# MAT 303 Project One Summary Report

Alyssa Basile

alyssa.basile@snhu.edu

Southern New Hampshire University

## 1. Introduction

The data set I will be exploring in this project is housing.csv. I will be analyzing the relationships between different variables of a house for a real estate company. The point of this analysis is to predict sale prices for houses based on critical variable factors and the goal is to set better prices when listing a home for a client to ensure the listing can be sold within a reasonable amount of time. I will be running several different analyses in this project:

* First Order Regression Model with Quantitative and Qualitative Variables
* Complete Second Order Regression Model with Quantitative Variables
* First Order Regression Model with Quantitative Variables
* Scatterplots
* Residuals against Fitted Values
* Normal Q-Q Plot
* Prediction and Confidence Intervals
* Nested Models F-Test

## 2. Data Preparation

There are several important variables in this data set, and they are:

* Price (price)
* Number of bathrooms (bathroom)
* Size of the living area in sqft (sqft\_living)
* Age of the home (age)
* Measure of craftsmanship and the quality of materials used to build the home (grade)
* Average age of all appliances in the home (appliance\_age)
* Crime rate per 100,000 people (crime)
* Home backs out to a lake (view = “2”), backs out to trees (view = “1”), or backs out to a road (view = “0”)

There are 22 columns and 2,692 rows in this data set.

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

### Correlation Analysis

### Chart, scatter chart Description automatically generatedChart, scatter chart Description automatically generated

In the scatterplot of price against the size of the living area, there is a linear relationship. The bigger the living area is, the higher the price will be.

In the scatterplot of price against the age of the home, there does not appear to be any correlation between the price and the age of the home. There is no direction, it is all over the place. The image of pearson correlation matrix table is below to demonstrate the strength and direction of the correlation.

Table

Description automatically generated

According to the pearson correlation, coefficient between price (price) and the size of the living area (sqft\_living), it is 0.6895. The strength of correlation is moderate (because it is between 0.40 < |R| < 0.80, it is considered moderate). Because it is positive, this is a positive correlation. The correlation between price (price) and age of the home (age), it is -0.0746. The strength of correlation is weak (because it is between 0 < |R| < 0.40, it is considered weak). Because it is negative, this is a negative correlation. I have also included scatterplot graphs above to show and prove my answers.

### Reporting Results

*Table

Description automatically generated with medium confidence*

The general form of the second-order regression model with three quantitative variables and 1 qualitative variable is above where y is price, is sqft\_living, is grade, is bathrooms, (view1) and (view2) are the dummy variables for view.

The form above is the multiple regression model equation with beta estimates. For every sqft of the area, the living area beta estimate will increase by 94.84. The home that backs out to a lake view is view2. Which would mean we add 228,700 and we will not add 165,700. This is because will be 1 and you multiply that with 228,700. will be 0, when you multiply that with 165,700, we will get 0.

The value of (R-squared) is 0.6475 and the value of (Adjusted R-squared) is 0.6469. Using the predictor of the size of living area (sqft\_living), measure of craftsmanship and the quality of materials used to build the home (grade), number of bathrooms (bathrooms), and the view (view) for this model, this data can explain 64.75% of the variation in price. Which means it explains a little over half of the variation in price (the goal is to have it be as close to 100% as possible). According to the , this model is kind of complicated since the percentage is low (64.69%).

*Chart, scatter chart

Description automatically generated*Chart, line chart

Description automatically generated

In the scatterplot of residuals against fitted values, the graph appears that the dots are clustered on the left side of the graph. Which would mean constant variance assumption may be violated. There is no evidence of non-linearity. The Q-Q plot showed normally distributed residuals. Majority of the dots are on the straight line. Because of the possibility constant variance assumption may be violated, this model does not appear to be valid.

### Evaluating Significance of Model

Overall F-test:

Null Hypothesis:

Alternative Hypothesis:

Overall F-statistic is F = 986.9 with a corresponding p-value of < 2.2 \* . Since the p-value is close to 0, sufficient evidence exists to reject the null hypothesis, which suggests that at least one of the predictors is linearly related to y.

Individual t-test:

Null Hypothesis: for some I = 1, …, n

Alternative Hypothesis:

Living Area predictor:

The t-statistic for living area estimate is 16.986 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and living area are linearly related.

Grade predictor:

The t-statistic for grade estimate is 22.545 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and grade are linearly related.

Number of Bathrooms predictor:

The t-statistic for number of bathrooms estimate is -4.238 with a corresponding p-value of 2.33 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and the number of bathrooms are linearly related.

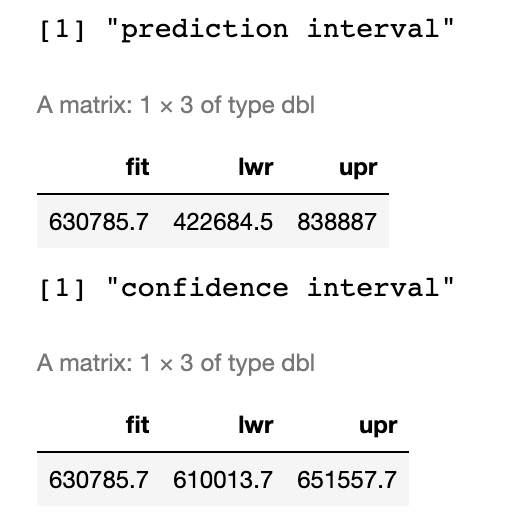
View1 (trees) predictor:

The t-statistic for view1 estimate is 16.559 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and view1 are linearly related.

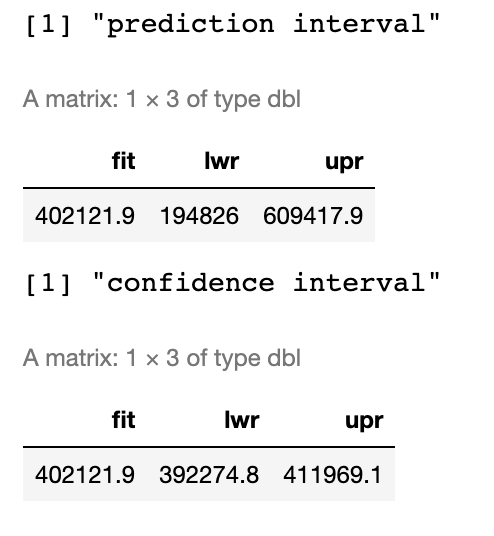
View2 (lake) predictor:

The t-statistic for view2 estimate is 20.170 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and view2 are linearly related.

### Making Predictions Using Model



The 90% prediction and confidence interval above is for the predicted price for a home that backs out to a lake and has a 2,150 sq ft living area, 7 grade, and three bathrooms. The predicted price is $630,785.70. The 90% prediction interval for the price identified is between $422,684.50 and $838,887. The 90% confidence interval for the price identified is between $610,013.70 and $651,557.70.



The 90% prediction and confidence interval above is for the predicted price for a home that backs out to a road and has a 2,150 sq ft living area, 7 grade, and three bathrooms. The predicted price is $402,121.90. The 90% prediction interval for the price identified is between $194,826 and $609,417.90. The 90% confidence interval for the price identified is between $392,274.80 and $411,969.10.

The reason why the prediction interval is wider than the confidence interval is because for one, prediction interval is best used for one sample and making it wider gives us some more room to make an estimate (added uncertainty). As for confidence interval, it is typically used for multiple samples.

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

### Correlation Analysis

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

In the scatterplot of price against the age of appliances, there is a quadratic relationship (the line is curved). The older the age of appliances are, the lower the cost will be. This also applies to the scatterplot of price against the crime rate per 100,000 people. The higher the rate is, the lower the cost will be. Because this is a quadratic relationship, second order model will be appropriate using these variables. Pearson correlation matrix table is shown below.

Table

Description automatically generated

According to the pearson correlation, coefficient between price (price) and the age of appliances, it is -07860. The strength of correlation is moderate (because it is between 0.40 < |R| < 0.80, it is considered moderate). Because it is negative, this is a negative correlation. Between the price and the crime rate, the strength of correlation is moderate (because it is between 0.40 < |R| < 0.80, it is considered moderate). Because it is negative, this is a negative correlation. This matches to the scatterplots above.

### Reporting Results

Text

Description automatically generated

The general form of the second-order regression model with two quantitative variables, 1 interaction term, and 2 squared terms is above where y is price, is appliance\_age, and is crime.

The form above is the second-order regression model equation with beta estimates.

The value of (R-squared) is 0.8088 and the value of (Adjusted R-squared) is 0.8084. Using the predictor of the age of appliances, crime rate per 100,000 people, interaction term of age of appliances and crime rate, and squared terms of both age of appliances and crime rate for this model, this data can explain 80.88% of the variation in price. The goal is to have the value be as close to 100% as possible. Which would mean this model is pretty accurate. This model appears to be more accurate than the first regression model.

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

In the scatterplot of residuals against fitted values, the graph appears to be random. While the dots are clustered on the left side of the graph, I do not see much of a pattern. There is no evidence of non-linearity. The Q-Q plot showed normally distributed residuals. Majority of the dots are on the straight line before veering off the line. All assumptions seem to hold so this model appears to be valid.

### Evaluating Significance of Model

Overall F-test:

Null Hypothesis:

Alternative Hypothesis:

Overall F-statistic is F = 2272 with a corresponding p-value of < 2.2 \* . Since the p-value is close to 0, sufficient evidence exists to reject the null hypothesis, which suggests that at least one of the predictors is linearly related to y.

Individual t-test:

Null Hypothesis: for some I = 1, …, n

Alternative Hypothesis:

Age of Appliances predictor:

The t-statistic for age of appliances estimate is -30.951 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the age of appliances and the price are linearly related.

Crime Rate predictor:

The t-statistic for crime rate estimate is -24.827 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and crime rate are linearly related.

Interaction term: Age of Appliances and Crime Rate predictor:

The t-statistic for the interaction term estimate is 1.072 with a corresponding p-value of 0.234. Since this p-value is greater than the significance level of 5% (0.05), we have insufficient evidence to reject the null hypothesis. Which suggests the price and interaction terms are not linearly related.

predictor:

The t-statistic for estimate is 10.501 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and are linearly related.

predictor:

The t-statistic for estimate is 8.782 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and are linearly related.

### Making Predictions Using Model

Table

Description automatically generated

The 90% prediction and confidence interval above is for the predicted price for a home that has one-year-old appliances and is in an area that has acrime rate of 81.02 per 100,000 individuals. The predicted price is $864,423.40. The 90% prediction interval for the price identified is between $711,566.60 and $1,017,280. The 90% confidence interval for the price identified is between $854,109.10 and $874,737.70.

Table

Description automatically generated

The 90% prediction and confidence interval above is for the predicted price for a home that has 15-year-old appliances and is in an area that has acrime rate of 200.50 per 100,000 individuals. The predicted price is $271,051.60. The 90% prediction interval for the price identified is between $118454.40 and $423,648.80. The 90% confidence interval for the price identified is between $265,846 and $276,257.20

## 5. Nested Models F-Test

### Reporting Results

Text

Description automatically generated

The general form of the first-order regression model with two quantitative variables and 1 interaction term is above where y is price, is appliance\_age, and is crime.

The value of (R-squared) is 0.7995 and the value of (Adjusted R-squared) is 0.7993. Using the predictor of the age of appliances, crime rate per 100,000 people, interaction term of age of appliances and crime rate for this model, this data can explain 79.95% of the variation in price.

### Evaluating Significance of Model

Overall F-test:

Null Hypothesis:

Alternative Hypothesis:

Overall F-statistic is F = 3573 with a corresponding p-value of < 2.2 \* . Since the p-value is close to 0, sufficient evidence exists to reject the null hypothesis, which suggests that at least one of the predictors is linearly related to y.

Individual t-test:

Null Hypothesis: for some I = 1, …, n

Alternative Hypothesis:

Age of Appliances predictor:

The t-statistic for age of appliances estimate is -49.65 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the age of appliances and the price are linearly related.

Crime Rate predictor:

The t-statistic for crime rate estimate is -48.17 with a corresponding p-value of < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and crime rate are linearly related.

Interaction term: Age of Appliances and Crime Rate predictor:

The t-statistic for the interaction term estimate is 31.63 with a corresponding p-value of 0.234. < 2 \* . Since this p-value is close to 0 and is less than the significance level of 5%, we have sufficient evidence to reject the null hypothesis. Which suggests the price and interaction terms are linearly related.

### Model Comparison

Table

Description automatically generated

Complete model:

Reduced model:

Null Hypothesis: (The reduced model is sufficient)

Alternative Hypothesis: (The coefficient for the interaction term is non-zero. The complete model should be used.)

Let us use a level of significance of 5%. If the p-value is less than 0.05, we will reject the null hypothesis. Otherwise, we will not reject the null hypothesis.

The F-value = 65.26001 and the p-value = 2.113594e-28.

Since the p-value 2.113594e-28 is less than the level of significance 0.05, we will reject the null hypothesis and conclude that the squared terms should be used in predicting the price. Therefore, the complete model should be used.

## 6. Conclusion

To predict house prices, I would choose the complete second order regression model with quantitative variables. We can confirm this in 2 ways.

First way:

Looking at the value of (R-squared) and the value of (Adjusted R-squared) for all 3 models. The first model (first order regression model with quantitative and qualitative variables), it only explains 64.75% of the variation in price. It is the lowest percentage of all, and we want the percentage to be as close as 100% to demonstrate how accurate the model is. In addition to it, is low as well (64.69%). Which means this model has too many variables used, and it is complicated. The second model (complete model), which is the model I would use to predict house prices, it explains 80.88% of the variation in price. Out of all 3 models, it is the closest to 100%. Lastly, the first order regression model with two quantitative variables and one interaction term (reduced model), it explains 79.95% of the variation in price. It is slightly lower than the second model.

Second way:

Performing the F-test for nested models confirmed which model to use. We were able to reject the null hypothesis and conclude the squared terms should be used in predicting the price. In doing so, it had meant complete model should be used. If we were unable to reject the null hypothesis, it would have meant the reduced model would have been the better model to use.

Without these analyses, we would have not been able to find the relationship between the price and other variables. It also helped us confirm which model is the best to use in predicting the house prices.