# MAT 303 Project One Summary Report

Nicholas Kreuziger

nicholas.kreuziger@snhu.edu

Southern New Hampshire University

## 1. Introduction

The dataset we are exploring is a collection of all house sale prices for King County Washington between May 2014 and May 2015 (Harlfoxem, 2016). This data will be used to create two Regression Models. One of them is a First Order Regression and the other is a Second Order Regression Model. These results will be used to help interpret the impact of different variables on house sale price.

## 2. Data Preparation

The full housing dataset consists of 22 columns and 2692 rows. We will only be utilizing 12 of these columns in our own dataset. Below is a table of those important columns and their representations. Graphical user interface, text, application

Description automatically generated with medium confidence

While there are many variables in this dataset we will be focusing on price as the response variable in all our models. The predictor variables that will be utilized to test their effects on the price variable will be sqft\_living, grade, bathrooms, view, appliance\_age and crime.

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

### Correlation Analysis

Chart, scatter chart

Description automatically generated

Above is a scatter plot of Housing Price and Living Area (in Square Feet). There is a positive linear trend, as Living Area increases Housing Price also increases.

Chart, scatter chart

Description automatically generated

Above is a scatter plot of Housing Price and Age of a Home. There is not any significant linear trend in this relationship.

Calculating a Pearson Correlation Coefficient Matrix of the variables’ housing price, living area and age helps to define the linear relationship the scatterplots visualize. Zybooks MAT 303: Applied Statistics II for Science Section 3.3 Table 3.3.1 defines a Strong correlation as anything above 0.80, and a Moderate Correlation between 0.40 and 0.80.

Table 3.3.1: Strength of correlation.

|  |  |
| --- | --- |
| Value of |R| | Strength of correlation |
| 0<|R|≤0.40 | Weak |
| 0.40<|R|≤0.80 | Moderate |
| 0.80<|R|≤1.00 | Strong |

Taking these definitions of Correlation Strength into account, we can analyze housing price and its independent correlation to living area and age.

|  |  |  |
| --- | --- | --- |
| Relationship | Pearson Correlation Coefficient | Strength |
| Price and Living Area | 0.6895 | Moderate, Positive |
| Price and Age | -0.0746 | Weak, Negative |

### Reporting Results

Our first model will be a first order multiple regression model using price as a response variable as influenced by Living Area, Grade of Home, Number of Bathrooms and View. These variables will be represented as below:

E(Y) = Estimated Price of the House

β0 = Intercept

X1 = Square Feet Living Area

β1 = Square Feet Living Area Coefficient

X2 = Grade

β2 = Grade Coefficient

X3 = Bathrooms

β3 = Bathrooms Coefficient

X4 = View1

β4 = View Coefficient

X5 = View2

β5 = View Coefficient

Since the View Coefficients are Qualitative, the data only communicates in binary 1, 0 values. As such multiple possibilities of the column are expanded in variables called dummy variables. View is a variable with 3 possible values of 0,1 and 2. The reference level of this model is 0. Below is a table decoding the use of the variables in this model.

|  |  |  |
| --- | --- | --- |
| View Value | X4 Value | X5 Value |
| 0 | 0 | 0 |
| 1 | 1 | 0 |
| 2 | 0 | 1 |

The general form of the model will appear:

Using Statistical Software for this model, the resulting summary is:

Text

Description automatically generated

From these computations the resulting R-Squared *R-Squared* and adjusted R-Squared *Adjusted R-Squared* values are,

64.75% and 64.69%. R-Squared is the variation in the Dependent Variable (House Price) explained by the independent variables. These numbers indicate you can predict 64.75% of the variance in housing price with this first order model. Adjusted R-Squared is the variation explained by only the independent variables that significantly help in explaining the dependent variable, Adjusted R-Squared penalizes adding independent variables that do not help in predicting the dependent variable. The Adjusted R-Squared value only differing by 0.06% (64.75%-64.69%) indicates that the variables currently residing in the model are not significantly penalizing variance accountability.

The beta estimates indicate that the general form of this model becomes:

The beta estimates for View and Living Area Square Footage are:

Sqft\_living = 9.384e+01

View 1 = 1.657e+05

View 2 = 2.287e+05

A positive value for beta estimates indicates an upward concavity for the variables Sqft\_living, View 1 and View 2.

These beta estimates tell us, with all other values fixed, that for every square foot of living space price increases by $93.84. For a lakeview you can expect price to increase by $228,700!

Using Residuals and Fitted Values we will create a scatter plot of Residuals and Fitted Values in addition to a Q-Q plot plotting sample quantiles against theoretical quantiles.

A scatterplot plotting Residual values against Fitted Values evaluates the Mean of Zero and Constance Variance assumptions of a multiple regression model. The Mean of Zero assumption assumes that the residual values in a model are zero (the response variable is linear in response to predictor variables), if they are not then the model may be nonlinear. Constant variance assumes that the residuals for each of the predictor variables should have equal or similar variance, called homoscedasticity. Heteroscedasticity is a condition where unequal and unrelated variance occurs. The scatterplot for our model does not appear to have any plot patterns obstructing linearity nor patterns in variance.

*Chart, scatter chart

Description automatically generated*

A Q-Q plot is a visualization of the distribution of errors. The Q stands for Quantiles, which are continuous intervals with equal probabilities across a probability distribution. A normalized set of errors will follow a linear pattern. The Q-Q plot for this model indicates normally distributed residuals. There does appear to be some divergence in the lower and upper quantiles in this plot.

Chart, line chart

Description automatically generated

The histogram below demonstrates that there appear to be some out of the norm values once you approach the realm of -/+$400,000. Considering the below 70% fit of model 1 it makes sense that there may be some error distribution outlying a normalized trend.

Chart, histogram

Description automatically generated

### Evaluating Significance of Model

Evaluating this model’s significance at a 5% level, we will be using the ANOVA (Analysis of Variance) F-Test. An F-test applied to a multiple regression model is used to determine if the overall model and its variables are collectively influencing the response variable in a statistically significant manner. After determining the results of the F-Test, if a relationship exists you can evaluate individual relationships in the model with a multiple regression individual t-test. This is useful when you want to troubleshoot or assess the value of certain predictor variables in a model.

The resulting model computations for model 1 using statistical software are below:

Text

Description automatically generated

The Overall F-Test is testing if the response variable has a relationship with at least one of the predictor variables. The Significance Level of our test is 5%, mathematically α = 0.05.

The Null hypothesis, H0 is that the predictor variables have no relationship with the response variable. Mathematically stated, the beta coefficient of predictor variables sqft\_living, grade, bathrooms, view 1 and view 2 are equal to 0.

H0 : β1 = β2 = β3 = β4 = 0

The Alternative hypothesis, Ha is that at least one predictor variable has a relationship with the response variable house price. Mathematically stated, the beta coefficient of predictor variables are not equal to 0.

Ha : βi ≠ 0 for *i* = 1, 2, 3, 4

The P-Value of the Overall F-Test is 2.2 e-16. Since P-Value < Significance Level the null hypothesis is rejected in favor of the alternative hypothesis. At least one predictor variable has a relationship to the response variable house price.

Individual t-tests can be used once the F-Test has determined the model has a predictor variable with a relationship to the response variable. Every individual t-test has a similar Null and Alternative Hypothesis. We will be testing with a significance level of 5%, mathematically α = 0.05.

The Null Hypothesis, H0, is that the predictor variable coefficient βi (i = 1, 2, 3, 4 for each individual test) = 0. The Alternative Hypothesis, Ha , is that the predictor variable coefficient βi  (i = 1, 2 , 3, 4 for each individual test) ≠ 0.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | p-value | Significance level | Result |
| Sqft\_living | 2e-16 | 0.05 | Reject Null Hypothesis |
| Grade | 2e-16 | 0.05 | Reject Null Hypothesis |
| Bathrooms | 2.33e-05 | 0.05 | Reject Null Hypothesis |
| View 1 | 2e-16 | 0.05 | Reject Null Hypothesis |
| View 2 | 2e-16 | 0.05 | Reject Null Hypothesis |

The t-test results indicate every variable is statistically significant in relation to house price.

### Making Predictions Using Model

Using this model, we can make predictions about house price given variable values. Let’s say we are given a house that backs out to a lake (view = 0), with 2,150 sq ft living area, a grade of 7 and three bathrooms. We can predict the cost of a house with these features by applying the variable values to the model.

A prediction interval is used to predict what range a future individual observation may fall in all outcomes. A confidence interval is used to predict the range the average of the response variable will fall. In short, a prediction interval tries to account for all possible values with the model for a single observation while a confidence interval tries to predict where the average of the probability distribution will fall given specific values for the predictor variables. Confidence intervals will always be narrower than Prediction intervals since a Confidence interval is the estimation of a range of an average while the predictor interval is the range of values for a single instance.

Using this model, we can construct the confidence and prediction intervals for House Price given a house that backs out to a lake (view = 0), with 2,150 sq ft living area, a grade of 7 and three bathrooms.

The 95% prediction interval for house price of a house with these qualities is,

Text

Description automatically generated with low confidence

The prediction interval lower and upper ranges are $382,791.60 to $878,779.90. If we were to observe a house with these qualities selling, we would observe a sale within this lower and upper range 95% of the time.

Text

Description automatically generated with low confidence

The confidence interval lower and upper ranges are $606,031.70 to $655,539.70. If we were to observe a house with these qualities selling repeatedly, we would arrive at an average sell price in this range. This range is predicted to be a part of the true population 95% of the time.

Let’s compare the price of the house above with the price of a house sharing the same qualities except the view is a dirt road instead of a lake. We are given a house that backs out to the road (view = 2), with 2,150 sq ft living area, a grade of 7 and three bathrooms. We can predict the cost of a house with these features by applying the variable values to the model.

Using this model, we can construct the confidence and prediction intervals for House Price given a house that backs out to a lake (view = 0), with 2,150 sq ft living area, a grade of 7 and three bathrooms.

The 95% prediction interval for house price of a house with these qualities is,

Text

Description automatically generated with low confidence

The prediction interval lower and upper ranges are $382,791.60 and $878,779.90. If we were to observe a house with these qualities selling, we would observe a sale within this lower and upper range 95% of the time.

Text

Description automatically generated with low confidence

The confidence interval lower and upper ranges are $606,031.70 and $655,539.70. If we were to observe a house with these qualities selling repeatedly, we would arrive at an average sell price in this range. This range is predicted to be a part of the true population 95% of the time.

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

### Correlation Analysis

**Chart, scatter chart

Description automatically generated**

The above scatterplot compares Appliance Age to Housing Price. Due to the curve of the trend, it cannot be taken to be a linear relationship. This relationship is more appropriately modelled with a Second Order model.

**Chart, scatter chart

Description automatically generated**

The above scatterplot compares Crime per 100,000 Citizens to Housing Price. Due to the curve of the trend, it cannot be taken to be a linear relationship. This relationship is more appropriately modelled with a Second Order model.

### Reporting Results

We will create a second-order regression model with the quantitative predictor variables appliance age and crime rate per 100,000 citizens. The response variable will be housing price. This will be model 2. These variables will be represented as below:

β0 = Intercept

X1 = Appliance Age

β1 = Appliance Age Coefficient

X2 = Crime Rate

β2 = Crime Rate Coefficient

X3 = Crime Rate: Appliance Age

β3 = Crime Rate: Appliance Age Coefficient

β4 = Appliance Age ^2 Coefficient

β5 = Crime Rate ^2 Coefficient

The general form of our regression model wherein y = house price and X1 = appliance age and X2 = crime rate is:

Using Statistical Software for this model, the resulting summary is:

Text, table

Description automatically generated

From these computations the resulting R-Squared *R-Squared* and adjusted R-Squared *Adjusted R-Squared* values are,

80.88% and 80.84%. R-Squared is the variation in the Dependent Variable (House Price) explained by the independent variables. These numbers indicate you can predict 80.87% of the variance in housing price with this second order model. Adjusted R-Squared is the variation explained by only the independent variables that significantly help in explaining the dependent variable, Adjusted R-Squared penalizes adding independent variables that do not help in predicting the dependent variable. The Adjusted R-Squared value only differing by 0.04% (80.88%-80.84%) indicates that the variables currently residing in the model are not significantly penalizing variance accountability.

The beta estimates indicate that the general form of this model becomes:

A scatterplot plotting Residual values against Fitted Values evaluates the Mean of Zero and Constance Variance assumptions of a multiple regression model. The Mean of Zero assumption assumes that the residual values in a model are zero (the response variable is linear in response to predictor variables), if they are not then the model may be nonlinear. Constant variance assumes that the residuals for each of the predictor variables should have equal or similar variance, called homoscedasticity. Heteroscedasticity is a condition where unequal and unrelated variance occurs. The scatterplot for our model does not appear to have any major plot patterns obstructing linearity nor patterns in variance. There does appear to be a slight curve in the positive direction of residuals as fitted values approach $1,000,000 house prices.

Chart, scatter chart

Description automatically generated

A Q-Q plot is a visualization of the distribution of errors. The Q stands for Quantiles, which are continuous intervals with equal probabilities across a probability distribution. A normalized set of errors will follow a linear pattern. The Q-Q plot for this model indicates normally distributed residuals up until 3+ quantiles. Divergence from a linear set of Quantiles begins at 3+. Meaning that the data is skewed to the right.

Chart, line chart

Description automatically generated The histogram below further reinforces and qualifies that the data is skewed right. Indicating the residuals tend to be -200,000 to 0.

**Chart, histogram

Description automatically generated**

### Evaluating Significance of Model

Evaluating this model’s significance at a 5% level, we will be using the ANOVA (Analysis of Variance) F-Test. An F-test applied to a multiple regression model is used to determine if the overall model and its variables are collectively influencing the response variable in a statistically significant manner. After determining the results of the F-Test, if a relationship exists you can evaluate individual relationships in the model with a multiple regression individual t-test. This is useful when you want to troubleshoot or assess the value of certain predictor variables in a model.

The resulting model computations for model 2 using statistical software are below:

Text, table

Description automatically generated

The Overall F-Test is testing if the response variable has a relationship with at least one of the predictor variables. The Significance Level of our test is 5%, mathematically α = 0.05.

The Null hypothesis, H0 is that the predictor variables have no relationship with the response variable. Mathematically stated, the beta coefficient of predictor variables appliance\_age, crime, appliance\_age ^2 and crime ^2 are equal to 0.

H0 : β1 = β2 = β3 = β4 = β5 = 0

The Alternative hypothesis, Ha is that at least one predictor variable has a relationship with the response variable house price. Mathematically stated, the beta coefficient of predictor variables are not equal to 0.

Ha : βi ≠ 0 for *i* = 1, 2, 3, 4, 5

The P-Value of the Overall F-Test is 2.2 e-16. Since P-Value < Significance Level the null hypothesis is rejected in favor of the alternative hypothesis. At least one predictor variable has a relationship to the response variable house price.

Individual t-tests can be used once the F-Test has determined the model has a predictor variable with a relationship to the response variable. Every individual t-test has a similar Null and Alternative Hypothesis. We will be testing with a significance level of 5%, mathematically α = 0.05.

The Null Hypothesis, H0, is that the predictor variable coefficient βi (i = 1, 2, 3, 4, 5 for each individual test) = 0. The Alternative Hypothesis, Ha , is that the predictor variable coefficient βi  (i = 1, 2 , 3, 4, 5 for each individual test) ≠ 0.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | p-value | Significance level | Result |
| Appliance\_age | 2e-16 | 0.05 | Reject Null Hypothesis |
| Crime | 2e-16 | 0.05 | Reject Null Hypothesis |
| Appliance\_age:Crime | 0.284 | 0.05 | Fail to Reject Null Hypothesis |
| Appliance\_age ^2 | 2e-16 | 0.05 | Reject Null Hypothesis |
| Crime ^2 | 2e-16 | 0.05 | Reject Null Hypothesis |

The t-test results indicate every variable is statistically significant in relation to house price except for the interaction term between Appliance Age and Crime. It can be concluded that for any change in either variable they will not affect the other.

### Making Predictions Using Model

Using this model, we can make predictions about house price given variable values. Let’s say we are given a house with one year old appliances located in an area with 81.02 per 100,000 individual crime rate. We can predict the cost of a house with these features by applying the variable values to the model.

Using this model, we can construct the confidence and prediction intervals for house price given a house with appliances on average a year old in an area with a crime rate of 81.02.

The 95% prediction interval for house price of a house with these qualities is,

Text

Description automatically generated with medium confidence

The prediction interval lower and upper ranges are $682,264 to $1,046,583. If we were to observe a house with these qualities selling, we would observe a sale within this lower and upper range 95% of the time.

Graphical user interface

Description automatically generated with medium confidence

The confidence interval lower and upper ranges are $852,131.80 to $876,714.90. If we were to observe a house with these qualities selling repeatedly, we would arrive at an average sell price in this range. This range is predicted to be a part of the true population 95% of the time.

Let’s do the same analysis of a house with 15-year-old appliances and a crime rate of 200.50. We can predict the cost of a house with these features by applying the variable values to the model.

Using this model, we can construct the confidence and prediction intervals for House Price given a house with 15-year-old appliances and a crime rate of 200.50.

The 95% prediction interval for house price of a house with these qualities is,

Graphical user interface, text, application

Description automatically generated

The prediction interval lower and upper ranges are $89,201.59 to $452,901.60. If we were to observe a house with these qualities selling, we would observe a sale within this lower and upper range 95% of the time.

Graphical user interface, application

Description automatically generated

The confidence interval lower and upper ranges are $264,848.10 and $277,255.10. If we were to observe a house with these qualities selling repeatedly, we would arrive at an average sell price in this range. This range is predicted to be a part of the true population 95% of the time.

## 5. Nested Models F-Test

### Reporting Results

The general form of our regression model (we will call this model 3) wherein y = house price and X1 = appliance age and X2 = crime rate with an interaction term between Appliance Age and Crime Rate is:

Performing a summary of model 3 results in:

Text

Description automatically generated with medium confidence

Indicating the coefficients of the general model should be:

### Evaluating Significance of Model

Evaluating this model’s significance at a 5% level, we will be using the ANOVA (Analysis of Variance) F-Test. An F-test applied to a multiple regression model is used to determine if the overall model and its variables are collectively influencing the response variable in a statistically significant manner. After determining the results of the F-Test, if a relationship exists you can evaluate individual relationships in the model with a multiple regression individual t-test. This is useful when you want to troubleshoot or assess the value of certain predictor variables in a model.

The resulting model computations for model 3 using statistical software are below:

Text

Description automatically generated with medium confidence

The Overall F-Test is testing if the response variable has a relationship with at least one of the predictor variables. The Significance Level of our test is 5%, mathematically α = 0.05.

The Null hypothesis, H0 is that the predictor variables have no relationship with the response variable. Mathematically stated, the beta coefficient of predictor variables appliance\_age, crime, appliance\_age:crime are equal to 0.

H0 : β1 = β2 = β3 = 0

The Alternative hypothesis, Ha is that at least one predictor variable has a relationship with the response variable house price. Mathematically stated, the beta coefficient of predictor variables is not equal to 0.

Ha : βi ≠ 0 for *i* = 1, 2, 3

The P-Value of the Overall F-Test is 2.2 e-16. Since P-Value < Significance Level the null hypothesis is rejected in favor of the alternative hypothesis. At least one predictor variable has a relationship to the response variable house price.

Individual t-tests can be used once the F-Test has determined the model has a predictor variable with a relationship to the response variable. Every individual t-test has a similar Null and Alternative Hypothesis. We will be testing with a significance level of 5%, mathematically α = 0.05.

The Null Hypothesis, H0, is that the predictor variable coefficient βi (i = 1, 2, 3 for each individual test) = 0. The Alternative Hypothesis, Ha , is that the predictor variable coefficient βi  (i = 1, 2 , 3 for each individual test) ≠ 0.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | p-value | Significance level | Result |
| Appliance\_age | 2e-16 | 0.05 | Reject Null Hypothesis |
| Crime | 2e-16 | 0.05 | Reject Null Hypothesis |
| Appliance Age:Crime | 2e-16 | 0.05 | Reject Null Hypothesis |

The t-test results indicate every variable is statistically significant in relation to house price.

### Model Comparison

When comparing models with a Nested Models F-Test two models are considered. The Complete Model is the original Model with the most factors. The Reduced Model is the Complete Model, absent factors as defined by the experiment. In this instance, we are comparing the second order model (Model #2) made earlier in this report with the first order model we just created. In both models X1 represents Appliance Age and X2 represents Crime Rate.

Complete Model:

Reduced Model:

To compare these models statistically we will conduct an Analysis of Variance (ANOVA) Test at a 5% level of significance. Mathematically, Significance Level α = 0.05. This F-Test will test if the Reduced Model is sufficient, or the addition of the squared terms is needed.

The Null Hypothesis is that the complete model’s Squared Term’s coefficients, β3 and β4 have no linear relationship to the response variable House Price. Mathematically, H0: β3 = β4 = 0.

The Alternative Hypothesis is that least one of the complete model’s Squared Term’s coefficients, β3 and β4 is non-zero meaning the squared terms are needed. Mathematically, Ha: βi ≠ 0 for i = 3, 4. The summary of the ANOVA test:

Graphical user interface, application

Description automatically generated

The p-value for this test is 2.113594e-28. This is lower than the significance value of 0.05. The Null Hypothesis is rejected, and the Alternative Hypothesis is concluded. At least one of the squared terms has a higher coefficient than 0 and the Complete model is necessary.

## 6. Conclusion

The Models used to model the house prices in this report had a few variations. Below is a table summarizing the key points and variables of each:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Coefficients | Order | Adj R-Squared | Model |
| 1 | 5 | 1 | 64.69% |  |
| 2 | 4 | 2 | 80.84% |  |
| 3 | 3 | 1 | 79.93% |  |

Using a Nested Models F-Test we concluded the necessity of the squared terms from Model 2. Additionally, Model 2 has the highest Adjusted R-Squared Value. I’d feel comfortable recommending Model 2. Practically speaking, these analyses help to cross reference between a few key variables from this housing dataset to determine which variables more actively influence housing price. We learned that the number of variables wasn’t as important as the variables selected. The second order model even enhanced a model consisting of two variables and accounted for upwards of 80% of the variability in housing price. This is impressive. It also warrants more research and begs the question if a model could be built that predicts 85% or higher variance in housing prices in King County. These analyses could be a starting point to analyzing larger housing price trends and provide larger scale economic insights.

## 7. Citations

Harlfoxem. (2016). House Sales in King County, USA [Data file]. Retrieved from https://www.kaggle.com/harlfoxem/housesalesprediction