# MAT 303 Project One Summary Report

Catherine Taft

Catherine.taft@snhu.edu

Southern New Hampshire University

## 1. Introduction

I really just can’t keep a job it seems. I’m now a data analyst for a real estate company in Seattle and we’ll be looking at some real estate data and looking at which attributes of a house may affect the sales price. We’ll be using a variety of multiple regression models, including both first order and higher order.

This’ll be useful for the real estate agency to evaluate the price ballpark for houses that we put on the market. If we set an unreasonably high price, obviously people are less likely to buy it, especially if it’s a buyer’s market. The same could be said perhaps for a suspiciously low price. So finding the right price range for the specific attributes of the house will make sure we have a good starting point for sales price.

## 2. Data Preparation

There are quite a few variables in this dataset, but the ones we’ll be working with are price, age of the house, crime rate, square footage, the grade (quality of craftsmanship), the view (lake or trees), number of bathrooms, and the age of the appliances.

This dataset has 22 columns and 2,692 rows.

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

### Correlation Analysis

First we’ll start with a scatterplot for price vs living area, and price vs age of house.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Price and square footage clearly have a positive correlation. As the square footage increases, so does the price. Most of them are clustered below $1,000,000 and under 4,000 square feet but there’s an obvious linear relationship. There doesn’t seem to really be a relationship between price and the age of the house. Maybe a very slight negative correlation. The plot just kind of shows us that most houses in general are below $1,000,000 (as the other plot also indicates) and that most of them are built less than 80 years ago, none of which is surprising, and doesn’t indicate a clear linear relationship between those variables. There does seem to be a curve to the plot but I’m not sure it’s relevant.

To find the correlation coefficients we can create a correlation matrix:

Table

Description automatically generated

The correlation coefficient for price vs square footage is 0.6895, which indicates a pretty strong positive correlation. The correlation coefficient for price vs age is -0.0746 which shows a negative correlation but the number is so close to 0 that it indicates a very weak relationship.

### Reporting Results

Now we’ll run a regression model with price as the response variable and living area, grade, bathrooms and views 1 and 2 as predictor variables.

The general equations is:

Where the variables are, in order, living area, grade, bathrooms, view 1 (tree view), view 2 (lake).

To write the equation for our scenario here we need to create and run the model first:

Table

Description automatically generated

So based on this we can sub in the values for our equation:

These are all in scientific notation (I misunderstood previously and thought the e notation in the output was just an exponent) so this evaluates to:

The R-squared and Adjust R-squared values are 0.6475 and 0.6469 respectively, which indicate a decently solid model where about 65% of the variation in price can be explained by these parameters.

The beta estimates for square footage and lake view (view2) are and respectively. This means that for every unit increase in square footage, the price will increase by $93.84, and that if the house has a lake view the price will increase by $228,700, which is a lot.

I don’t really understand what it means when it says include tables for residuals or why it would say that because there’s a huge number of them and it doesn’t seem useful to include that. So I guess I’ll just include this, which is what I get when I print it. Obviously I can’t include all of them.

Table, calendar

Description automatically generated

Table

Description automatically generated

More useful than that are the scatterplots of residuals against fitted values, and the QQ plot:

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

Based on this, I’m not quite sure about the residuals vs fitted graph. I’m inclined to say it shows homoscedasticity but the distinct clumping on the left gives me some pause. The QQ plot shows a normal distribution as there’s very little deviation from the regression line.

### Evaluating Significance of Model

To test the model at a 5% significance level we can do the overall F-test. The null hypothesis is that the coefficients for the predictor variables (the beta estimates) are all 0, indicating no linear relationship between them and the response variable. The alternative hypothesis is that at least one beta estimate is not 0, indicating that it has a meaningful relationship to the response variable.

In other words:

The p-value for the overall F-test is 2.2 \* 10-16 which is much less than 0.05, so we will reject the null hypothesis and accept the alternative hypothesis that at least one of the terms is significant to the model.

To determine which terms are relevant we can carry out individual beta tests. The null hypothesis is that the term’s beta estimate is 0 and therefore the term is not significant. The alternative hypothesis for each term is that it doesn’t equal 0 and therefore is significant. In other words, where i is the subscript of the term we’re testing:

The p-values for the terms are: square footage: , grade: , number of bathrooms: , view 1 (tree view): , view 2 (lake view):

In each of these cases, the p-value is much lower than the significance level of 0.05, so we’d reject the null hypothesis in each case, and accept the null hypothesis. All of these terms are statistically relevant to the model.

### Making Predictions Using Model

**Scenario 1**

To find the predicted price for a home with a lake view and a 2,150 sq ft living area, grade of 7 and 3 bathrooms, we can substitute those values into the equation.

Here’s the original equation again:

And here it is with the values of the predictor variables replaced:

This evaluates to roughly $630,875.70. In New York I feel like that would be pretty cheap.

Next we’ll find the 90% prediction and confidence intervals for this house:

Table

Description automatically generated

The prediction interval for this scenario is [422,684.5 ; 838,337]. That means that 90% of the time the value of a house with these specific attributes and parameters will fall within that interval.

The confidence interval here is [610,013.7 ; 651,557.7] which means that 90% of the time the average price of a group of houses with these attributes will fall within that range.

The prediction interval is wider (significantly) than the confidence interval because it accounts for a lot of different kinds of potential error, including variation in individual data points, sampling error, etc.

**Scenario 2**

Now to find the predicted price for a house that backs out onto a road (in other words has neither a lake nor tree view), has a 2,150 sq ft living area, grade 7 and 3 bathrooms, we can just adjust the previous equation, because the only thing different is that neither view 1 nor view 2 will be relevant:

This evaluates to roughly $402,121.90.

Table

Description automatically generated

Here we have the prediction and confidence interval. The prediction interval is [194,826 ; 609,417.9] which means that 90% of the time the value of a house with these exact parameters will fall within this interval. The confidence interval is [392,274.8 ; 411,969.1] which means that 90% of the time, the mean of a group of houses with these parameters will fall within this interval.

As I said before, the prediction interval is wider than the confidence interval because it’s accounting for various errors in the sampling and estimating process.

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

### Correlation Analysis

Next we’re going to look at price against age of appliances, and price against the crime rate.

Chart

Description automatically generated Chart, scatter chart

Description automatically generated

There’s a clear negative correlation here, but there’s an obvious curve to these plots so a higher order regression model would be appropriate.

### Reporting Results

The general form of a second order model using age of appliance and crime rate as predictors is:

is the slope parameter and the subsequent terms are: appliance age, crime rate, the interaction term for appliance age:crime rate, appliance squared and crime rate squared. Because this is a higher order equation, it needs to include all interaction terms and squared terms.

In order to write the equation for this specific scenario, we need to run the model:

Table

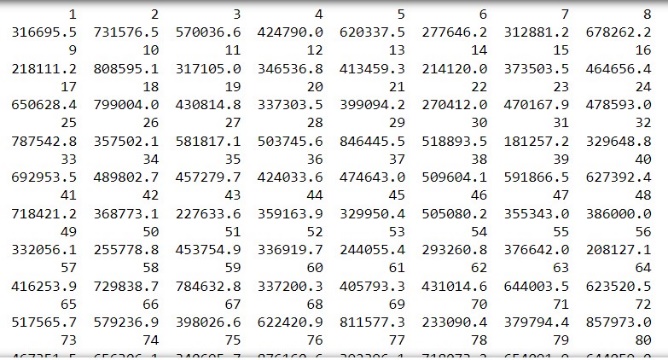
Description automatically generated

So based on this, the equation for this model is:

The R-squared and Adjusted R-squared values are 0.8088 and 0.8084 respectively. This means that approximately 80% of the variance in price can be accounted for by a model with these specific parameters.

Again it says to include tables for the residual and fitted values but I don’t know what information can be gleaned from that in this context considering that we’re plotting them so I’ll just smush them in really small.

Table

Description automatically generated 

Here are the plots for residuals vs fitted values, and the QQ plot:

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

Based on these plots, we can assume homoscedasticity and a linear relationship.

### Evaluating Significance of Model

To find the significance of the model we can do the overall F-test. The null hypothesis is that the beta estimates for each term is zero, meaning that the model is not significant. The alternative hypothesis is that at least one of the beta estimates is not zero, making the model significant.

The p-value for this model is 2.2 \* 10-16 which is much less than our significance level of 0.05, so we will accept the alternative hypothesis that at least one of the terms is significant to the model.

Now to determine which terms are significant at the 5% significance level, we’ll do individual beta tests. The null hypothesis is that the beta estimate is equal to zero, making it not relevant to the model. The alternative hypothesis is that it is not zero, making it relevant.

The p-values for the terms are: appliance age: 2.2\*10-16, crime rate: 2.2\*10-16, interaction term for appliance age and crime: 0.284, appliance age squared: 2.2\*10-16, crime rate squared: 2.2\*10-16.

The only one that is not below the significance level of 0.05 is the interaction term for appliance age and crime, so that is not relevant to the model. The others are all below the significance level so for those we would accept the alternative hypothesis that they’re relevant.

### Making Predictions Using Model

For the following two examples using appliance age and crime rate, we’ll use this general form:

Where the variables are appliance age, crime rate and the interaction term for the two, in order.

**Scenario 1**

For the first scenario it’s asking about a house with one year old appliances in an area with a crime rate of 81.02.

This evaluates to roughly $864,423.40. The confidence and prediction intervals are:

Table

Description automatically generated

The prediction interval is [711,566.6 ; 1,017,7280] which means that we can be 90% certain that a the price of a house with these specific model parameters will fall within that interval.

The confidence interval is [854,109.1 ; 874,737.7] which means that we can be 90% certain that the average price of a group of houses with these exact parameters will fall within that interval.

**Scenario 2**

The second scenario is asking for the predicted price of a house that has 15 year old appliances and a crime rate of 200.5. The equation for that would be:

This evaluates to roughly $271,051.60, which makes sense in comparison to the previous.

Table

Description automatically generated

The prediction interval here is [118,454.40 ; 423,648.80] which means that 90% of the time the price of a house with these specific model parameters will fall within that range.

The confidence interval is [265,846 ; 276,257.2] which means that 90% of the time the average price of a group of houses will fall within that interval.

## 5. Nested Models F-Test

### Reporting Results

As I mentioned above, the first order model for price using age of appliances and crime rate is:

Where the variables are appliance age, crime rate, and the interaction term between them. So this is reduced from the full version of the equation that also includes the squared terms.

Here is the model output:

Text, table

Description automatically generated with medium confidence

So based on this information, our equation for this model would be:

### Evaluating Significance of Model

To see if this reduced model is significant we’ll run an overall F-test on it. The null hypothesis will be that all of the beta estimates are equal to 0 and therefore are not significant to the model. The alternative hypothesis is that they’re not equal to zero and therefore are.

Based on the information above, the p-value for this model is 2.2\*10-16 which is much less than the significance level of 0.05, so we will accept the alternative hypothesis that at least one of the terms is relevant to the model.

To determine which one(s) are relevant we’ll do individual beta tests at the same level of significance. The null hypothesis for each is that the coefficient is 0, making the term insignificant. The alternative hypothesis is that it’s not 0, making it significant.

The p-values are: appliance age: -4.159 \* 104, crime rate: -3.418 \* 103, and the interaction term for appliance age and crime: 1.51 \* 102. These are all below the significance level of 0.05 so we can accept the null hypothesis and say that each of them is relevant to the model.

### Model Comparison

Now we want to compare this reduced model with the original one (the quadratic model), so we can perform the appropriately named nested model F-test. The reduced model in general is composed of features of the original full model. Comparing them in this instance will basically tell us if the squared terms are significant because the reduced model does not feature those terms.

The general form of the reduced model here is:

*Y*

Where the variables are appliance age, crime rate, and the interaction term for the two.

The general form for the complete model is:

Which is the same except we see the squared terms for appliance age and crime rate added at the end.

To run the nested model F-test, the null hypothesis will be that the coefficients are all 0 and that the model is insignificant. The alternative hypothesis is that at least one coefficient is not 0 and therefore it’s significant.

Table

Description automatically generated

Here’s the output for the F-test. The p-value is 2.11359 \* 10-28 which is most certainly less than the significance level of 0.05, so we can accept the alternative hypothesis that the model is significant. This means that the squared terms may be significant, so we can’t drop them and use the reduced model*.*

## 6. Conclusion

For this we did a range of models, including a first order regression, a second order model, and then built a reduced one and compared that to the complete second order model. We did that by performing a nested F-test, which told us that we couldn’t use the reduced model, because the squared terms were potentially relevant to the model.

The model of these three that I’d use to predict house prices is the complete second order model we did, because it’s more appropriate and… complete… than the first order model due to the squared terms, and the result of the nested F-test told us that we couldn’t use the reduced version of that model due to potential significance of the squared terms, so it’s really the only model that could be recommended here I think.

The practical importance here is that we can evaluate house prices with an appropriate model given the non-linear shape of some of those variable relationships, and the model itself accounts for some very good aspects that people would consider when buying a house. It seems like this is intended to be specific to Seattle, but a model like this would be appropriate for a quick ballpark estimate anywhere.