# MAT 303 Project One Summary Report

[Luna Saccoia]

[Luna.Saccoia@snhu.edu]

Southern New Hampshire University

## 1. Introduction

*The data set being explored for the statistical analyses is the housing data set. Information contained in this data set will be used to calculate the effects different variables, such as number of bedrooms and age of the home, can have on the sale price of the home.*

*I will be running 3 different analyses. I will be creating and analyzing a first order regression model with quantitative and qualitative variables, a second order regression model with quantitative variables and a nested model of the second order regression model.*

## 2. Data Preparation

*The important variables that we will be working with in this data set are price, age of home, square feet of living space, grade of the home, number of bathrooms, view, age of appliances, and crime rate.*

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

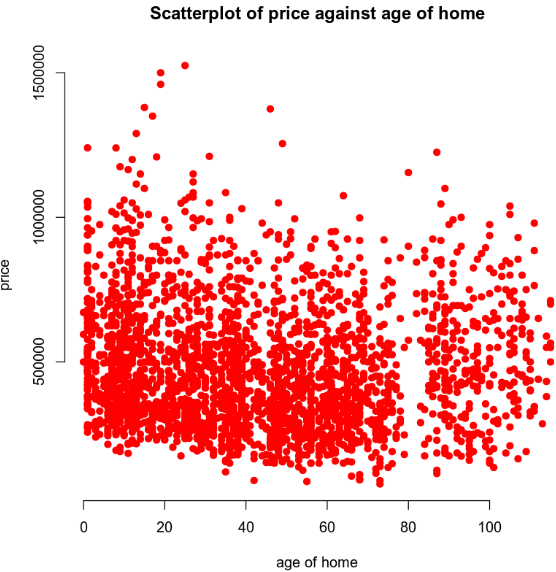
## 3. Model #1 - First Order Regression Model with Quantitative and Qualitative Variables

### Correlation Analysis

*Scatterplot of price (price) vs. living area (sqft\_living):*

**

*Scatterplot of price (price) vs. age of the home (age):*

**

*There is a positive trend between living space and price. If the living area increases, the price of the home increases. There is no discernible trend between age of home and price.*

*The correlation coefficient for living area is and for age is .*

*The strength of the correlation coefficients for living area and age of the home are:*

* *Moderate positive correlation for living area because the correlation coefficient for living area is positive at between 0.40 and 0.80.*
* *Weak negative correlation for age because the correlation coefficient is negative at between 0.0 and 0.40.*

### Reporting Results

*The general form equation for the multiple regression model using price as the response variable and living area, grade of the home, number of bathrooms and view as predictor variables is:*

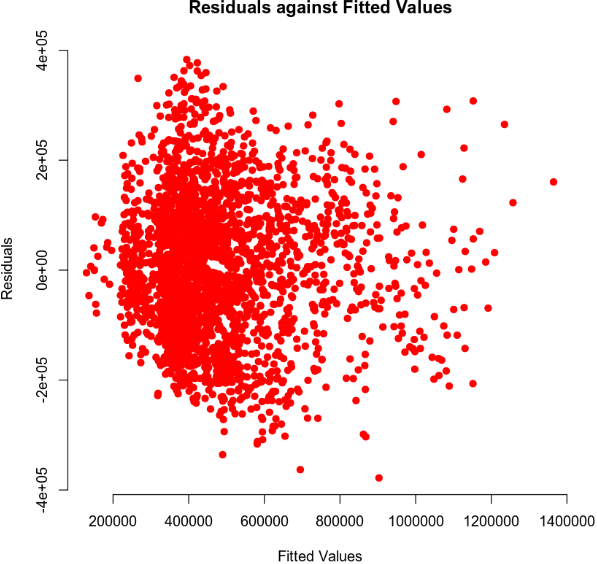
*The is the intercept. The values are the popular regression terms for living area, grade of the home, number of bathrooms, view1 and view2. The terms are the values that can be input for the variables living area, grade of the home, number of bathrooms, dummy variable for view=trees, and dummy variable for view=lake.*

*The model equation using these variables is (first is scientific notation, second is E-notation, third is expanded form):*

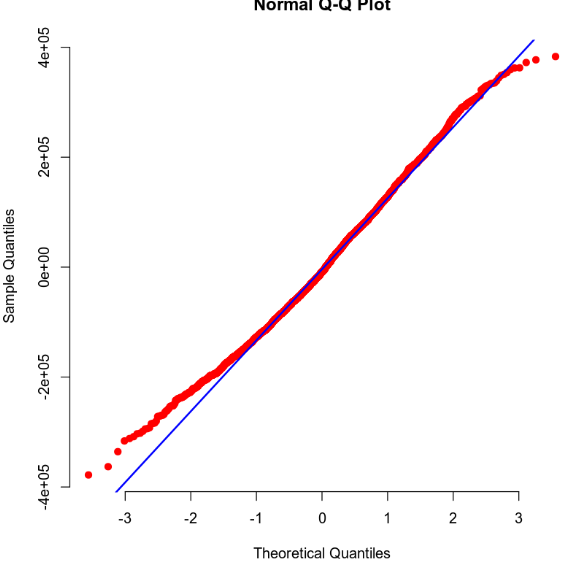
*The value of (R-squared) is and (Adjusted R-squared) is 0.6469. This means that approximately 65% of the variance in price can be explained by a model that uses living area, grade of the home, number of bathrooms, and dummy variables for view=trees and view=lake as predictor variables.*

*The beta estimate for living area is . This means that for every 1 unit increase in living area, the price will increase by . The beta estimate for lake view is . This means that if the house has a lake view, the price will increase by .*

*The plot for residual against fitted values:*

**

*The normal Q-Q plot:*

**

*(As per our discussion, I am omitting the list of fitted values and residuals here and in the code. The list was extensive and affected readability.)*

*The residuals against fitted values plot supports an assumption of homoscedasticity and linearity as no discernible pattern exists in the plot.*

### Evaluating Significance of Model

*To evaluate if the model is significant at a 5% level of significance, an overall F-test is used. The null and alternative hypotheses for this are:*

*The null hypothesis is that the beta estimates are equal to zero and therefore the model is not significant, meaning there is no statistically significant relationship between the predictor variables and the response variable.*

*The alternative hypothesis is that at least one of the beta estimates is not equal to zero and therefore the model is significant, meaning at least one of the predictor variables has a statistically significant relationship with the response variable.*

*The P-value for this model is (0.00000000000000022). This is significantly lower than the level of significance of 0.05. As such, the null hypothesis should be rejected in favor of the alternative hypothesis.*

*With this information, it can be concluded that at least one of the predictors has a slope coefficient that is significantly different than 0 and therefore a statistically significant relationship exists between at least one of the predictor variables and the response variable in this model.*

*To evaluate which terms are significant at a 5% level of significance, an individual T-test is used. The null and alternative hypotheses for this are:*

*= Living area (sqft\_living):*

*The P-value for sqft\_living is (0.0000000000000002)*

*= grade:*

*The P-value for grade is (0.0000000000000002)*

*= bathrooms:*

*The P-value for bathrooms is (0.0000233)*

*= tree view (view1):*

*The P-value for view1 (trees) is (0.0000000000000002)*

*= lake view (view2):*

*The P-value for view2 (lake) is (0.0000000000000002)*

*Each null hypothesis is that the predictor variable being tested is equal to zero, is not significant and has no statistically significant relationship with the response variable.*

*The alternative hypothesis is that the predictor variable is not equal to zero, is significant and has a statistically significant relationship with the response variable.*

*Every variable’s P-value is significantly lower than the level of significance of 0.05. As such, the null hypothesis should be rejected in favor of the alternative hypothesis.*

*With this information, it can be concluded that every predictor variable in the model, individually, has a statistically significant relationship with the response variable.*

### Making Predictions Using Model

***Prediction scenario 1***

*To make predictions using the multiple regression model, hypothetical values will be used. In this scenario the home backs out to a lake, has 2,150sq ft of living area, is a grade 7, and has 3 bathrooms. A 90% prediction and confidence interval for price will be obtained.*

*The equation for this scenario is:*

*The predicted price for a home that backs out to a lake, has 2,150 sq ft of living area, a grade of 7, and 3 bathrooms is $630,785.70.*

*The 90% prediction interval for the price of this home is (422684.5, 838887).*

*The prediction interval for an individual response means that we can be 90% certain that the price of a house will fall within these bounds, if it has a lake view, 2,150sq ft of living area, is a grade 7 and has 3 bathrooms. This is taking into account uncertainty related to to Y varying according to the regression error, , as well as sampling uncertainty related to estimating the regression parameters.*

*Taking the regression error into account as well as sampling uncertainty relating to estimating the regression parameters is the reason the prediction interval for an individual response is wider than the confidence interval for the mean. (Zybooks, 2016)*

*The 90% confidence interval for the price of this home is (610013.7, 651557.7).*

*The confidence interval for the mean tells us that we can be 90% certain that the average home price of a group of homes will fall within these bounds; if they have lake view, 2,150sq ft of living area, are a grade 7, and have 3 bathrooms.*

***Prediction scenario 2***

*To make predictions using the multiple regression model, hypothetical values will be used. In this scenario the home backs out to a road, has 2,150sq ft of living area, is a grade 7 and has 3 bathrooms. A 90% prediction and confidence interval for price will be obtained.*

*The equation for this scenario is:*

*The predicted price for a home that backs out to a road, has 2,150 sq ft of living area, a grade of 7, and 3 bathrooms is $402,121.90.*

*The 90% prediction interval for the price of this home is (194826, 609417.9).*

*The prediction interval for an individual response means that we can be 90% certain that the price of a house will fall within these bounds, if it has a road view, 2,150sq ft of living area, is a grade 7 and has 3 bathrooms. This is taking into account uncertainty related to to Y varying according to the regression error, , as well as sampling uncertainty related to estimating the regression parameters.*

*The 90% confidence interval for the price of this home is (392274.8, 411969.1).*

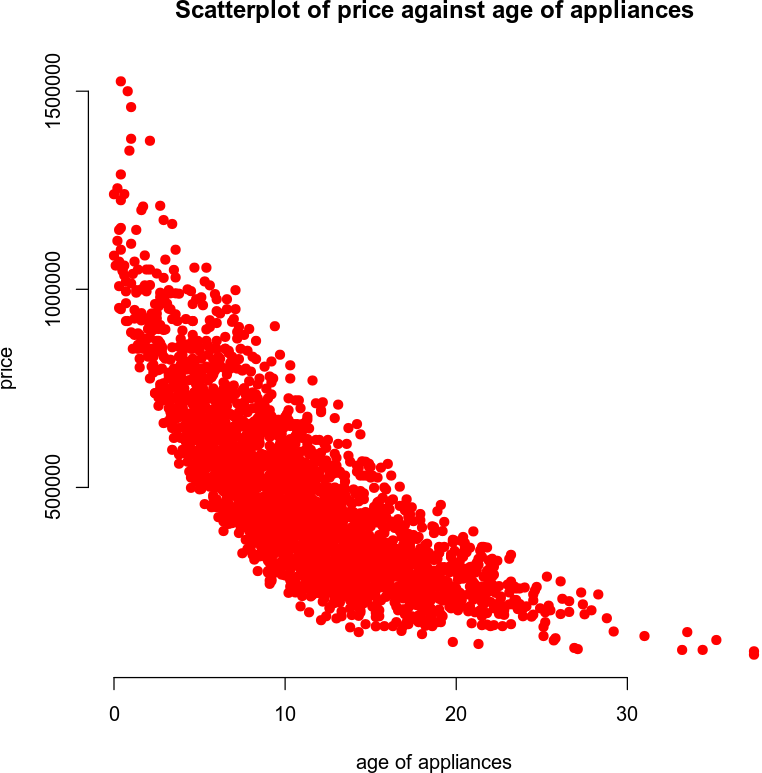
*The confidence interval for the mean tells us that we can be 90% certain that the average home price of a group of homes will fall within these bounds; if they have road view, 2,150sq ft of living area, are a grade 7, and have 3 bathrooms.*

*Taking the regression error into account as well as sampling uncertainty relating to estimating the regression parameters is the reason the prediction interval for an individual response is wider than the confidence interval for the mean. (Zybooks, 2016)*

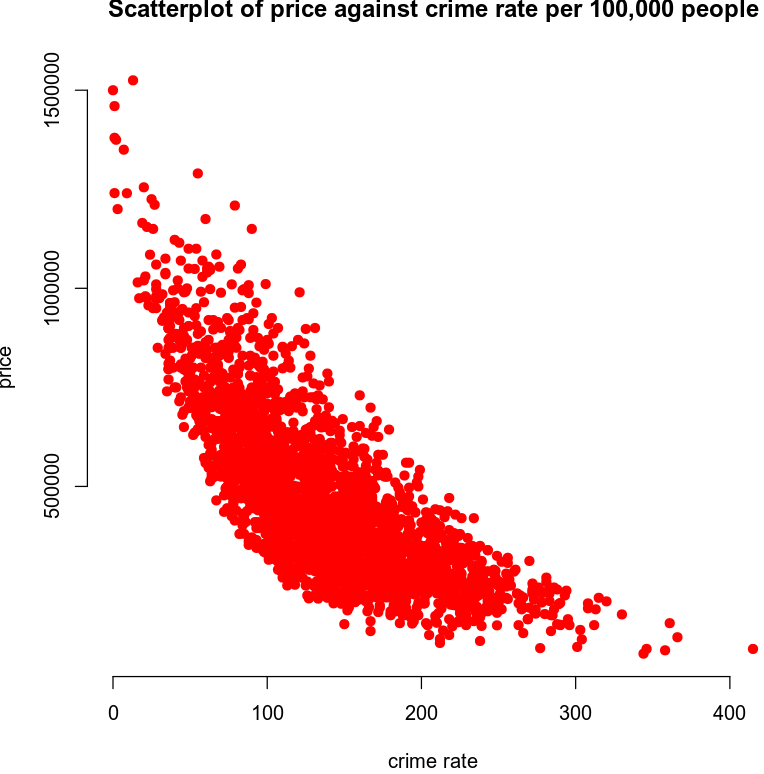
## 4. Model #2 - Complete Second Order Regression Model with Quantitative Variables

## Correlation Analysis

*Scatterplot for price vs. age of appliances (appliance\_age):*

**

*Scatterplot of price vs. crime rate per 100,000 people (crime):*

**

*Each of these scatterplots has a noticable curve to it. They do not follow a linear pattern and a second order model would be appropriate using these variables.*

### Reporting Results

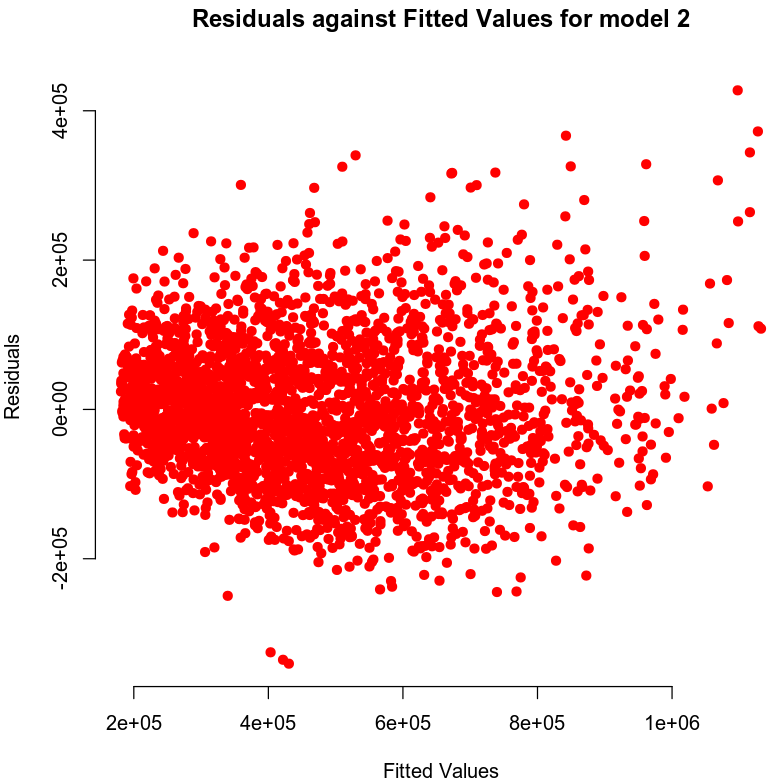
*The general equation for a complete second order model for price, using age of appliances and crime rate per 100,000 people as predictors is:*

*In this equation is the slope parameter. are the beta estimates for appliance age, crime, the interaction between appliance age and crime, appliance age squared and crime rate squared. are where we can input our hypothetical values for appliance age and crime to predict price based on those values. are the values for appliance age squared and crime rate squared.*

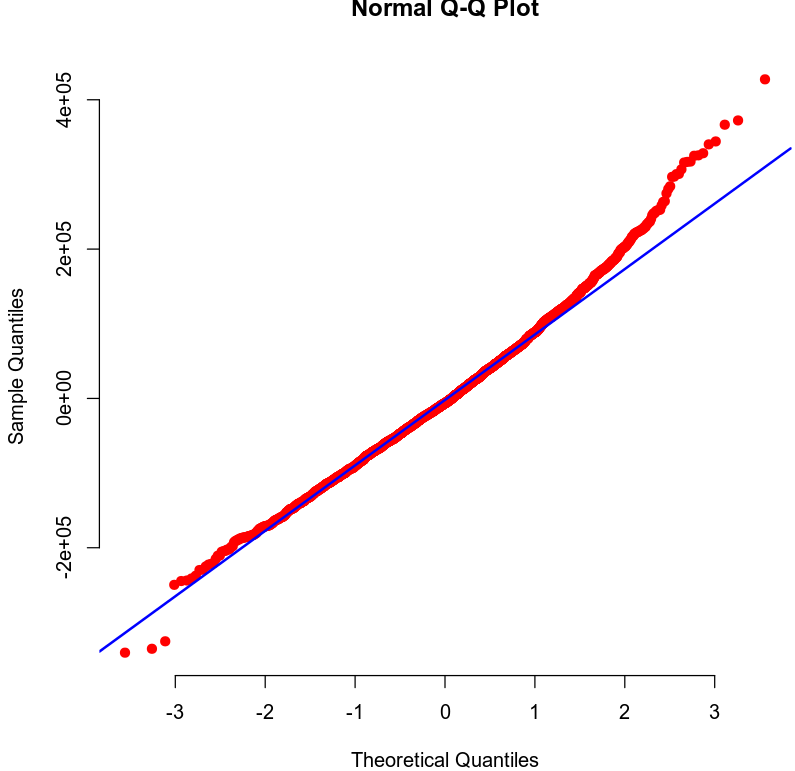
*The equation with the beta estimates from the model is: (first is scientific notation, second is E-notation, third is expanded form)*

*The value of (R-squared) is . The value of (adjusted R-squared) is . The is the coefficient of determination. It tells us we can conclude that approximately 81% of the variance in home price can be explained by a model that uses appliance age and crime rate as predictor variables.*

*The residuals against fitted values plot:*

**

*The normal Q-Q plot:*

**

*(As per our discussion, I am omitting the list of fitted values and residuals here and in the code. The list was extensive and affected readability.)*

*The residuals against fitted values plot supports an assumption of homoscedasticity and linearity. There is some deviation from the line on the Q-Q plot, but the deviation is not significant.*

### Evaluating Significance of Model

*To carry out an overall F-test our null and alternative hypotheses are:*

*The null hypothesis is that the beta estimates are equal to zero and therefore the model is not significant, meaning there is no statistically significant relationship between the predictor variables and the response variable.*

*The alternative hypothesis is that at least one of the beta estimates is not equal to zero and therefore the model is significant, meaning at least one of the predictor variables has a statistically significant relationship with the response variable.*

*The P-value is or 0.00000000000000022. This value is significantly lower than our level of significance of 0.05. Therefore the null hypothesis should be rejected in favor of the alternative hypothesis.*

*With this information, it can be concluded that a statistically significant relationship exists between at least one of the predictor variables and the response variable in this model.*

*To evaluate which terms are significant at a 5% level of significance, an individual T-test is used. The null and alternative hypotheses for this are:*

*The null hypothesis is that the predictor variable being tested is equal to zero, is not significant and has no statistically significant relationship with the response variable.*

*The alternative hypothesis is that the predictor variable is not equal to zero, is significant and has a statistically significant relationship with the response variable.*

*The P-value for appliance age is or 0.00000000000000022. The P-value for crime is or 0.00000000000000022. The P-value for the interaction between appliance age and crime is 0.284. The P-value for appliance age squared is or 0.00000000000000022. The P-value for crime rate squared is or 0.00000000000000022.*

*The P-values for appliance age, crime rate, appliance age squared, and crime rate squared are all well below the level of significance of 0.05. The P-value for the interaction of appliance age and crime is not below the level of significance.*

*It can be concluded that a statistically significant relationship exists between home price and the predictor variables of appliance age, crime rate, appliance age squared and crime rate squared. There is not a statistically significant relationship between home price and the interaction of appliance age and crime rate.*

### Making Predictions Using Model

***Prediction 1***

*To predict the price of a home that has 1 year old appliances and is in an area that has a crime rate of 81.02 per 100,000 people, the following equation is used:*

*The predicted price of a home that has 1 year old appliances and is in an area that has a crime rate of 81.02 per 100,000 people is $864,423.40.*

*The 90% prediction interval is (711566.6, 1017280). The prediction interval for an individual response means that we can be 90% certain that the price of a house will fall within these bounds, if it has 1 year old appliances and is in an area with a crime rate of 81.02 per 100,000 people.*

*The 90% confidence interval is (854109.1, 874737.7). The confidence interval for the mean tells us that we can be 90% certain that the average home price of a group of homes will fall within these bounds; if they have 1 year old appliances and are in an area with a crime rate of 81.02 per 100,000 people.*

***Prediction 2***

*To predict the price of a home that has 15 year old appliances and is in an area that has a crime rate of 200.50 per 100,000 people, the following equation is used:*

*The predicted price of a home that has 15 year old appliances and is in an area that has a crime rate of 200.50 per 100,000 people is $271,051.60.*

*The 90% prediction interval is (118454.4, 423648.8). The prediction interval for an individual response means that we can be 90% certain that the price of a house will fall within these bounds, if it has 15 year old appliances and is in an area with a crime rate of 200.50 per 100,000 people.*

*The 90% confidence interval is (265846, 276257.2). The confidence interval for the mean tells us that we can be 90% certain that the average home price of a group of homes will fall within these bounds; if they have 15 year old appliances and are in an area with a crime rate of 200.50 per 100,000 people*

## 5. Nested Models F-Test

### Reporting Results

*The general form equation of a first order model for price using appliance age and crime rate per 100,000 people as predictor variables is:*

*In this equation is the slope parameter. are the beta estimates for appliance age, crime, and the interaction between appliance age and crime. are where we can input our hypothetical values for appliance age and crime to predict price based on those values.*

*The equation with the beta estimates from the model is: (first is scientific notation, second is E-notation, third is expanded form)*

### Evaluating Significance of Model

*To carry out an overall F-test our null and alternative hypotheses are:*

*The null hypothesis is that the beta estimates are equal to zero and therefore the model is not significant, meaning there is no statistically significant relationship between the predictor variables and the response variable.*

*The alternative hypothesis is that at least one of the beta estimates is not equal to zero and therefore the model is significant, meaning at least one of the predictor variables has a statistically significant relationship with the response variable.*

*The P-value is or 0.00000000000000022. This value is significantly lower than our level of significance of 0.05. Therefore the null hypothesis should be rejected in favor of the alternative hypothesis.*

*As such, it can be concluded that a statistically significant relationship exists between the response variable for home price and at least one of the predictor variables for appliance age, crime rate and the interaction of appliance age and crime rate.*

*To evaluate which terms are significant at a 5% level of significance, an individual T-test is used. The null and alternative hypotheses for this are:*

*The null hypothesis is that the predictor variable being tested is equal to zero, is not significant and has no statistically significant relationship with the response variable.*

*The alternative hypothesis is that the predictor variable is not equal to zero, is significant and has a statistically significant relationship with the response variable.*

*The P-value for appliance age is or 0.00000000000000022. The P-value for crime is or 0.00000000000000022. The P-value for the interaction between appliance age and crime is or 0.00000000000000022.*

*The P-values for appliance age, crime and the interaction of appliance age and crime are all well below the level of significance of 0.05. As such, the null hypothesis should be rejected in favor of the alternative hypothesis. It can be concluded that each of the predictor variables for appliance age, crime rate and the interaction of appliance age and crime rate, individually, have a statistically significant relationship with the response variable of home price.*

### Model Comparison

*To compare this reduced model with the complete second order model for price using appliance age and crime per 100,000 people as predictors, to test whether the quadratic (squared) terms contribute in predicting the price of homes, a nested model F-test can be performed.*

*In general the reduced model is the model without the terms being tested for necessity. The complete model is the model that includes these terms being tested. The nested model F-test is used to test whether we can omit the variables being tested, in this case the quadratic terms, and just use the reduced model, or if we need to include these terms and use the complete second order model.*

*The equation for the complete second order model is:*

*In this equation is the slope parameter. are the beta estimates for appliance age, crime, the interaction between appliance age and crime, appliance age squared and crime rate squared. are where we can input our hypothetical values for appliance age and crime to predict price based on those values. are the values for appliance age squared and crime rate squared.*

*The equation for the reduced model is:*

*In this equation is the slope parameter. are the beta estimates for appliance age, crime, and the interaction between appliance age and crime. are where we can input our hypothetical values for appliance age and crime to predict price based on those values.*

*To run the nested model F-test (ANOVA) at a 5% level of significance to determine if the quadratic terms are needed, the null and alternative hypotheses are:*

*The null hypothesis is that the beta estimates are equal to zero and therefore the model is not significant, meaning there is no statistically significant relationship between the predictor variables and the response variable.*

*The alternative hypothesis is that at least one of the beta estimates is not equal to zero and therefore the model is significant, meaning at least one of the predictor variables has a statistically significant relationship with the response variable.*

*The P-value for the nested model F-test (ANOVA) is . This is well below the level of significance of 0.05. The null hypothesis should be rejected in favor of the alternative hypothesis. As such, it can be concluded that the quadratic terms are significant and should be included. The reduced model is not sufficient and the complete second order model must be used.*

## 6. Conclusion

*The statistical analyses performed included a first order regression model, a complete second order regression model, a reduced model test and a nested F-Test (ANOVA) between the reduced model and the complete second order model to determine whether the reduced model is sufficient or is the complete second order model must be used.*

*I would use the complete second order model with predictors for appliance age, crime rate, interaction between appliance age and crime rate, appliance age squared and crime rate squared. The value is the highest for this model. This tells us that the highest percentage of variance can be explained using this model. It shows a strong correlation between these predictors and the response variable of home price.*

*The practical importance of the analyses performed is that it can be used to predict a home’s worth depending on the factors included, in the case of the model I chose, appliance age and crime rate. The analyses show that these variables have a strong relationship with home price and as such, are good predictors to use when appraising a home’s value and predicting the price it could sell for.*

## 7. Citations

*Zybooks MAT 303: Applied Statistics II for Science, (2016, August).*

*Retrieved March 16, 2020, from https://learn.zybooks.com/zybook/SNHUMAT303v1*