MAT 303 Project One Summary Report

Creston Getz
Southern New Hampshire University

1. Introduction

The dataset we are using in this analysis is called “housing.” It contains various variables we will use to create three regression models to predict the sale price of homes. We can derive many useful relationships from this data, such as how the number of bedrooms affects the price of a home. The results of this report will be used to set better prices for clients. We will complete three types of analysis in this report: a first order regression model with both quantitative and qualitative variables, a second order regression model with only quantitative variables, and a nested F-test for the models. The nested F test will help us determine what should be kept in the model.

2. Data Preparation

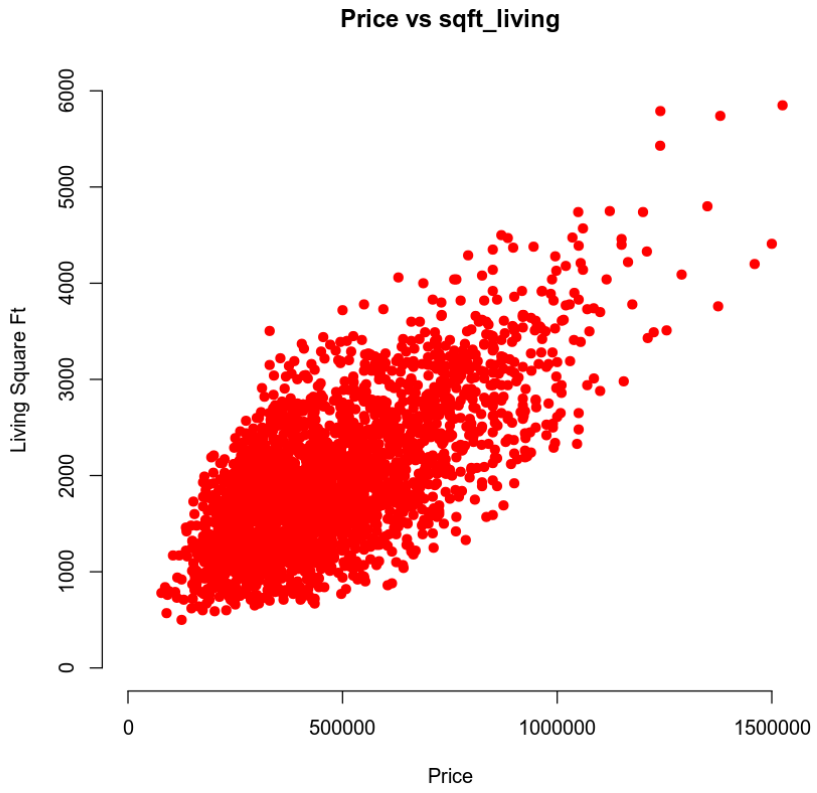
There are many important variables in this dataset. The actual names of the variables are in italics. You can view the full table of the main variables in the notebook. The main variables we will be working with in this report are *price*, which is the sale price of the home. *Sqft_living* which is the square footage of the living area. *Sqft_above* is the upper-level living area square footage. *Age* which is the age of the home (presumably in years). *Bathrooms* tells us how many bathrooms the home has. *View* is a qualitative variable which tells us if the home backs out to a lake, trees, or a road. *School_rating* gives an average rating of the schools in the area. *Crime* measures the crime rate per 100,000 people. These are all the variables we will be using in our regression models.

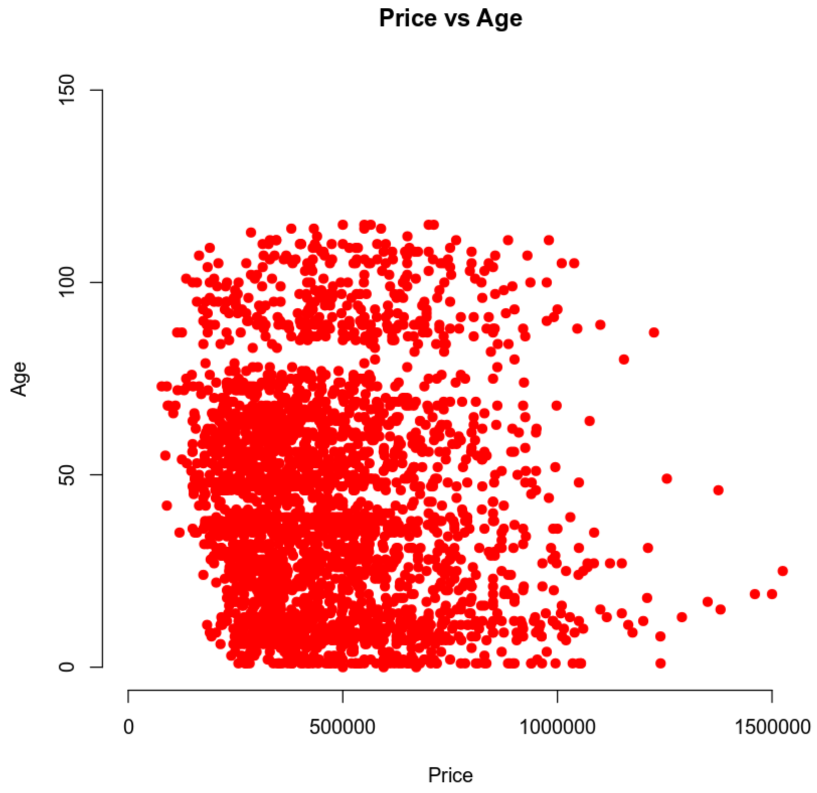
In this report we will refer to the names of the variables as their full names for readability. For example, *sqft_living* and the living area square foot mean the same thing. When we say *crime*, we are referring to the crime rate per 100,000 people. This dataset has 23 columns and 2692 rows of data.

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

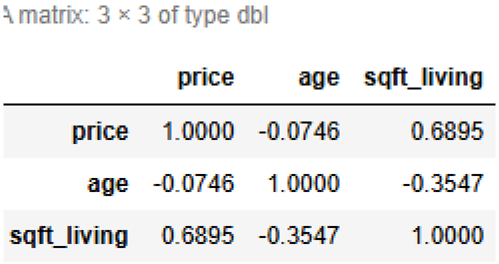
Correlation Analysis

In this section we are going to create two scatterplots of age and living area square footage with price to find any trends. The scatterplot of price vs the living area square footage shows a clear linear trend. As the price increases, so does the square footage of the living area. This trend is not true for price vs age. Based on that scatterplot, there does not seem to be any linear trend. There are brand new homes at the same price as 100-year-old homes and vice versa. However, the more expensive homes are in the newer range.


Price vs sqft_living
Living Square Ft
0
1000
2000
3000
4000
5000
6000
0
500000
1000000
1500000
Price


Price vs Age
Age
0
50
100
150
0
500000
1000000
1500000
Price

Next, we are going to report the correlation coefficients between these variables to mathematically get their correlation. Our Pearson correlation matrix below lines us with our observations from the scatterplots. The slight negative relationship between age and price is very weak. However, it is there and as age increases, prices will decrease marginally. Price and living area square footage, however, have a strong positive correlation. Larger homes or ones with larger living areas will have higher prices. Also, the negative moderate correlation between age and living area square footage shows that older homes tend to have smaller living areas.



\ matrix: 3 × 3 of type dbl

	price	age	sqft_living
price	1.0000	-0.0746	0.6895
age	-0.0746	1.0000	-0.3547
sqft_living	0.6895	-0.3547	1.0000

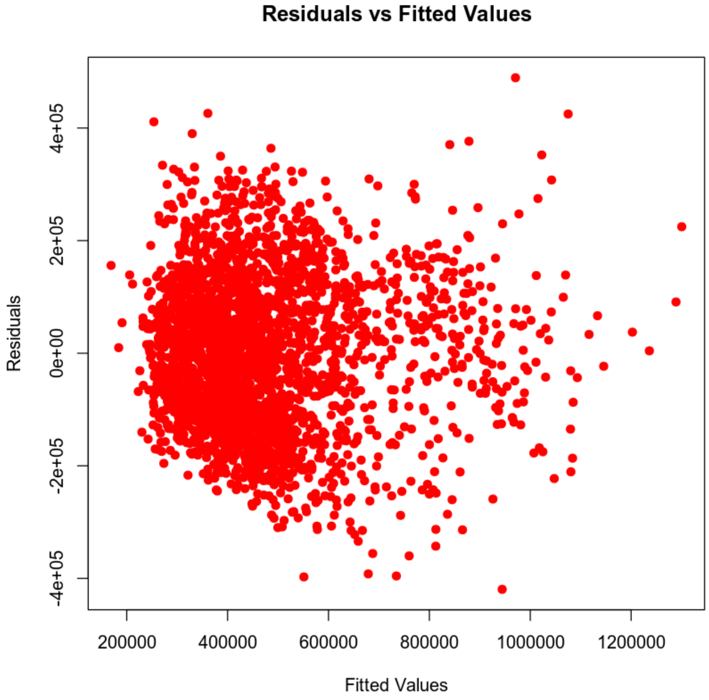
Reporting Results

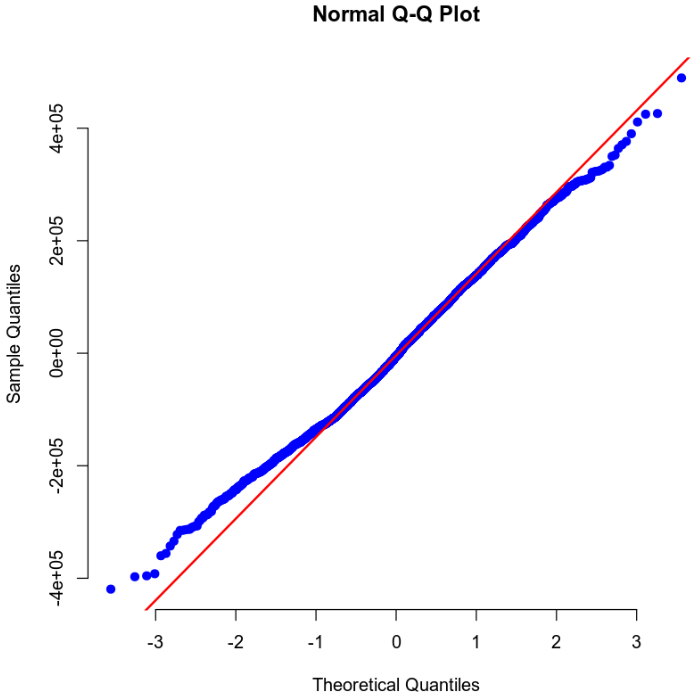
In this section, we will create a regression model to predict price using living area, upper-level area, age, and bathroom, and view as the predictors. This model's general form will look like: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$. And the prediction equation will look like: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \widehat{\beta_3} x_3 + \widehat{\beta_4} x_4 + \widehat{\beta_5} x_5 + \widehat{\beta_6} x_6$. Where y is the predicted price of the home, x1 is living area, x2 is upper-level area, x3 is age, x4 represents the number of bathrooms, and x5 and x6 represent the house's view. The base level for view is when both x5 and x6 are 0. If x5 is 1 then the house backs out to trees, 0 if not. And if x6 is 1 then the house backs out to a lake, 0 if not. $\widehat{\beta_i}$ where i is the number of predictors are estimates of, β_i respectively.

Using the output from R and our prediction equation above, we can create the model. $\Pr ice = 7709 + 129.3x_1 + 19.51x_2 + 1451x_3 + 43{,}970x_4 + 167{,}500x_5 + 249{,}000x_6$. The R-squared for this model is 60.26% and the adjusted R-squared is 60.02%. The adjusted R-squared is remarkably close to the R-squared value, meaning that the number of predictors in this model is good and that they are all useful. A R-squared value of 60% is a moderate fit; however, there is a lot of uncertainty when it comes to housing prices. So, our model may have much lower R-squared values compared to engineering or physics. These R-squared values tell us the amount of variation in price. So about 60% of the price variation is from living and upper area square footage, age, bathrooms, and view.

The beta estimates for the living area and lake view are not the same. Living area is a quantitative variable, so the price will change 129.3 per 1 unit of change of living area square footage. The qualitative or categorical predictor view represents whether the house back out to a lake, trees or a road. The values for x5 and x6 will be 0 or 1. This is known as dummy coding which a way to use categorical variables in models. It is not used per 1 unit change like living space, the house backs to a lake, or it does not. The effect on price only happens when the category is present or true. View2 in our model will be 0 if there is no lake, or 1 if there is a lake, it is true or false. The price either increases by 249,000 if the house has a lake, or the price is unaffected if the house does not have a lake.

Next is a scatterplot of the residual vs fitted values and a Q-Q plot. The plots can be used to determine whether our regression model is appropriate and helps us validate assumptions. Our Q-Q plot shows our points follow the line very closely in the middle, with some tailing on the end. This tells us that the residuals are about normal, which for our model is acceptable. The scatterplot shows a fan shape, and the residuals are not randomly placed. This tells us that heteroscedasticity is present in this model, or that the variance is not constant. We would ideally want a random scatter of points with constant spread across the x-axis, which is not shown in this plot. We will keep this in mind for the rest of our analysis, but heteroscedasticity may affect our confidence intervals and p-values.





Evaluating Significance of Model

To determine if our model is statistically significant, we will perform an overall F test and Individual t-tests to test each predictor variable. For each test, we are going to use a 5% level of significance. First the F –test, our null hypothesis is that all the slope parameters are equal to zero or mathematically H_0: $\beta_1 = \beta_2 = \beta_n = 0$ where n is the number of regression parameters (our model has 6). The alternative hypothesis is that at least one parameter is not zero or mathematically H_a: *at least one* $\beta_i \neq 0$ for i = 1,2,3,4,5,6. The p-value for our F-test is 2.2e-16 which is much lower than 0.05. We reject the null hypothesis in favor of the alternative hypothesis. There is a significant linear relationship between the response variable and the predictor variables.
To determine which of the predictors are statistically significant, we will perform an individual t-test again using a significance level of 5%. We will use the same null and alternative hypothesis for each test. The null is that the regression parameter is 0 or H_0: $\beta_i = 0$ and the alternative hypothesis is that the regression parameter is not 0 or mathematically H_a: $\beta_i \neq 0$. The p-value for sqft_living is 2e-16, we reject the null hypothesis in favor of the alternative hypothesis. The p-value for sqft_above is 0.00896, so we reject the null hypothesis in favor of the alternative hypothesis. Age's p-value is 2e-16 so again we reject the null hypothesis in favor of the alternative hypothesis. Bathroom's p-value is 9.13e-13, so we reject the null in favor of the alternative hypothesis. The p-value for both view1 and view2 is 2e-16. So, we reject the null hypothesis in favor of the alternative for each. All the predictor's p-values are lower than the significance level. Meaning that the regression parameters are significantly different from 0. There is a significant relationship between each predictor variable and the price.

Making Predictions Using Model

Using our model, the predicted price for a home with 2150 sq ft living space, 1050 sq ft upper living space, is 15 years old, has 3 bathrooms, and backs out to a road is 459,828.2. The 90% prediction interval for this home is (239,563-680,093.4). We are 90% confident that a new home with the same predictors (living space, age, bathrooms, and view) would fall into this range. The 90% confidence interval for this home is (446,087.9-473,568.5), meaning we are 90% confident that the average price for a home with the predictors given falls into this range.

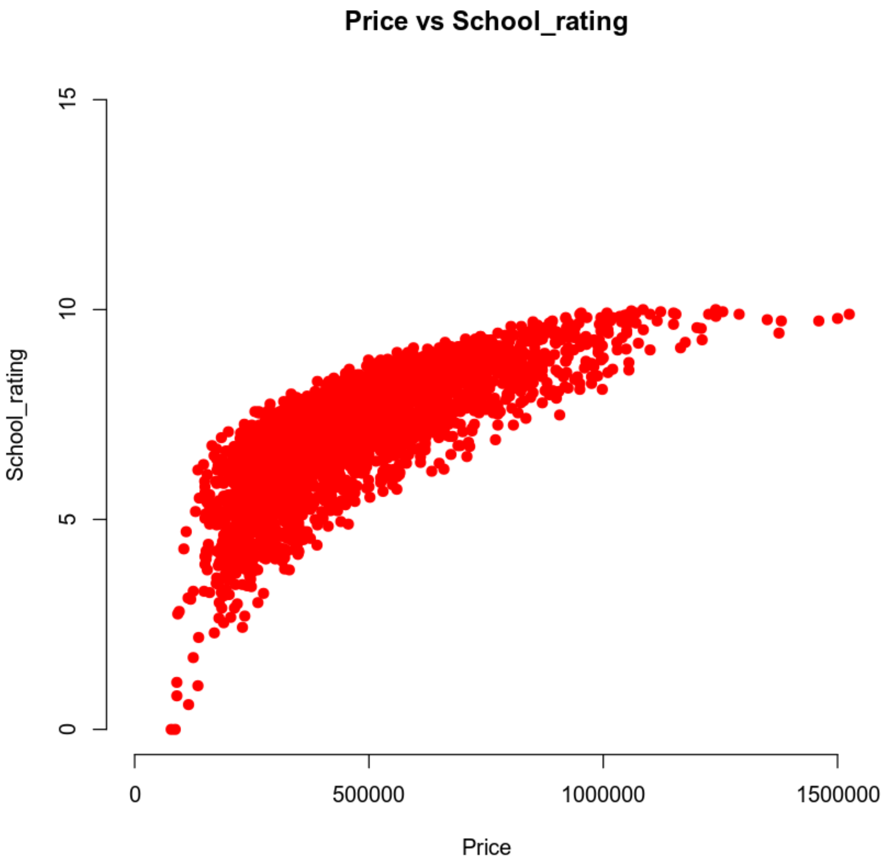
The predicted price for a home with 2150 sqft living area, 2100 upper living area, is 5 years old, has 5 bathrooms, and backs out to a lake is 1,074,285. The 90% prediction interval is (852,522.6-1,296,048), so we are 90% confident that a new home with the same predictors will fall into this range. The confidence interval for this home is (1,045,117-1,103,454). We are 90% confident that the true average price of a home with the same predictors will fall into this range.

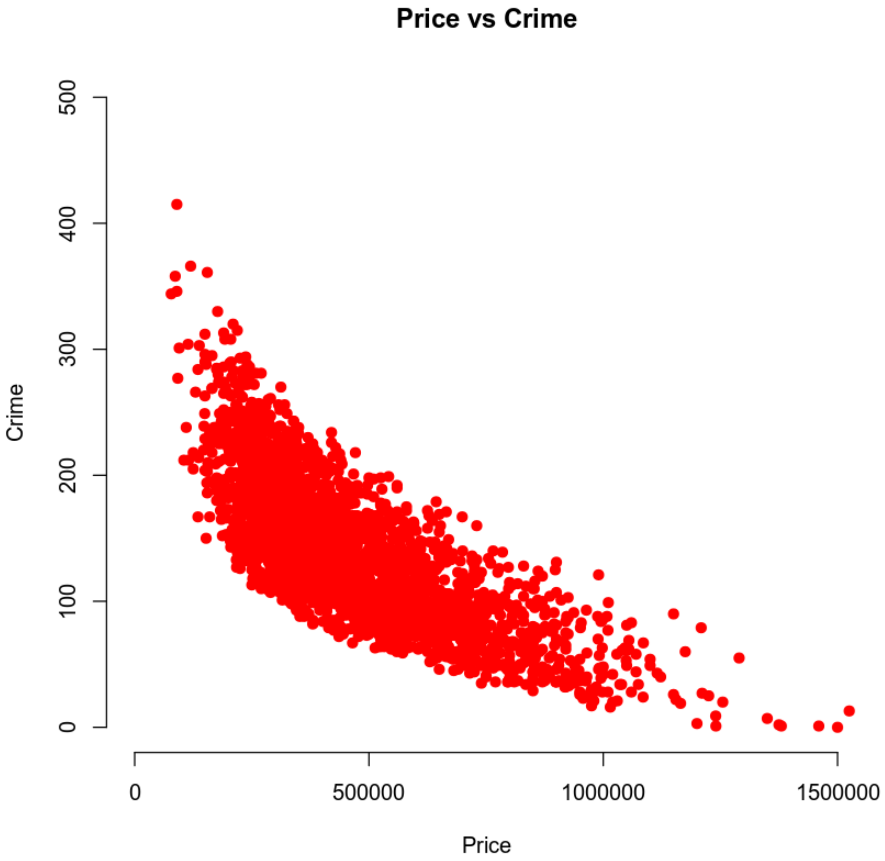
In both homes we predicted the prediction interval was wider than the confidence interval. This is mostly because the prediction interval must account for significantly more variability. There are many variables our model does not account for, and the prediction interval must account for that when using our predictors to predict the price of a home. The confidence interval is looking for the average price of the home which can account for less uncertainty.

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

Correlation Analysis

For our second model we will be using school rating and crime. The scatterplots below show that relationships between price, school rating and crime are non-linear. Because of this, we should create a 2nd order model using these variables. As school ratings increase, price increases but maxes around a school rating of 10. As for crime, as it decreases, the price of the home increases. Both variables have a strong relationship with price.



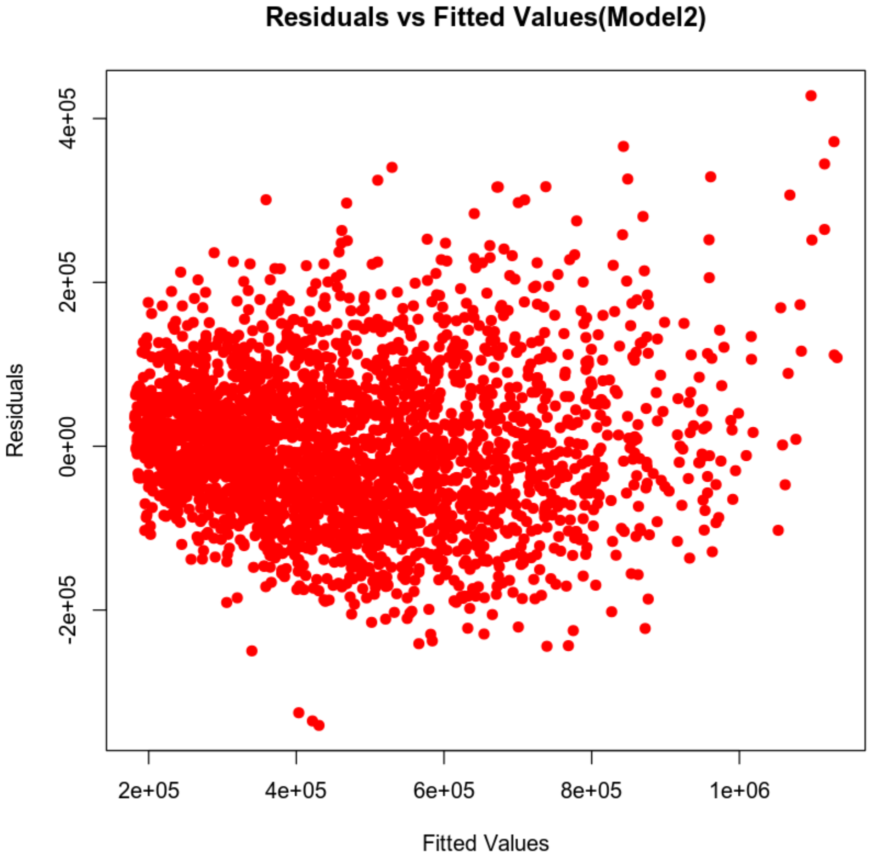


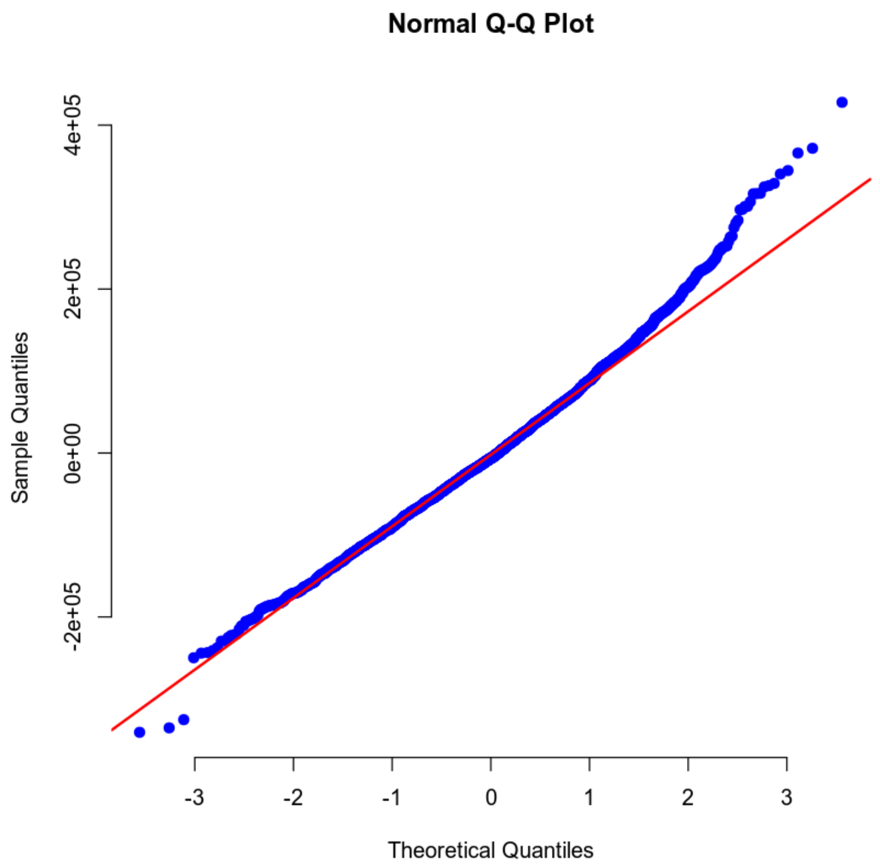
Reporting Results

In this section, we are going to create a second order regression model using school_rating and crime as predictors with price as the response. A second order model includes the interaction term for all terms and the squared values for each term. The general form for this second order regression model is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$. The prediction equation is: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \widehat{\beta_3} x_1 x_2 + \widehat{\beta_4} x_1^2 + \widehat{\beta_5} x_2^2$. Where y is the predicated value of price, x1 is school rating and x2 is crime. $\widehat{\beta_i}$ where i is number of betas are estimates of β_i.
After making our model in R we finish our prediction equation which is:
$price = 733{,}900 - 73{,}750 school - 3{,}155 crime + 11{,}650 school^2 + 637.7 crime^2 - 52.27(crime \cdot school)$.

The value of R squared is 80.88% while the adjusted R squared is 80.84%. The adjusted R squared is almost identical to the normal R squared which tells us that the number of predictors in the model is good and that they are all doing something in the model. Both values of R squared tells us that about 80% of the variance in price can be explained by this model (a second order model with school rating and crime). 80% indicates a good fit, which is better than the 60% from our first model.

Below is a scatterplot of our model's fitted values and residuals and a Q-Q plot. The data appears mostly normal based on the Q-Q plot. Most of the points fall onto the line in the middle with some tailing at the ends. However, like the first model, homoscedasticity may not hold true for this model either. While the points in the scatterplot look randomly placed about the y axis, on the x axis the data is clearly left clustered. This may indicate that heteroscedasticity is present. We will keep this in mind for the rest of our analysis.





Evaluating Significance of Model

It is important to see if our model is significant by running a hypothesis test which we will do in this section. We will perform an overall F-test for the entire model, then run individual t-tests for each predictor variable. For every test, we will use a 5% level of significance. The null hypothesis for the F test is that all regression parameters are equal to zero; there would be no relationship between the predictor and response variables. Mathematically this is $H_0\colon \beta_n = 0$. The alternative hypothesis is that at least one regression parameter is not equal to zero, or at least one variable is related to the predictor. Mathematically this is $H_a\colon at\ least\ one\ \beta_i \neq 0$. The p-value for the overall F test is 2.2e-16, which is a lot lower than 5%. We reject the null hypothesis in favor of the alternative hypothesis. Our model is significant; at least one predictor has a significant relationship to the price of a home. Which based on the scatterplots of school rating, crime, and price, sounds right.

Next, we will perform an individual t-test for each predictor variable using a 5% level of significance. Each T-test will use the same null and alternative hypothesis. The null hypothesis is the regression parameter for that predictor variable is 0 or $H_0\colon \beta_i = 0$. The alternative hypothesis is that the parameters are not zero or $H_a\colon \beta_i \neq 0$. If the null is more likely, the variable has no effect on the home price. But if the alternative hypothesis is true then it has some effect. The p-value for school_rating is 0.000406 which is lower than our significance level, so we reject the null hypothesis in favor of the alternative hypothesis. Crime's p-value is 1.90e-09, again we reject the null in favor of the alternative hypothesis. Both school_rating and crime on their own are significant to the model. Next, the quadratic terms (school rating and crime squared) for school rating and crime are both 2e-16 which is a lot lower than 5%. For both, we reject the null hypothesis in favor of the alternative hypothesis. The interaction term between crime and school rating's p-value is 0.281513 which is higher than 5% so we fail to reject the null hypothesis. The interaction term is the only predictor variable in our model that is not statistically significant. We may consider removing it.

Making Predictions Using Model

Using our model, the predicted price of a home with a 9.80 school rating and a crime rate of 81.02 is 874,497. The 90% prediction interval is (721,606.2-1,027,388). Meaning we are 90% confident that a new home with the same ratings will fall into this range. Again, using our model, the predicted price for a home with a 4.28 school rating and a 215.50 crime rating is 199,706.7. This makes sense that a home with a higher crime rate and lower school rating costs a lot less. The 90% prediction interval is (46,991-352,421.7). This is a very wide prediction interval ranging from a home valued at about 46 thousand to 350 thousand. We are 90% sure a new home with these ratings will fall into this range. The 90% confidence interval for this home is (191,753.5-207,659.9), meaning we are 90% confident that the average price of a home with the same ratings will be in this range.

5. Nested Models F-Test

Reporting Results

In the next few sections, we will be performing a nested model F-test. We will use this test to compare the predictor variables in the second model we created in this report and determine which is more suitable for our needs. First, we need a model to compare the second order model with. In the notebook we created a first order model with school-rating and crime as predictors. The general form for this model is: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. The prediction equation is: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \widehat{\beta_3} x_1 x_2$. Where y is the predicted value for price, x1 is school rating, and x2 is the crime rate. $\widehat{\beta_i}$ where i is the number of betas (1 and 2) are the estimates of β_i. After we make the model in R, we can finish the prediction equation. This model does not include the squared terms. Its prediction equation is: $price = -410{,}233.37 + 155{,}559.97 x_1 + 2230.07 x_2 - 564.85 x_1 x_2$.

Evaluating Significance of Model

Next, we must determine if our model is significant by doing an overall F-test (not to be confused with a nested F-test) and individual t-tests. For each test, we will use a 5% level of significance. The null hypothesis for the F-test is that all regression parameters are equal to zero or mathematically H_0: $\beta_n = 0$. The alternative hypothesis is that at least one of the regression parameters in the model is not equal to zero or H_a: $\beta_i \neq 0$. The p-value for the F-test is 2.2e-16, which tells us that the model is statistically significant. We reject the null hypothesis in favor of the alternative hypothesis.

Now we will perform a few individual t-tests to determine which predictors in the model are significant if any. For each test, our null hypothesis is that the beta or regression parameter is equal to zero, mathematically $H_0\colon\ \beta_i\ =\ 0$. While the alternative hypothesis is $H_0\colon\ \beta_i\ \neq\ 0$, the regression parameter is not equal to zero. Interestingly, every predictor in this model has the same p-value of 2e-16. So, for school_rating, crime, and the interaction term we reject the null hypothesis in favor of the alternative hypothesis. Each predictor is useful for the model.

Model Comparison

In this section we will use the model we just created and the 2nd order model to perform a nested F-test. When comparing two models, the complete model is the original model with more predictors. And the reduced model has every predictor in it other than the ones we want to compare. The reduced model is a subset of the full model, and we must remove at least one predictor. The complete model in this report is the 2nd-order model we made. The reduced is the model we just made with no squared terms.

The general form for the reduced model is: $E(y)\ =\ \beta_0\ +\ \beta_1 x_1\ +\ \beta_2 x_2\ +\ \beta_3 x_1 x_2$. The prediction equation for the reduced model is $\hat{y}\ =\ \widehat{\beta_0}\ +\ \widehat{\beta_1} x_1\ +\ \widehat{\beta_2} x_2\ +\ \widehat{\beta_3} x_1 x_2$, where y is the predicted value of price, x1 is school rating, and x2 is crime. $\widehat{\beta_\iota}$ where i is the number of parameters is estimates of β_i. The general form for the complete model is: $E(y)\ =\ \beta_0\ +\ \beta_1 x_1\ +\ \beta_2 x_2\ +\ \beta_3 x_1 x_2\ +\ \beta_4 x_1^2\ +\ \beta_5 x_2^2$. The prediction equation for the complete model is: $\hat{y}\ =\ \widehat{\beta_0}\ +\ \widehat{\beta_1} x_1\ +\ \widehat{\beta_2} x_2\ +\ \widehat{\beta_3} x_1 x_2\ +\ \widehat{\beta_4} x_1^2\ +$ $\widehat{\beta_5} x_2^2$. Where y is the predicted value of price, x1 is school rating, and x2 is crime. $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$, $\widehat{\beta_3}$, $\widehat{\beta_4}$, $\widehat{\beta_5}$ are estimates of β_1, β_2, β_3, β_4, β_5.

Our nested model F-test is trying to determine if the quadratic terms for school_rating and crime should be kept. It does this by comparing the reduced model to the complete model with the terms we want to test. We are using a 5% or 0.05 level of significance for our test. Our null hypothesis is that the additional parameters in the complete model are all equal to 0, or the two squared terms betas are equal to zero. Mathematically the null hypothesis is $H_0\colon\ \beta_j\ =\ \beta_{j+1}\ =\ \beta_n\ =\ 0$. The alternative hypothesis is that at least one additional regression parameters are not equal to zero. Mathematically, $H_a\colon\ at\ least\ one\ \beta_i\ for\ i = j, \ldots., n$ where is the number of parameters that have been split into the two groups. If the null hypothesis is not rejected, the reduced model is enough, and we do not need to add the quadratic terms. But if the alternative is more likely, then the quadratic terms should be added. And the only two regression parameters we have for this test are the ones for school_rating squared and crime squared. The p-value for nested model F-test is 2.22716e-28 which is astronomically smaller than our significance level of 5%. We reject the null hypothesis in favor of the alternative hypothesis. This means that at least one of the two quadratic terms should be kept in the model.

6. Conclusion

The model I would choose to best predict home prices is the second model. This model has a higher R squared value than the first model and a lower residual standard error or RSE (the difference between fitted values and residuals). The second model also includes the quadratic terms for school_rating and crime which we added because of the curved relationship found in the scatterplots. We justified adding the squared terms by performing a nested model F-test to determine if the squared terms should be kept. In addition to that, the second model still has a higher R squared and a lower RSE than the third model, which is a reduced version of the second model with no squared terms.

The analyses performed in this report have a lot of practical uses. Most notably by using the second model we can predict the price of homes in an area. Data on crime rates and school ratings can be gathered in mass for various locations. Using the second model we can predict the average price of homes in given areas without knowing the sq ft of the home, the upper living area, the number of bathrooms, and so on. This can be an immensely powerful analysis tool for our real estate company to set listing prices more accurately for our clients. It also, as mentioned, allows us to predict the price of homes we are trying to buy. This model could be particularly useful for an agent to use to get an idea of the house prices in an area.