# MAT 303 Project One Summary Report

Zachary Tankersley

Zachary.tankersley@snhu.edu

Southern New Hampshire University

## 1. Introduction

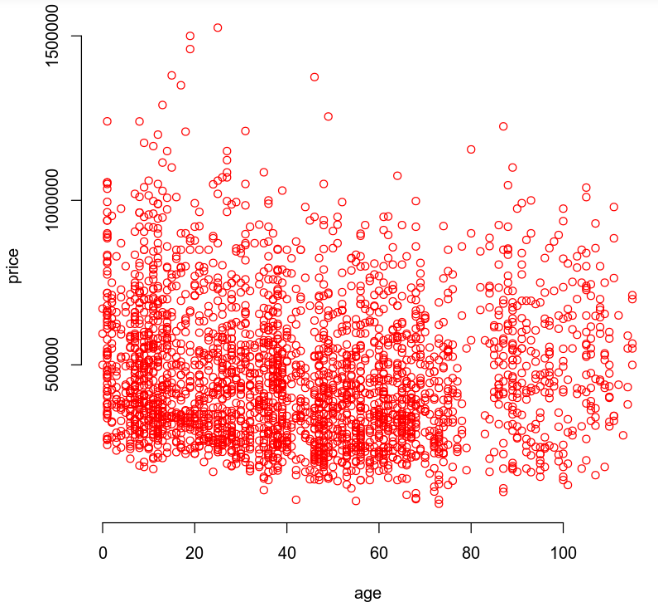
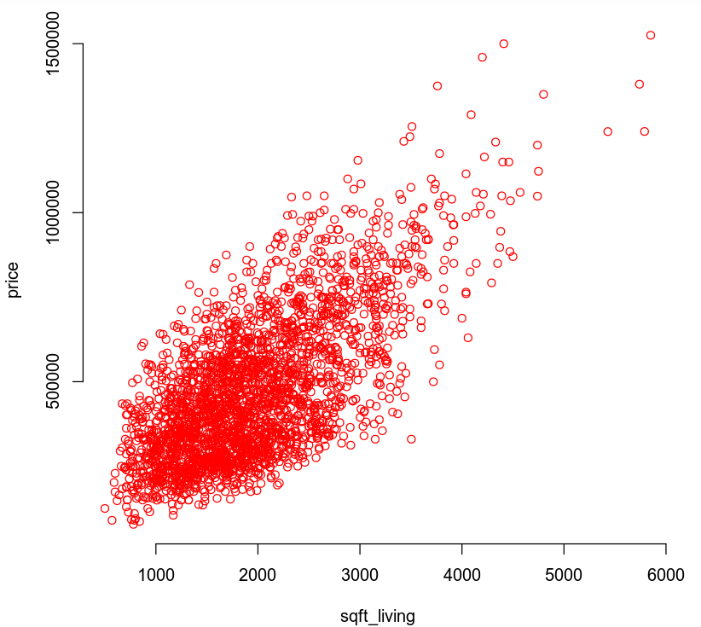
To leverage a historical data set containing information on several housing attributes, I will be building several predictive models that will be used to predict a home’s projected selling price more accurately. The data set contains information on 2,692 homes and has variables such as the number of bedrooms and bathrooms, square footage, and condition. The aim is to help the company set better prices when listing homes for our clients. A first-order regression model with both quantitative and qualitative variables, a second-order multiple regression model with quantitative variables, and a nested model F-test will be created.

## 2. Data Preparation

When building the different models, I will use several variables to predict price, including living area (sqft\_living), grade of the home (grade), number of bathrooms (bathrooms), view (view) age of the home (age), age of the appliances (appliance\_age), and the crime rate per 100,000 people (crime). There are 2,692 rows of data (2,693 if you include the header row with variable names), and 22 columns containing individual variable information.

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

### Correlation Analysis

****

For a visual comparison, two scatterplots have been created, one showing price vs living area and one showing price vs age. By inspecting visual trends, we can see that living area is likely correlated with price, *positively* meaning that as the square footage of living space increases, so too does the price. However, looking at the price vs age scatterplot, we see no clear positive or negative linear relationship between price and age. Real-world knowledge can also be applied here, a 100-year-old house may not necessarily be worth a million dollars, quite the opposite. The correlation coefficients for price vs square feet and price vs age are 0.6895 and -0.0746 respectively. The coefficient for age is negative hinting at a negative relationship between age and price.

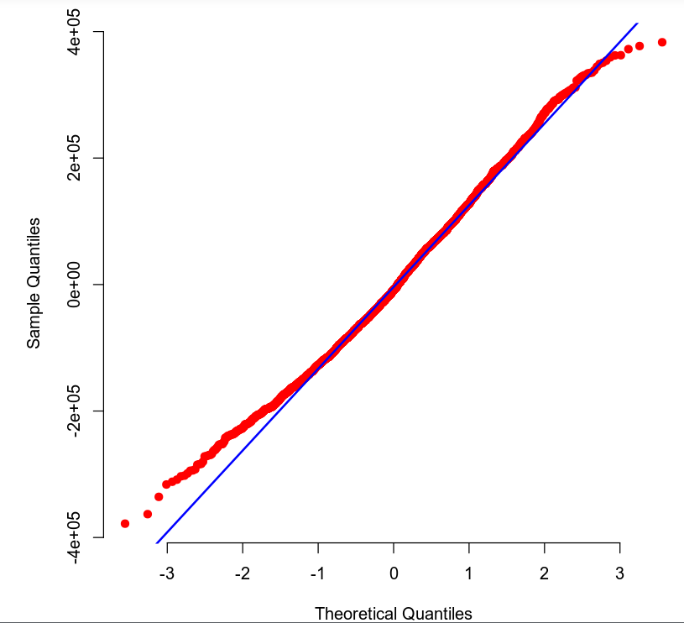
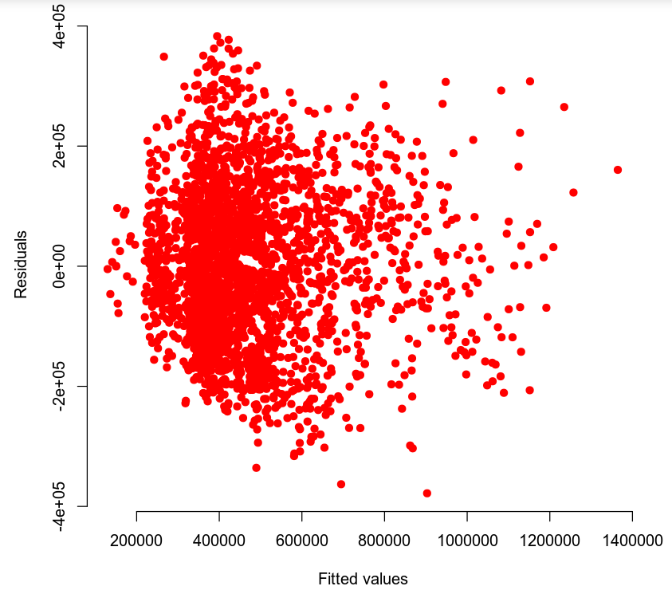
### Reporting Results

To represent a first-order regression model using price as the response variable, living area, grade, number of bathrooms, and view as predictor variables and in its general form, we can use:

Using a summary analysis function in R, we are returned with the beta estimates and can build our model in its true form as follows:

**+165700(view1) +228700(view2)**

We can see that the R2 is 0.6475 and the Adjusted R-Squared is 0.6469 meaning that only about 64% of the variance in price is explained by this model. The beta estimates for living area and lake view are 93.84x and 228700x respectively, meaning that the price increases by 93.84 for every unit increase in square footage of the living area, and that if the property has a lake view the price will increase by an additional 228,700. Below you can view the scatterplot for residuals vs fitted values and the normal Q-Q plot:



The scatterplot seems to confirm homoscedasticity, no pattern is apparent. And the Q-Q line in the Q-Q plot seems to have little or no variance until you get well into the 2nd and 3rd quantiles.

### Evaluating Significance of Model

For evaluating our model significance against a 5% level of significance, we need only look at the summary analysis function results in R where we can see that the P-value is 2.2e-16 which is less than the 0.05 threshold. The null and alternative hypotheses are:

This means that we can reject the null hypothesis in favor of the alternative by saying a statistically significant relationship does exist between price and at least one predictor variable. To determine which individual variables are significant, individual beta tests can be performed and compared to the null and alternative hypotheses:

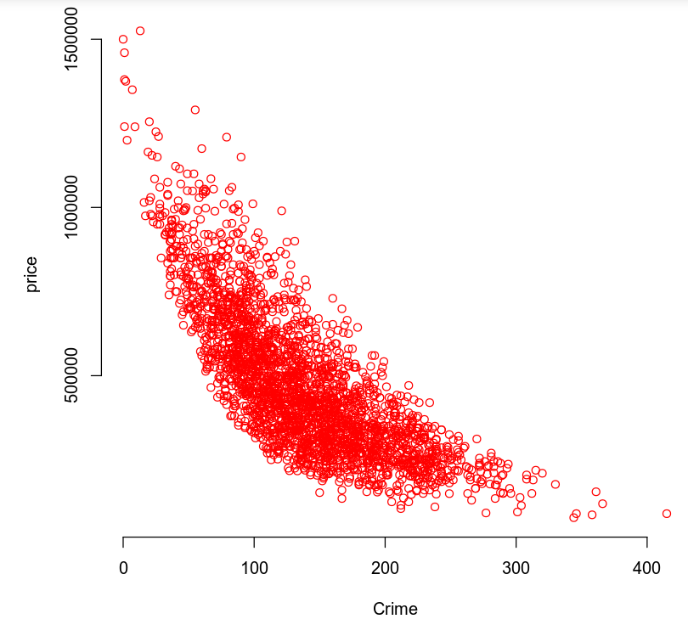
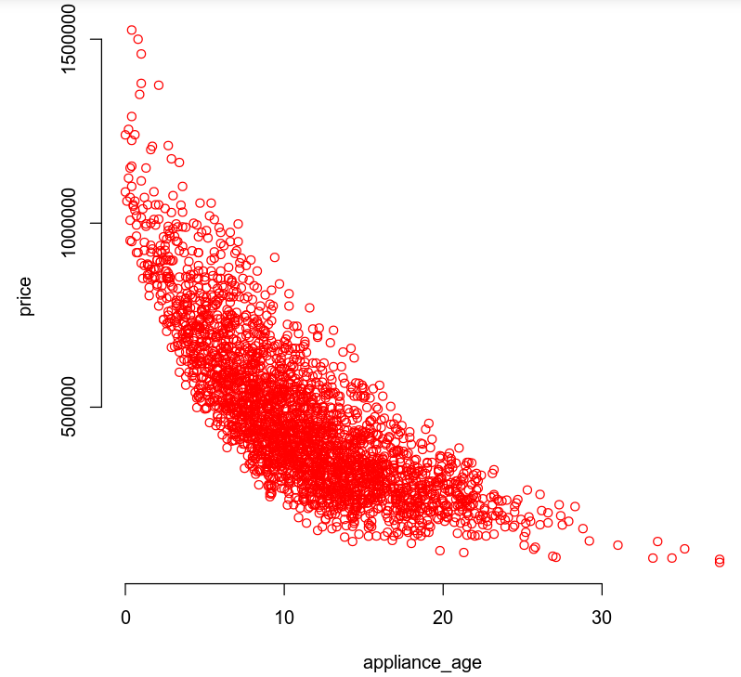
The beta value of the bathrooms variable is 2.33e-05 while all others are 2e-16 which are all well under the 5% significance threshold, meaning in this model, each predictor variable *does* have a statistically significant relationship with price and can be included.

### Making Predictions Using Model

Using the previously supplied model, we can predict a home price of $630,785.7 for a home that backs out to a lake and has a 2,150 sqft living area, a grade of 7, and three bathrooms. The 90% prediction interval and confidence intervals are [422684.5, 838887] and [610013.7, 651557.7] respectively. This means that 90% of predicted values should fall within the prediction interval range and that the average of those predicted values should fall within the confidence interval range 90% of the time. For the same house backing out to a road, the predicted price would be $402,121.90 with a confidence and prediction interval of [194826, 609417.9] and [392274.8, 411969.1] respectively. In all cases, the prediction interval predicts a range of values based on the model, the confidence interval is predicting the mean of that range of values and therefore will be narrower.

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

### Correlation Analysis

****

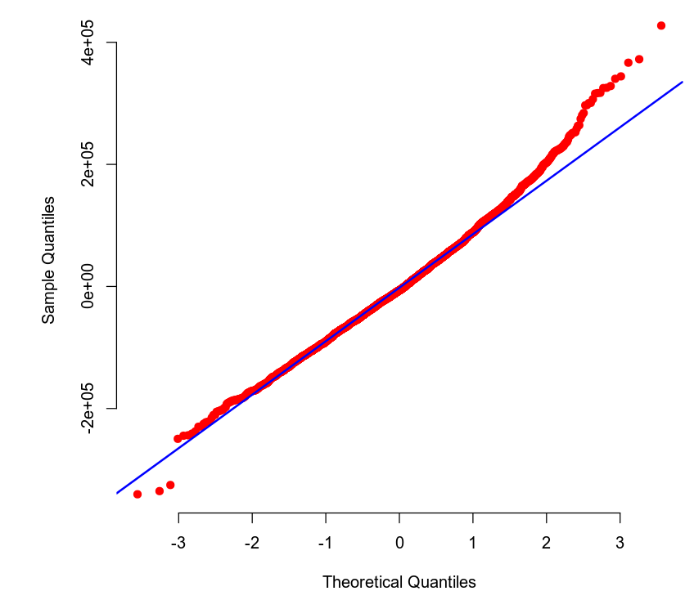
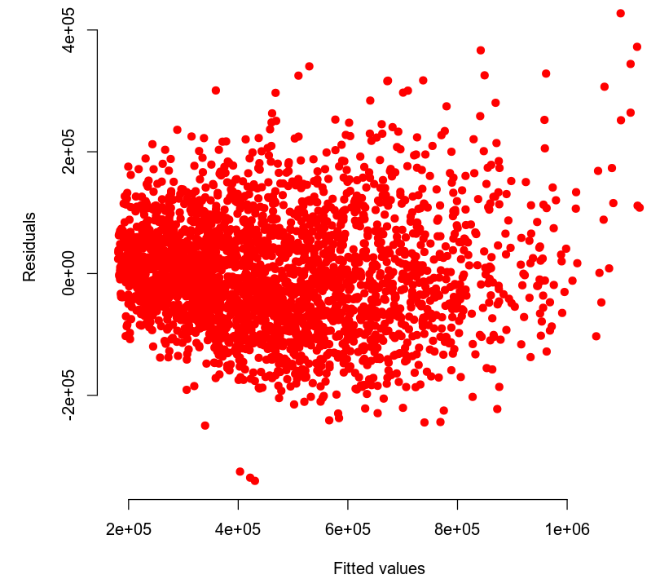
Here I have created two scatter plots showing price vs appliance age and price vs crime per 100,000 people. A single, clearly downward concavity exists in both indicating a need for a second order regression term with a negative relationship to price.

### Reporting Results

To represent a complete second-order regression model using price as the response variable and age of appliances and crime rate per 100,000 people as predictor variables, in its general form I will use:

To show the model in its true form I ran a summary analysis function on the data with this new model and came up with the following:

Using the results of the summary analysis function I received an of 0.8088 and an value of 0.8084 meaning that about 81% of the variance in price can be explained by this model. By creating scatter plots of the residuals vs the fitted values and a Q-Q plot, the homoscedasticity and normality of predicted values can be visualized as below:



The data appears to have no discernable patterns and the Q-Q plot shows normalized data until the upper 2nd quantile, and lower 3rd quantile. Both indicate data normality for predicted values.

### Evaluating Significance of Model

To determine the significance of this model we first need to identify the null and alternative hypotheses which are as follows:

When viewing the summary analysis results, we can see a P-value of 2.2e-16 which is well under the 5% significance threshold we are seeking. With that said, we can safely reject the null hypothesis and state that there is a statistically significant relationship between price and one of the predictor variables in this model. To find the significance of each term we need to carry out individual beta tests on the variables. The hypotheses to compare against would be:

As can be seen from the summary analysis results, all beta values are 2e-16 *except for* the interaction term/variable “appliance\_age:crime” which has a beta value of 0.284. This means for each variable other than the interaction term, we can reject the null in favor of the alternative and say that each variable *is* significant to the model at a 5% significance threshold. The interaction term “appliance\_age:crime” should be thusly rejected and not used in this model, while the others should.

### Making Predictions Using Model

Using the previously generated model the prediction for the price of a home that has one-year-old appliances and is located in an area with a crime rate of 81.02 per 100,000 people is $864,423.40 and has a 90% prediction and confidence interval of [711566.6, 1017280] and [854,109.1, 874737.7] respectively. The prediction interval should contain 90% of predicted values within its range and the confidence interval should contain the mean of those values 90% of the time. For a home that has 15-year-old appliances and is located in an area with a crime rate of 200.50 per 100,000 people, the predicted price would be $271,051.60 with 90% prediction and confidence intervals of [118454.4, 423648.8] and [265846, 276257.2] respectively. We can see obviously that with a crime rate of more than double and aged appliances, the predicted price drops drastically. On the low end of the first home’s predicted values, you have a price of about 711,000 vs the low end of the second home predicted at 118,000. On the high end of prices on those same two homes the predicted prices are *almost* three times higher between the lower crime and higher crime areas.

## 5. Nested Models F-Test

### Reporting Results

To create a model that uses the age of appliances and crime rate per 100,000 people to predict the price of a home, we must come up with the general form of the model which is:

Using the summary analysis function to build the model equation, we get:

### Evaluating Significance of Model

By further analyzing the results of the summary function for this model, the resultant P-value is observed to be 2.2e-16 which is well below the 5% significance threshold of 0.05. We know this by comparing the null and alternative hypotheses for the model:

We can safely reject the null hypothesis in favor of the alternative and say that at least one predictor is significantly related to the response variable (price). For individual beta tests on the variables, we return to the results of the summary analysis function and see that each variable has a beta parameter of 2e-16, which are all below the 5% significance level meaning that they are safe to include in this model, and once they are compared against the null and alternative hypotheses for the individual terms shown below, have their null hypotheses rejected as well:

### Model Comparison

To compare this reduced model above with the complete model we can initially represent their general form(s) as:

*And:*

The general idea of a reduced versus a complete model is easier to see when compared side to side. The idea is that a reduced model is a “reduced” variation of a “complete” model with terms/variables that have been removed to compare statistical differences between models that exclude those variables. In this case, a model that includes the squared higher-order terms *appliance\_age2* and *crime2* is compared to a model without them. The null and alternative hypotheses for this test are:

The resulting P-value on the ANOVA test comes out as 2.113594e-28 which is well below the 5% significance threshold of 0.05, causing the rejection of the null hypothesis in favor of the alternative hypothesis. The complete model appears to be more accurate and *should* be used in this case.

## 6. Conclusion

To summarize, the model of choice here would seem to be the second “complete” model. The and values of the second model are higher, pointing to a more accurate model, as well as the results of the nested models F-test which shows that the complete model is also more accurate than a reduced version of itself. It doesn’t seem intuitive that appliance age would be a more apt predictor of price than the square footage of living space, number of bathrooms, view, or even grade of the property, but the math doesn’t lie. The variance of the dependent variable price is indeed better explained by crime rates and appliance age, to the tune of almost 16% more accuracy. This data could be used to take a comprehensive approach to predict the projected sales price, and therefore asking price, of the homes that our customers want us to market for them, which in turn could lead to better margins for the company, and our customers.