# MAT 303 Project One Report

Brandon Hobbs

Brandon.Hobbs@snhu.edu

Southern New Hampshire University

## Introduction

The data set being analyzed consists of 2691 rows and 23 columns. Each row contains data about a particular home, e.g., square footage, areal crime rate, age, etc. See Figure 1 for the first few rows for data.

The data will be used to build a three linear regression models with the purpose of predicting a home’s sales price (price) from the other data.  
  
First, the data in the csv-file will be ingested into a data frame so the R-language may be used for the stated purpose. Next, it will be plotted to provide a sense of the data and then the regression models, and their appropriateness, will be calculated. Finally, the models will be used to make predictions.

Table

Description automatically generated

**Figure 1: First 5 Rows of Data used for Analysis**

## Data Preparation

To begin the analysis the data, all 2691 rows and 23 columns, were imported into a data frame for consumption in the R-language. Of particular interest are the *price*, square footage of the living space (*sqft\_living*)*,* upstairs square footage(*sqft\_above*)*,* home age (*age*), number of baths (*bathrooms*), what the rear of the house looks out upon (*view)*, the crime rate (*crime*), and the area’s school ratings (*school\_rating*). Three regression models will be created.

The first model with be a first-order model and will try and predict the potential sales price from the quantitative variables, *sqft\_living*, *sqft\_above*, *age*, *bathrooms*, and the qualitative [0,1] variable *view*.

The second model with be a complete second-order model and will try and predict the potential sales price from the quantitative variables *crime* and *school\_rating*.

The last model with be a reduced form of the second model and will try and predict the potential sales price from the quantitative variables *crime* and *school\_rating*. This and model two will be compared using a Nested Model F-test.

## First-Order Model

This model will try and predict the sales price from 5 other variables. It will be a first-order polynomial.

### Correlation Analysis

To begin, *price* was plotted against *sqft\_living* andagainst *age* to see if any trends could be spotted visually, Figure 2.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

**Figure 2: Scatterplot of *price* versus *sqft\_living* (left); Scatterplot of *price* versus *age* (right)**

Square footage (*sqft\_living*) appears to be positively, linearly related to price. *Age*, however, is not certain – if there is a linear relationship it is weak, i.e., there is no defined trend to the data, nor does it seem to have any directionality.To determine how strongly *price* is dependent on *sqft\_living* or *age* the Pearson correlation coefficient (*R*) was computed. A positive correlation between two variables means that as one variable increases, the other variable increases as well. A negative correlation between two variables means that as one variable increases, the other variable decreases. Figure 3 provides guidance on how to interpret the magnitude of this value.

*Table

Description automatically generated*

**Figure 3: Pearson Correlation Coefficient Magnitude and its Interpretation**

Table 1 contains the Pearson correlation matrix which may be interpreted using Figure 3.

**Table 1: Pearson Correlation Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***price*** | ***sqft\_living*** | ***age*** |
| ***price*** | 1 | 0.6895 | -0.0746 |

Table 1 confirms the suspicions from visual inspection. Square footage (*sqft\_living*) is related to price with a moderate, positive, linear relationship and *age* has a very weak relationship. Table 1 does show that the relationship between *age* and *price* is negative, which makes sense, i.e., as a house ages (*age* increases) one would expect the price to fall.

### Reporting Results

The previous section showed that *price* and two of the predictor variables are linearly related but not very strongly. Due to this fact the model will be built with more variables so that price may be predicted more accurately. This model will be of the form:

Because view is qualitative, it must, first, have dummy variables created to handle all its states. The data has three possible states {0, 1, 2} so two dummy variables are needed. Table 2 shows these states and associated variables.

**Table 2: Qualitative Predictors with Road as the Reference**

|  |  |  |
| --- | --- | --- |
|  | ***view1*** | ***view2*** |
| **road** | 0 | 0 |
| **trees** | 1 | 0 |
| **lake** | 0 | 1 |

With the final model being:

With *sqft\_living* as X1, *sqft\_above* as X2, *age* as X3 , *bathrooms* as X4 , *view1* as X5 , *view2* as X6. The model says that *sqft\_living* is valuable; for example, if all variables are held constant but additional square footage is added the price will increase 129.3 units. Lake view is also very valuable, i.e., the beta term of X6 is large, so price will change 2.49E5 units if all other variables are constant.

This model has a coefficient of determination (*R2*) of 0.603 – meaning that 60.3% of the variability in sales price is explained by the predictor variables. The model also has an adjusted *R2* of 0.602. The adjusted *R2* tends to only increase when a worthwhile predictor variable is added. This value should not be used in isolation but could be used if a new predictor variable, e.g., *backyard*, was added to the model. The adjusted *R2* could evaluate if it was a valuable addition.

To further determine if the model was relevant an F-test is used. An F-test is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = β2 =…= β6=0*

*Ha: At least one βi ≠ 0 for i = 1 to 6*

The null hypothesis states that *β1* through *β6* are zero; meaning there is no correlation between *price* andthe predictor variables. The alternative states at least one beta term, *β1* through *β6,*are not zero; meaning there is a correlation between *price* and at least one predictor. This will be evaluated against an α of 5% or a 95% confidence interval. Table 3 shows the F-Test statistic and its associated P-value:

**Table 3: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 679.3 |
| P-value | 2.2E-16 |

The P-value confirms that the null value may be rejected, 2.2E-16 << 0.05; thus, at least one variable is linearly correlated to sales price. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

What the F-test does not reveal is how many of the predictor variables are relevant or which ones. To determine which variables are relevant an individual t-test is conducted on each variable. Each t-test will have a similar null hypothesis and alternative hypothesis. The null hypothesis and alternative hypothesis will be of this form:

*H0: βi =0*

*Ha: βi ≠ 0 for i = 1…n*

As before, the null hypothesis states that *βi* is zero; meaning there is no correlation between its predictor variable and *price*. The alternative states that *βi* is not zero; meaning there is a correlation between its predictor variable and *price.* Based on these hypotheses the P-values can be used to determine statistical relevance, see Table 4.

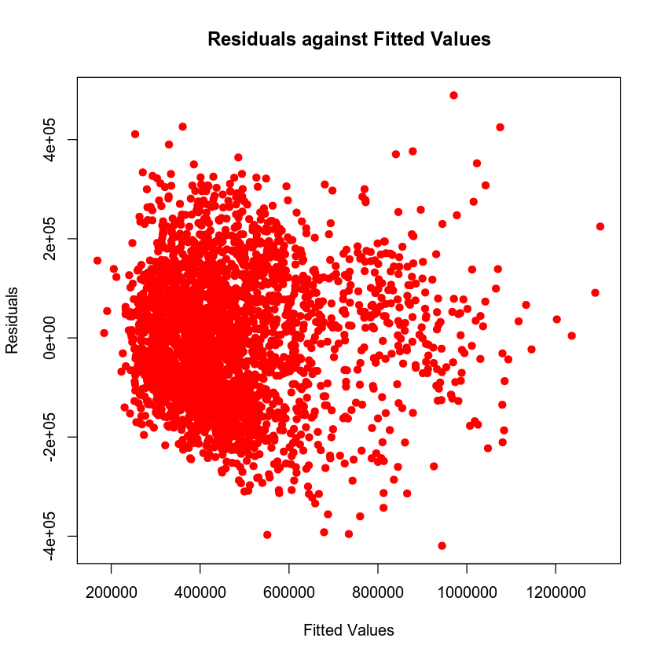
**Table 4: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| *sqft\_living* | 2.00E-16 |
| *sqft\_above* | 0.00894 |
| *age* | 2.00E-16 |
| *bathrooms* | 9.13E-13 |
| *view1* | 2.00E-16 |
| *view2* | 2.00E-16 |

All P-values are less than the 5% significance level, i.e., P-value << 0.05. Therefore, all variables, are shown to be statistically relevant.

The final test to determine whether linear regression is appropriate is to plot the residuals versus the fitted values, Figure 4. This plot allows someone to diagnostically examine if the necessary linear regression assumptions are valid for the sample data. These assumptions are:

* Mean of zero
* Independence
* Normality
* Constant variance

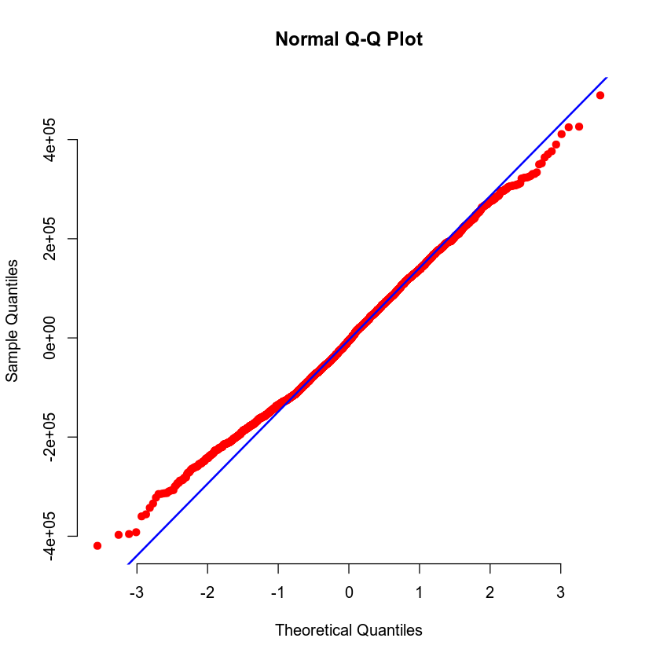


**Figure 4: Fitted Value versus Residuals**

The fitted values are the model’s predictions, and the residuals are the difference between the actual data and the model’s prediction. Figure 4 confirms the assumptions of mean of zero, data is centered around a y-value of zero, and constant variance, there are no “fan-shaped” patterns. Constant variance is also known as homoscedasticity – Figure 4 confirms the data is homoscedastic.

An independence test is not needed as there are no time varying variables in the data set.

Normality is confirmed with a QQ plot, Figure 5.



**Figure 5: QQ Plot of Residuals**

A QQ plot confirms normality if the residuals lie on the blue line. If the points deviate “significantly” from the diagonal line, then the assumption of normality is violated. The residuals do not appear to deviate and thus normality is confirmed.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, the expected sales price for a home with 2,150 ft2, 1,050 ft2 upper story, is 15 years old, has 3 baths, and opens to a view of the road is 459,828.20 units.

While a home with 4,250 ft2, 2,100 ft2 upper story, is 5 years old, has 5 baths, and opens to a view of the lake is 1,074,285 units.

However, these sales values represent the mean sales price, someone can be 90% confidant that the actual average sales price is in the range of 446,087.90 – 473,568.50 or 1,045,117.00 – 1,103,454.00 for all homes sold with those attributes, respectively.

Due to the uncertainty in estimating the mean value and the random variation in what was already observed someone could be certain, to 90%, that a single home with the same attributes will have a sales price between 23,9563.00 – 680,093.40 or 852,522.60 – 1,296,048.00, respectively. This wider range is known as the prediction interval. The prediction interval is wider because it considers the variability of the individual points around the predicted mean in addition to the uncertainty in sampling.

The ranges and means are inline with what most people would anticipate. As the house becomes larger, has more amenities, is younger, and has a better view the more valuable the home should be.

## Second-Order Regression Model Using *crime* and *school\_rating*

The second model with be a complete second-order model and will try and predict the potential sales price from the quantitative variables *crime* and *school\_rating*.

### Correlation Analysis

To begin, *price* was plotted against *crime* andagainst *school\_rating* to see if any trends could be spotted visually, Figure 6.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

**Figure 6: Scatterplot of *price* versus *crime* (left); Scatterplot of *price* versus *school\_rating* (right)**

From visual inspection the both *crime* and *school\_rating* data appears to be non-linear, first-order would inappropriate, and have an upwards concavity. The upwards concavity would suggest that the model should have positive second-order terms.

### Reporting Results

The previous section showed that *price* and *crime* or *school\_rating* appear to be non-linearly related. This suggest that a higher order variable is needed. This model will be of the form:

With the final model being:

With *crime* as X1 and *school\_rating* as X2. The positive (+)6.377 and (+)1.165E4 confirms the suspicion from visual inspection of upwards concavity.

This model has a coefficient of determination (*R2*) of 0.809 – meaning that 80.9% of the variability in sales price is explained by the predictor variables. The model also has an adjusted *R2* of 0.808. The adjusted *R2* tends to only increase when a worthwhile predictor variable is added. This value should not be used in isolation but could be used if a new predictor variable, e.g., *backyard*, was added to the model. The adjusted *R2* could evaluate if it was a valuable addition.

To further determine if the model was relevant an F-test is used. An F-test is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = β2 =…= β5=0*

*Ha: At least one βi ≠ 0 for i = 1 to 5*

The null hypothesis states that *β1* through *β5* are zero; meaning there is no correlation between *price* andthe predictor variables. The alternative states at least one beta term, *β1* through *β5,*are not zero; meaning there is a correlation between *price* and at least one predictor. This will be evaluated against an α of 5% or a 95% confidence interval. Table 5 shows the F-Test statistic and its associated P-value:

**Table 5: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 2272 |
| P-value | 2.2E-16 |

The P-value confirms that the null value may be rejected, 2.2E-16 << 0.05; thus, at least one variable is linearly correlated to sales price. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

What the F-test does not reveal is how many of the predictor variables are relevant or which ones. To determine which variables are relevant an individual t-test is conducted on each variable. Each t-test will have a similar null hypothesis and alternative hypothesis. The null hypothesis and alternative hypothesis will be of this form:

*H0: βi =0*

*Ha: βi ≠ 0 for i = 1…n*

As before, the null hypothesis states that *βi* is zero; meaning there is no correlation between its predictor variable and *price*. The alternative states that *βi* is not zero; meaning there is a correlation between its predictor variable and *price.* Based on these hypotheses the P-values can be used to determine statistical relevance, see Table 6.

**Table 6: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| *crime* | 1.90E-09 |
| *crime2* | 2.00E-16 |
| *school\_rating* | 0.000406 |
| *school\_rating2* | 2.00E-16 |
| *school\_rating\*crime* | 0.281513 |

All P-values are less than the 5% significance level, i.e., P-value << 0.05, except the interaction term, *school\_rating\*crime.* Therefore, all variables, except the single interaction term, are shown to have a statistically relevant. Removing the *school\_rating\*crime* interaction term and recomputing the model should be evaluated but will be left for another day.

As before, to test if linear regression is appropriate the residuals versus the fitted values were plotted, Figure 7. Using the same assumptions from before:

* Mean of zero
* Independence
* Normality
* Constant variance

Chart, scatter chart

Description automatically generated

**Figure 7: Fitted Value versus Residuals**

The fitted values are the model’s predictions, and the residuals are the difference between the actual data and the model’s prediction. Figure 7 confirms the assumptions of mean of zero, data is centered around a y-value of zero, and constant variance, there are no “fan-shaped” patterns. Constant variance is also known as homoscedasticity – Figure 7 confirms the data is homoscedastic.

An independence test is not needed as there are no time varying variables in the data set.

Normality is confirmed with a QQ plot, Figure 8.

Chart, line chart

Description automatically generated

**Figure 8: QQ Plot of Residuals**

A QQ plot confirms normality if the residuals lie on the blue line. If the points deviate “significantly” from the diagonal line, then the assumption of normality is violated. The residuals do not appear to deviate and thus normality is confirmed.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, the expected sales price for a home in an area with 81.02 crime reports per 100,000 and schools having a rating of 9.8 is 874,497.00 units.

While a home in an area with 215.5 crime reports per 100,000 and schools having a rating of 4.28 is 199,706.70 units.

However, these sales values represent the mean sales price, someone can be 90% confidant that the actual average sales price is in the range of 863,681.4 – 885,312.7 or 191,753.5 – 207,659.9 for all homes sold with those attributes, respectively.

Due to the uncertainty in estimating the mean value and the random variation in what was already observed someone could be certain, to 90%, that a single home with the same attributes will have a sales price between 721,606.2 – 1,027,388 or 46,991.65 – 352,421.7, respectively. This wider range is known as the prediction interval. The prediction interval is wider because it considers the variability of the individual points around the predicted mean in addition to the uncertainty in sampling.

These ranges and means are inline with what most people expect. As crime increases and schools worsen, the value of the home should drop.

## Reduced Form Model and Nested Tests

This last model with be a reduced form of the second model and will try and predict the potential sales price from the quantitative variables *crime* and *school\_rating*. This and model two will be compared using a Nested Model F-test.

A nested models F-test considers *a subset* of the regression parameters other than the intercept parameter, *β0.* For this model 2 will be reduced to the first-order factors and the interaction term only.

The model will be of this form:

The final model becomes:

With crime as X1 and school\_rating as X2.

This model has an R2 value of 0.799 meaning that 79.9% of the sales price variability is accounted for.

As for the previous models, an F-test and individual t-tests were conducted to evaluate the whole model and individual terms. The F-test will have the same null and alternative hypotheses as shown for the prior models:

*H0: β1 = β2 = β3=0*

*Ha: At least one βi ≠ 0 for i = 1 to 3*

The t-tests will also have the same null and alternative hypotheses as the prior models:

*H0: βi =0*

*Ha: βi ≠ 0 for i = 1…n*

Tables 7 and 8 show the F-test and individual t-tests, respectively.

**Table 7: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 3573 |
| P-value | <2.2E-16 |

**Table 8: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| *crime* | <2E-16 |
| *school\_rating* | <2E-16 |
| *school\_rating\*crime* | <2E-16 |

The P-value for the F-test, 2.2E-16 << 0.05, confirms at least one beta is not zero, i.e., reject the null hypothesis.

All P-values from the t-tests are less than the 5% significance level, i.e., P-value << 0.05. Therefore, all variables, are shown to be statistically relevant.

**Model Comparison**

As mentioned before, Model 3 is a reduced form of Model 2 – the second-order terms were removed.

Complete second-order model:

Reduced model:

Both models had F-tests that proved relevancy, so which model should be used?

To determine if the additional, second-order terms should be added a Nested F-test is run. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = β2 = β3=0*

*Ha: At least one βi ≠ 0 for i = 1 to 3*

The null and alternative hypotheses are trying to state the same conclusions as for all other models: the null states that the slope parameters for the reduced model are zero; the alternative states that at least one is not zero. The results of the Nested F-test are shown in Table 9.

**Table 9: Nested F-test Comparing the Complete Second-Order Model and the Reduced Model**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Residual**  **Degrees of Freedom** | **F-Value** | **P-Value** |
| Model 3 | 2688 | NA | NA |
| Model 2 | 2686 | 65.20513 | <2.3E-28 |

Because the P-value is less than the 5% level of significance, i.e., 2.3E-28 << 0.05, the null hypothesis can be rejected, meaning the additional predictors should be added.

## Conclusion

Three models were built to try and accurately predict the sales price of a home. The first model used only first-order terms, the second was a complete, second-order model, and the third was a reduced form (eliminated the second-order terms) of the second model. All three models were shown to be statistically relevant via their overall F-tests, but not all terms of the individual models were, i.e., some predictor terms were shown to be statistically irrelevant.

Furthermore, the *R2* values for each of the models were quite varied. Table 10 shows the *R2* values of the three models.

**Table 10: Each of the Three Models and Their *R2* Values**

|  |  |
| --- | --- |
| Model | R2 Value |
|  | 0.603 |
|  | 0.809 |
|  | 0.799 |

The second model accounts for more variability than the other models. Model 2 was also shown to be more valuable than its reduced form. Therefore, Model 2 should be chosen as the preferred way to predict a home’s sales price based on this data and the work contained in this document.

Model 2 could help a real estate broker or sales team with any of their comparative analyses for clients or a broader market. It could also help them to determine if they are outperforming or underperforming the historic market.

## Citations

Hobbs, B. (2022). *MAT 303 module one summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 303 module two summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 303 module three summary report*. [Unpublished report]. SNHU.