# MAT 303 Project One Summary Report

Jarrale Butts

Jbutts1234@gmail.com

Southern New Hampshire University

## 1. Introduction

The dataset being explored pertains to historical data that can be used to analyze relationships between different attributes of a house (such as square footage or the number of bathrooms) and the house’s selling price. My results may be used in making predictions or calculating house price fluctuation based on the characteristics of the house. The analysis I will be running in this problem set involve first and second order regression models with both quantitative and qualitative variables, as well as a nested F-model.

## 2. Data Preparation

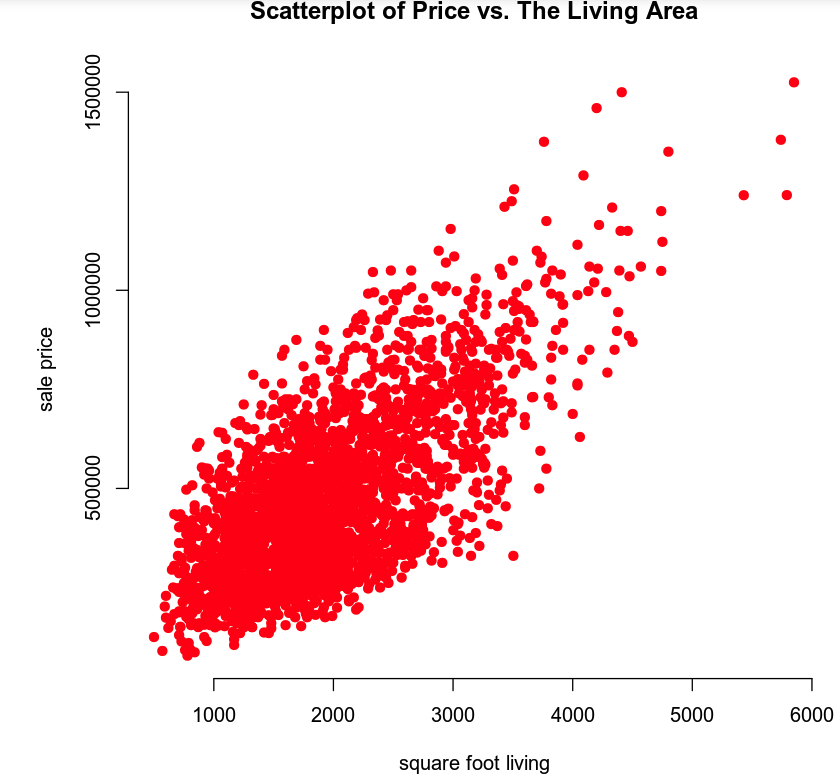
There are many variables included in the large set of historical data I will be analyzing. Amongst those variables the most important in which I will be working with include sale price of a home (price), number of bathrooms (bathrooms), size of the living area in sqft (sqft\_living), size of the upper level in sqft (sqft\_above), age of the home (age), crime rate per 100,000 people (crime), average rating of schools in the area (school\_rating), and different types of views (view).

The dataset also contains 23 columns and 2692 rows.

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

### Correlation Analysis

A couple of trends can be seen in the two scatterplots below. Starting with the scatterplot to the left (Price vs The Living Area) shows a clear and strong linear and positive correlation between the sale price of a home and the square foot associated with that home. On the right scatterplot (Price vs Age) it on the other hand shows a nonlinear and no correlation between the sale price of a home and the square foot associated with that home. Both scatter plots do possess a few outliers in the top region of the plot.

A picture containing text, screenshot, diagram

Description automatically generated

Below I have included a correlation table for a better visual representation of both the strength and direction for each correlation coefficient. Price and size of the living area in sqft have a positive moderate correlation of roughly 0.7. While Price and age of the home have a negative very weak correlation of nearly 0 at -0.07.

|  |  |
| --- | --- |
| Correlation Table | |
|  | price |
| price | 1.0000 |
| sqft\_living | 0.6895 |
| age | -0.0746 |

### Reporting Results

The general form equation of the multiple regression for model1 using price as the response variable and living area, upper level area, age of the home, number of bathrooms, and view as predictor variables is: . The prediction equation for model1 is: .

The multiple regression model for model1 is : lm(price ~ sqft\_living + sqft\_above + age + bathrooms + view). The prediction model equation for medel1 using outputs obtained from my R script is: .

The values of R-squared and adjusted R-squared for model1 are 0.603 and 0.602. Both the R-squared and adjusted R-squared values are at 60%. This means that 60% of how price fluctuates in model1 can be attributed linearly to the predictor variables living area, upper level area, age of the home, number of bathrooms, and view. Even though there are several predictors in model1, the adjusted R-square hardly penalizes the R-squared value.

In model1, size of the living area in sqft beta estimate of 129.3 suggests that on average, for every one unit increase in the sqft\_living variable the price is estimated to increase by $129.3, with the assumption that all other variables are held constant. A one unit increase in the variable view2 (lake view) suggest the price is estimated to increase by $249,000, with the assumption that all other variables are held constant.

Regarding the plots below a few assumptions can be made. Due to a lack of rising or falling patterns as well as no strong linear pattern the mean of zero assumption holds. The constant variance assumption does appear to hold due to a lack of clear change of variability. The left side of the residuals against fitted plot is clustered with many data points, but it doesn’t appear to change in variability. The independence assumption may not hold due to the leftward cluster in the residuals against fitted plot. Lastly, though it's not a strong argument looking at the Q-Q plot, the normality assumption does hold with enough residuals lying reasonably close to the diagonal line.

A picture containing text, screenshot, colorfulness

Description automatically generatedA picture containing text, line, plot, diagram

Description automatically generated

### Evaluating Significance of Model

The overall F-test for model1 showed that the model is significant at a 5% level of significance. The F-statistic has a value of 679.3 and the P-value of model1 is . The null hypothesis can be identified as , and the alternative hypothesis can be identified as (. Due to the P-value being less that 0.05 the null hypothesis can be rejected in favor of the alternative hypothesis.

As depicted in the table below showing the results of the individual beta test, I can conclude that all terms in the model are significant at a 5% level of significance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Individual Beta Test | | | | |
| Variable | Hypotheses | Test Statistic | P-value | Conclusion |
| Sqft\_living |  | t = 15.916 | p < 0.001 | Reject the null |
| Sqft\_above |  | t = 2.616 | p = 0.008 | Reject the null |
| age |  | t = 12.098 | p < 0.001 | Reject the null |
| bathrooms |  | t = 7.178 | p < 0.001 | Reject the null |
| view1 |  | t = 15.640 | p < 0.001 | Reject the null |
| view2 |  | t = 20.739 | p < 0.001 | Reject the null |

### Making Predictions Using Model

The predicted price 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 is: $459,828.2 The 90% prediction and confidence intervals for the price of this home are: $239,563 – $680,093.4 and $446,087.9 - $473,568.5 on the low and high end respectively. Given the prediction interval we can be 90% certain the actual price of a home with the given variables (characteristics) will fall in the price range of ($239,563 – $680,093.4). Alternatively, with the given confidence interval we can be 90% certain the average price of a home with the given variables (characteristics) will fall in the price range of ($446,087.9 - $473,568.5).

The predicted price 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 is: $1,074,285. The 90% prediction and confidence intervals for the price of this home are: $852,522.6 – $1,296,048 and $1,045,117 – $1,103,454 also on the low and high end. Similarly, to the explanation given previously, given the prediction interval we can be 90% certain the actual price of a home with the given variables (characteristics) will fall in the price range of ($852,522.6 – $1,296,048). Alternatively, with the given confidence interval we can be 90% certain the average price of a home with the given variables (characteristics) will fall in the price range of ($1,045,117 – $1,103,454).

The prediction interval is wider than the confidence interval because the prediction interval is specific to an individual’s home with the given variables (characteristics). The confidence interval on the other hand pertains to the average price of homes with similar variables (characteristics) in the population.

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

### Correlation Analysis

Both scatterplots below (Price and Average School Rating and Price and Crime Rate) show a non-linear relationship. Price and Average School Rating shows a positive non-linear trend, while Price and Crime Rate shows a negative non-linear trend. A second order model is appropriate using these variables because of the lack of linearity between the relationships of the response and predictor variables.

A picture containing text, screenshot, plot, diagram

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated

### Reporting Results

The general form of a complete second order model for model2 for price using average school rating in the area and crime rate per 100,000 people as predictors is: . The prediction equation for model2 is: .

The complete second order model for model2 is: lm(price ~ school\_rating + crime + school\_rating:crime + I() + I(). The prediction model equation for medel2 using outputs obtained from my R script is:

The values of R-squared and adjusted R-squared for model2 are 0.809 and 0.808. Both the R-squared and adjusted R-squared values of model2 are nearly 81%. The R-squared value is the percentage of the variation in price that can be attributed to the average rating of schools in the area and crime rate per 100,000 people, as well as their interaction term and squared terms. In this instance the adjusted R-square hardly penalizes the R-squared value.

Regarding the plots below a few assumptions can be made. Similarly, to the results in model1, due to a lack of rising or falling patterns as well as no strong linear pattern the mean of zero assumption holds. The constant variance assumption does appear to hold due to a lack of clear change of variability. The independence assumption seems to hold due to a lack of a strong relationship between one residual and the next. Lastly, looking at the Q-Q plot, the normality assumption does not hold. Though most points lie reasonably close to the diagonal line, forwards the end they trail off, and if given more data points I believe that trail off would continue.

A picture containing text, screenshot, red

Description automatically generatedA picture containing text, screenshot, line, plot

Description automatically generated

### Evaluating Significance of Model

The overall F-test for model2 showed that the model is significant at a 5% level of significance. The F-statistic has a value of 2272 and the P-value of model2 is . The null hypothesis can be identified as , and the alternative hypothesis can be identified as (. Due to the P-value being less that 0.05 the null hypothesis can be rejected in favor of the alternative hypothesis.

As depicted in the table below showing the results of the individual beta test, I can conclude that all terms in model2 except the interaction term (school\_rating \* crime) are significant at a 5% level of significance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Individual Beta Test | | | | |
| Variable | Hypotheses | Test Statistic | P-value | Conclusion |
| school\_rating |  | t = -3.541 | p < 0.001 | Reject the null |
| crime |  | t = -6.027 | p < 0.001 | Reject the null |
| school\_rating \* crime |  | t = -1.077 | p = 0.281 | Fail to reject |
|  |  | t = 10.497 | p < 0.001 | Reject the null |
|  |  | t = 8.777 | p < 0.001 | Reject the null |

### Making Predictions Using Model

The predicted price for a home in an area with average school rating of 9.80 and a crime rate of 81.02 per 100,000 individuals is: $874,497. The 90% prediction and confidence intervals for the price of this home are: $721,606.2 – $1,027,388 and $863,681.4 - $885,312.7 on the low and high end respectively. Given the prediction interval we can be 90% certain the actual price of a home with the given variables (characteristics) will fall in the price range of ($721,606.2 – $1,027,388). Alternatively, with the given confidence interval we can be 90% certain the average price of a home with the given variables (characteristics) will fall in the price range of ($863,681.4 - $885,312.7).

The predicted price for a home in an area with average school rating of 4.28 and a crime rate of 215.50 per 100,000 individuals is: $199,706.7. The 90% prediction and confidence intervals for the price of this home are: $46,991.65 – $352,421.7 and $191,753.5 – $207,659.9 also on the low and high end. Similarly, to the explanation given previously, given the prediction interval we can be 90% certain the actual price of a home with the given variables (characteristics) will fall in the price range of ($46,991.65 – $352,421.7). Alternatively, with the given confidence interval we can be 90% certain the average price of a home with the given variables (characteristics) will fall in the price range of ($191,753.5 – $207,659.9).

## 5. Nested Models F-Test

### Reporting Results

The general form of a first order model for model3 for price using average school rating in the area and crime rate per 100,000 people as predictors, including the interaction term between average school rating and crime rate is: . The prediction equation for model3 is: .

The first order regression model for model3 is: lm(price ~ school\_rating + crime + school\_rating:crime. The prediction model equation for medel3 using outputs obtained from my R script is:

### Evaluating Significance of Model

The overall F-test for model3 showed that the model is significant at a 5% level of significance. The F-statistic has a value of 3573 and the P-value of model3 is . The null hypothesis can be identified as , and the alternative hypothesis can be identified as (. Due to the P-value being less that 0.05 the null hypothesis can be rejected in favor of the alternative hypothesis.

As depicted in the table below showing the results of the individual beta test, I can conclude that all terms in model2 except the interaction term (school\_rating \* crime) are significant at a 5% level of significance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Individual Beta Test | | | | |
| Variable | Hypotheses | Test Statistic | P-value | Conclusion |
| school\_rating |  | t = 49.65 | p < 0.001 | Reject the null |
| crime |  | t = 17.20 | p < 0.001 | Reject the null |
| school\_rating \* crime |  | t = -31.63 | p < 0.001 | Reject the null |

### Model Comparison

When comparing two models a complete model is a regression model with all necessary and testable predictors included. A reduced model is a smaller subset of the complete model with one or more fewer predictors.

The general form of the reduced model is: . The prediction equation for the reduced model is: .

The general form of the complete model is: . The prediction equation for the complete model is: .

The nested model F-test at a 5% level of significance shows that at least one of the quadratics (squared) terms are needed. The model F-test has a P-value of . The null hypothesis can be identified as , and the alternative hypothesis can be identified as (. Due to the P-value being less that 0.05 the null hypothesis can be rejected in favor of the alternative hypothesis.

## 6. Conclusion

Based on the analysis that I have performed the model I would choose to predict house prices would be model2 having the highest adjusted R-squared value of nearly 81%. This indicates the percentage of the variation in price that can be attributed to the average rating of schools in the area and crime rate per 100,000 people. In model2 and crime are used as predictors. The overall F-test of model2 also showed that the model is significant at a 5% level of significance while have a P-value of . Price and Average School Rating and Price and Crime Rate scatterplots show a non-linear relationship. Price and Average School Rating shows a positive non-linear trend, while Price and Crime Rate shows a negative non-linear trend. Regarding the residual against fitted and normal Q-Q plots the mean of zero assumption holds, the constant variance assumption holds, the independence assumption holds, though the normality assumption does not hold.

The practical importance of the analyses that were performed is that they offer insights into factors that impact sale price of a home based on characteristics of that home. The prediction and confidence intervals provide information about the range of expected sale price of a home for individuals and populations with could be valuable in making decisions in the home buying market.