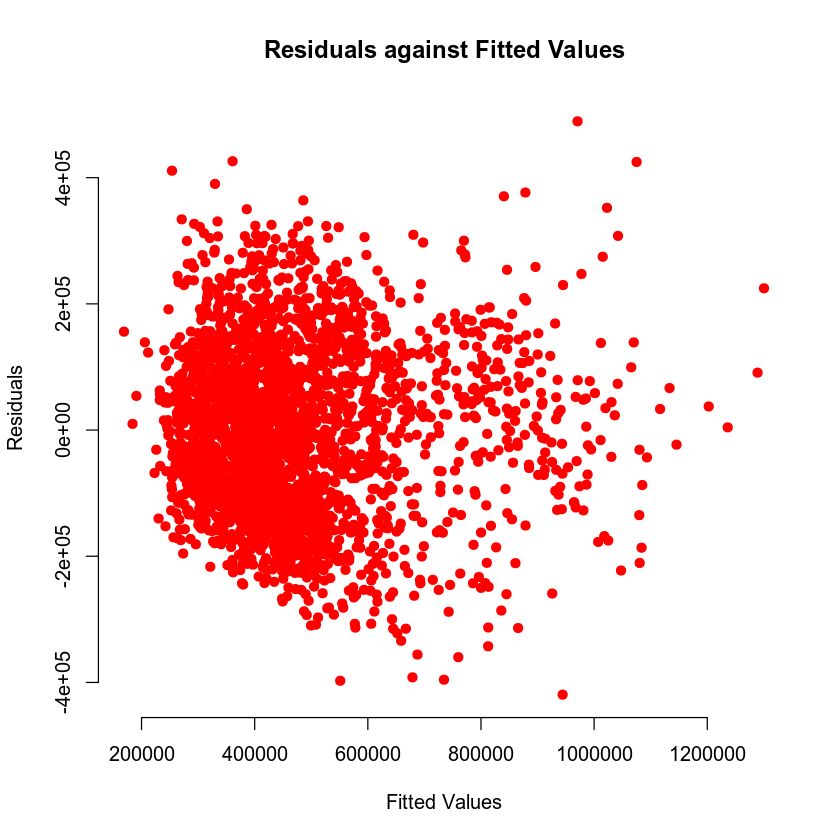
The general form of the multiple regression model is:

The prediction equation is:

In this model, represents the response variable price, represents living area, represents upper-level area, represents age of the home, represents number of bathrooms, and represent the view which is a qualitative predictor variable. The view can face a lake represented by view=2, trees represented by view=1, or a road represented by view=0. Once the R script has run, we can place the beta estimates in the equation:

The value of R-squared is 0.6029 and is the coefficient of multiple determination. It accounts for roughly 60% of the variation for the response variable, price, can be explained by the predictor variables and the qualitative predictor variables in this model. The Adjusted R-squared is the adjusted coefficient of determination and its value is 0.602 in this model.

The beta estimate for the living area is 1.293E+02 in scientific notation. This means that for every unit increase in living every, price will increase by 1.293E+02. The beta estimate for lake view is 2.490E+05 in scientific notation. This means that for every unit increase in the lake view the price will increase by 2.490E+05.



Chart, line chart

Description automatically generated

The first plot portrays the residuals against fitted values and provides no discernable pattern which indicates homoscedasticity. The second plot is the Normal Q-Q Plot which indicated a normal distribution as the values are on or near the trend line.

### Evaluating Significance of Model

In order to determine if the model is statistically significant at a 5%, or 0.05, level of significance, we can conduct an overall F-test. The null hypothesis is that no relationship exists between the price of a home and the qualitative and quantitative predictor variables. The alternative hypothesis is that a relationship does exist between the response variable and at least one of the variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | F-statistic | P-value | Conclusion |
| Price  (price) | for *i* = 1,2 | *f* = 679.3 | *p* = < 2.2E-16 | Reject the null hypothesis |

The p-value is < 2.2E-16 which is less than the level of significance of 5%. This indicates that this model is statistically significant and that a relationship does exist between at least one of the variables or terms. There is enough evidence to reject the null hypothesis.

In order to find which variables and/or terms in the model are significant at a level of significance of 5%, or 0.05, we will conduct individual beta tests. The null hypothesis is that the null hypothesis is that no relationship exists between the response variable and any of the predictor variables, interaction terms, or qualitative variables. The alternative hypothesis is that a relationship does exist.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| Living Area Size  (sqft\_living) |  | *t* = 15.916 | *p* = < 2E-16 | Reject the null hypothesis |
| Upper-Level Size  (sqft\_above) |  | *t* = 2.616 | *p* = 0.00894 | Reject the null hypothesis |
| Age of House  (age) |  | *t* = 12.098 | *p* = < 2E-16 | Reject the null hypothesis |
| Number of Bathrooms  (bathrooms) |  | *t* = 7.178 | *p* = 9.13E-13 | Reject the null hypothesis |
| View of Lake  (view1) |  | *t* = 15.640 | *p* = < 2E-16 | Reject the null hypothesis |
| View of Trees  (view2) |  | *t* = 20.739 | *p* = < 2E-16 | Reject the null hypothesis |

The P-values for all the predictor variables are less than the level of significance of 5%, or 0.05. Therefore, we have sufficient evidence to reject the null hypothesis and conclude that each variable is statistically significant in the model.

### Making Predictions Using Model

The predicted price for a home that has 2150ft2 living area, 1050ft2 upper level living area, is 15 years old, and backs out to a road is $459,828.20 and is expressed by the equation below:

=

The 90% prediction interval for the price of this home is (239563 – 680093.4). This indicates that there is a 90% likelihood a house with these characteristics will have a price that falls between these bounds. The confidence interval for the price of this home is (446087.9 – 473568.5). This means that we are 95% confident that the price for a house fall between this range if the sample were conducted repeatedly using these features.

The predicted price for a home that has 4250ft2 living area, 2100ft2 upper-level living area, is 5 years old, has 5 bathrooms, and backs out to a lake is $1,074,285.00 and is expressed in the equation below:

= $1,074,285.00

The 90% prediction interval for the price of this home is (852522.6 - 1296048). This indicates that the likelihood a house with these characteristics will have a price that falls between these bounds is 90%. The confidence interval for the price of this home is (1045117 - 1103454) which indicates that we are 90% confident that the price for a house fall between this range if the sample were conducted repeatedly using these features. Prediction intervals generally have larger sample sizes because they include the random variation of values as well as uncertainty in the population mean, making them wider compared to confidence intervals.

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

### Correlation Analysis

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

The first scatterplot portrays the price of a home vs. the average school rating. There is a positive correlation showing that as the school rating increases so does the price. The second scatterplot portrays the price of a home vs. the crime rate per 100,000 people. There is a negative correlation showing that as the crime rate increases, the price of a home decreases. Both charts present a slight concavity indicating that a second order model is appropriate using these variables.

### Reporting Results

The general form of the complete second order regression model:

The prediction equation of the complete second order regression model:

In this model, represents the response variable, price, represents the average school rating for the area, and represents the crime rate per 100,000 people. Once the R script has ran, we can place the beta estimates in the equation:

R-squared is the coefficient of multiple determination and its values is 0.8088 in this model. This means that roughly 80% of the variation for the response variable, price, can be explained by the predictor variables school rating and crime rate. The Adjusted R-squared value is 0.8084 in this model.

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

The first plot is The Residuals Against Fitted Values and portrays no discernable pattern which indicates homoscedasticity. The second plot is the Normal Q-Q Plot which indicates a normal distribution as the values are on or near the trend line.

### Evaluating Significance of Model

We will conduct an overall F-test in order to determine if the model is statistically significant at a 5% level of significance. The null hypothesis is that no relationship exists between the response variable, price, and the predictor variables, interaction terms, or qualitative variables. The alternative hypothesis is that a relationship does exist between the response variable and at least one of the variables or terms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | F – statistic | P-value | Conclusion |
| Price  (price) | for *i* = 1,2 | *f* = 2272 | *p* = < 2.2e-16 | Reject the null hypothesis |

Since the P-value is less than the 5% level of significance, we can reject the null hypothesis as this model is statistically significant. We will now conduct individual beta tests in order to find which variable or terms are significant and have a correlation to price.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| School Rating  (school\_rating) |  | *t* = -3.541 | *p* = 0.000406 | Reject the null hypothesis |
| Crime Rate  (crime) |  | *t* = - 6.027 | *p* = 1.90E-09 | Reject the null hypothesis |
| school\_rating:crime |  | *t* = -1.077 | *p* = 0.281513 | Fail to reject the null hypothesis |
| school\_rating2 |  | *t* = 10.497 | *p* = < 2E-16 | Reject the null hypothesis |
| crime2 |  | *t* = 8.777 | *p* = < 2E-16 | Reject the null hypothesis |

The P-value for school\_rating, crime, school\_rating2, and crime2 is less than the 5% level of significance which indicates that these variables are statistically significant and do have a correlation to the price of a home. The interaction term school\_rating:crime has a P-value greater than the 5% level of significance and is therefore not statistically significant.

### Making Predictions Using Model

The predicted value of a home in an area with an average school rating of 9.80 and a crime rate of 81.02 per 100,000 individuals is $874,497 and is expressed in the equation below:

= $874,497

The 90% prediction interval for this model is (721606.2 – 1027388). This means that there is a 90% likelihood that a house with these characteristics will have a price that falls between these bounds. The confidence interval for the price of this home is (863681.4 – 885312.7). This indicates that we are 90% confident that the price for a house fall between this range if the sample were conducted repeatedly using these features.

The predicted value of a home in an area with an average school rating of 4.28 and a crime rate of 215.50 per 100,000 individuals is $199,706.70 and is expressed in the equation below:

= $199,706.70

The 90% prediction interval for this model is (46991.65 – 352421.7) which signifies that the likelihood that the price will be between these bounds is 90%. The confidence interval is (191753.50 – 207659.9). This indicates that we are 90% confident that if the sample were conducted repeatedly with these features that the price will fall between these bounds.

## 5. Nested Models F-Test

### Reporting Results

The general form of a first order model:

The prediction equation of a first order model:

In this model, represents the response variable, price, represents the school rating, and represents the crime rate per 100,000 people. Once the R script has ran, we can add our beta estimates to the equation:

### Evaluating Significance of Model

We will now carry out an overall F-test to determine if this model is statistically significant at a 5% level of significance. The null hypothesis is that no relationship exists between the response variable, price, and the predictor variables. The alternative hypothesis is that a relationship does exist between the response variable and at least one of the predictor variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | F – statistic | P-value | Conclusion |
| Price  (price) | for *i* = 1,2 | *f* = 3573 | *p* = < 2.2E-16 | Reject the null hypothesis |

The P-value for price is lower than the level of significance of 5% which indicates that this model is statistically significant. We have sufficient evidence to reject the null hypothesis. We will now conduct individual beta tests to determine which variables are statistically significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| School Rating  (school\_rating) |  | *t* = 49.65 | *p* = < 2.2E-16 | Reject the null hypothesis |
| Crime Rate  (crime) |  | *t* = 17.20 | *p* = < 2.2E-16 | Reject the null hypothesis |
| school\_rating:crime |  | *t* = -31.63 | *p* = < 2.2E-16 | Reject the null hypothesis |

All the variables in this model have a P-value that is less than the 5% level of significance which indicates that the variables school\_rating, crime, and the interaction term school\_rating:crime have a correlation to the price of a home.

### Model Comparison

In general, a complete model is the full version of a model that contains all the possible terms including interaction terms and a reduced model is a smaller version in which the terms are a subset of the complete model. A reduced model may be used when only statistically significant terms are needed.

The general form of the reduced model:

The prediction equation for the reduced model:

The general form for the complete model:

The prediction equation for the complete model:

We will now conduct the nested model F-test to evaluate if the quadratic terms are needed. We need to determine if the first order terms are sufficient or if we need to use the quadratic terms from the complete model in order to predict the price of a home. The null hypothesis is that we do not need the quadratic terms and that the reduced model is sufficient. The alternative hypothesis is that we do need to use the quadratic terms and that we need to use the complete model.

|  |  |  |  |
| --- | --- | --- | --- |
| Hypothesis | F – statistic | P – value | Conclusion |
| for *i* = 4,5 | *f* = 65.20513 | *p* = 2.22716E-28 | Reject the null hypothesis |

The P-value is less than the 5% level of significance which indicates that we need to reject the null hypothesis as at least one of the quadratic terms are statistically significant. This means that we need to use the complete model instead of the reduced model.

## 6. Conclusion

In conclusion, different models were created to find how certain variables correlate to the price of a home. Model 1 used first order regression with quantitative and qualitative variables. The quantitative variables used include: the number of bathrooms (*bathrooms*), area of the living space (*sqft\_living*), area of the upper living space (*sqft\_upper*), and the age of the home (*age*). The qualitative variable used in this model was the view which can back to a road, trees, or a lake. We conducted an overall F-test to determine if the model was statistically significant and found that it was. Next, we conducted individual beta tests to find which variables were statistically significant. All the variables tested had a P-value less than the 5% level of significance which indicates that they each are statistically significant and have a correlation to the price of a home.

Model 2 used complete second order regression with quantitative variables. This model used the quantitative variables school rating (*school\_rating*) and the crime rate per 100,000 people (*crime*). The complete model uses the interaction term between these variables as well as the quadratic or squared version of the variable. An overall F-test was conducted to find if this model was statistically significant, and we found that it was. Individual beta tests were conducted to find which variables and/or terms were statistically significant. We found that all the variables tested except for the interaction term, school\_rating:crime, are statistically significant and have a correlation to the price of a home.

We then conducted an F-test using nested models to find if the complete model, (Model 2 in this project), was needed or if the reduced model was sufficient. The reduced model does not contain quadratic variables. Once the test was ran, the P-value was less than the 5% level of significance which indicates that we the quadratic variables are statistically significant and that the complete model was needed over the reduced model.

Overall, I would use Model 2 as the quadratic terms are statistically significant and can assist in providing insight when predicting the price of a home. I would also add some variables used in the first model such as the age of the home, size of the living space, and view as I feel these variables used in the complete model can further assist in accurately predicting the price a home. The analyses that were performed can serve as a guide for a real estate company to determine the worth of a home or for potential buyers to ensure they are getting a fair deal on a home and not overspending.

## 7. Citations

Berrier, J. (2016). MAT 303: Applied Statistics 2 for Science. Zyante Inc. (zyBooks.com)