# MAT 303 Project One Summary Report

Andrea Plunkett Jackson

[andrea.jackson11@snhu.edu](mailto:andrea.jackson11@snhu.edu)

Southern New Hampshire University

## 1. Introduction

In this project, I will be exploring a comprehensive historical dataset containing various attributes of houses, such as square footage, number of bathrooms, and other relevant factors, along with their corresponding selling prices. The primary objective is to develop regression models that can accurately predict a house’s selling price based on these attributes. The results of these analyses will be instrumental in helping our real estate company set more accurate and competitive prices when listing homes for clients. To achieve this, I will be conducting several types of regression analyses using the R programming language. These analyses include a First Order Regression Model that incorporates both quantitative and qualitative variables, a Complete Second Order Multiple Regression Model with quantitative variables, and Nested Models F-Test to compare the fit of different models. By interpreting the findings and describing their practical implications, I aim to provide valuable insights that will inform our pricing strategies and enhance our company’s market positioning.

## 2. Data Preparation

I am tasked with analyzing a dataset that encompasses several key variables that are pivotal in predicting the sale price of homes. The dataset includes the following significant variables:

* Price: This represents the sale price of the home.
* Bedrooms: This signifies the number of bedrooms in the home.
* Bathrooms: This denotes the number of bathrooms in the home.
* Sqft\_living: This is the size of the living area, measured in square feet.
* Sqft\_above: This is the size of the upper level, also measured in square feet.
* Sqft\_lot: This is the size of the lot, measured in square feet.
* Age: This represents the age of the home.
* Grade: This is a measure of the craftsmanship and the quality of materials used in the construction of the home.
* Appliance\_age: This is the average age of all appliances in the home.
* Crime: This represents the crime rate per 100,000 people in the area.
* Backyard: This indicates whether the home has a backyard (1) or not (0).
* School\_rating: This is the average rating of schools in the area.
* View: This indicates the view from the home, with values representing a lake (2), trees (1), or a road (0).

The dataset is comprised of 23 columns and 2692 rows, providing a substantial amount of data for my regression analyses. My aim is to analyze these variables to develop robust models that can accurately predict home prices. This will assist our company in setting competitive and realistic listing prices for our clients.

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

### Correlation Analysis

The scatterplot, which plots price against living area in square feet, offers significant insights into the correlation between a house’s price and its living area. This graphical representation illuminates how the price of a house is influenced by its living area, also measured in square feet. A key observation is the positive correlation between these two variables. This suggests that as the living area of a house increases, its price tends to rise correspondingly. This trend is commonly observed in the real estate market, where larger houses, boasting more living area, often command higher prices.

A diagram of a living area

Description automatically generated

The scatterplot price against age of home illustrates the relationship between a house’s price and its age. Upon analysis, several trends emerge:

There is no strong correlation between the age of a home and its price within this dataset. The data points are dispersed across all price ranges without a clear trend, suggesting that the age of a home does not strongly influence its price.

Also, despite the lack of a strong overall trend, there is significant variability in prices across homes of all ages. Suggesting that factors not represented in this scatterplot could also be influencing the price. Therefore, while age is a factor, it is one of many variables that determine a house’s price.

A graph of blue dots

Description automatically generated

The correlation matrix offers a comprehensive view of the relationships between various housing characteristics. Specifically, the correlation coefficient between the price and the living area (sqft\_living) is 0.6895, a value closer to 1. This indicates a fairly strong positive correlation, suggesting that as the living area of a home increases, its price also tends to increase correspondingly. This positive correlation signifies that while one variable increases, the other variable also tends to increase.

On the other hand, the correlation coefficient between the price and the age of the home is -0.0746, a value close to 0. This suggests a very slight negative correlation, meaning the age of a home has a minimal downward impact on its price. A negative correlation means that while one variable increases, the other variable tends to decrease. However, due to the very weak correlation, the age of a home has an insignificant influence on its price.

### 

### Reporting Results

The general form of a multiple regression model is:

E(

where:

E(Y) is the response variable.

is the y-intercept.

ᵢ​ (where ᵢ = 1,2,…) are the slope parameters for each predictor variable xᵢ.

The prediction equation of the multiple regression model with price as the response variable and living area, upper level area, age of the home, number of bathrooms, and view as predictor variables would be:

Price=​+ (Living Area) +​(Upper Level Area) +​(Age of Home) + ​(Number of Bathrooms)+​(View)

where:

Price is the response variable (the variable I want to predict).

​ is the y-intercept (the predicted value of Price when all predictor variables are 0).

​,​,​,​,​ are the slope parameters for the predictor variables Living Area, Upper Level Area, Age of Home, Number of Bathrooms, and View respectively (the change in Price for a one-unit change in the predictor variable, holding all other variables constant).

Based on the outputs from your R script, the multiple regression model for predicting the price of a home can be written as follows:

price=7709 + 129.3 \* sqft\_living + 19.51 \* sqft\_above + 1451 \* age + 43970 \* bathrooms + 167500 \* view1 + 249000 \* view2

The breakdown of the coefficients:

Intercept: 7709

Living Area (sqft\_living): 129.3

Upper Level Area (sqft\_above): 19.51

Age of the Home (age): 1451

Number of Bathrooms (bathrooms): 43970

View (view1): 167500

View (view2): 249000

The statistical analysis of the model reveals the following key metrics:

The R-squared value for the model is 0.6029. This metric suggests that approximately 60.29% of the variability in home prices can be attributed to the predictor variables incorporated in the model, which include living area, upper level area, age of the home, number of bathrooms, and view. This implies that the model is capable of explaining a significant proportion of the variation in home prices

The Adjusted R-squared value, on the other hand, stands at 0.602. This value, which is marginally lower than the R-squared value, adjusts for the number of predictors in the model. It offers a more precise measure of the model’s goodness-of-fit, particularly when comparing models with varying numbers of predictors. An Adjusted R-squared value of 0.602 signifies that, after adjusting for the number of predictors, the model continues to explain about 60.2% of the variability in home prices.

In summary, these statistics indicate that the model exhibits a good fit.

The beta estimates for the living area and lake view variables in the model provide valuable insights into their impact on home prices.

The beta estimate for the Living Area (sqft\_living) is 129.3. This suggests that for each additional square foot of living area, the price of the home is projected to increase by $129.30, provided all other variables in the model remain constant. The positive coefficient underscores the correlation between larger living areas and higher home prices.

The beta estimate for the Lake View (view1) is 167,500. This indicates that homes with a lake view are anticipated to command a price that is $167,500 higher than homes without a lake view, assuming all other variables in the model are held constant. This substantial positive coefficient implies that the presence of a lake view significantly enhances the value of a home.

In summary, these beta estimates highlight the importance of living area size and lake view in determining home prices. They provide a quantitative measure of the expected increase in home price for each unit increase in these variables, while holding all other factors constant. This information can be particularly useful for homeowners, buyers, and real estate agents in understanding and predicting home prices.

The interpretation of the Residuals against Fitted Values Plot is as follows:

Homoscedasticity: Under ideal circumstances, the residuals should be randomly dispersed around the horizontal axis (zero line) without any noticeable pattern. However, in this plot, there appears to be an increase in the spread of residuals as the fitted values increase. This suggests potential heteroscedasticity, implying that the variability of the residuals is not constant across all levels of fitted values.

Linearity: The residuals should be centered around zero without any systematic patterns. Although the residuals are centered around zero, the increasing spread indicates that the assumption of linearity might be violated for higher fitted values.

In summary, the plot suggests that there might be issues with heteroscedasticity and potentially non-linearity in the model.

A diagram of a number of blue dots

Description automatically generated

The interpretation of the Normal Q-Q Plot is as follows:

Linearity: Ideally, the points in the Q-Q plot should align along the reference line (depicted in red) if the residuals are normally distributed. In the given plot, the points conform to the reference line, but there are noticeable deviations at both ends (tails) of the plot. This suggests that while the residuals are normally distributed, there may be some deviations in the tails.

Symmetry: The plot suggests that the residuals are symmetrically distributed around the mean. However, the deviations in the tails imply the presence of some outliers or extreme values that do not conform to the normal distribution.

In conclusion, the Normal Q-Q plot indicates that the residuals are normally distributed, although with some deviations in the tails.

A graph with a line

Description automatically generated

### Evaluating Significance of Model

The model’s significance can be evaluated using an overall F-test. The F-test is used to test the null hypothesis that all of the regression coefficients are equal to zero, against the alternative hypothesis that at least one of them is not equal to zero.

≠ 0 for ὶ = 1,2,….,6

The null hypothesis () is: All regression coefficients are equivalent to zero. This implies that none of the predictors (sqft\_living, sqft\_above, age, bathrooms, view1, view2) have any effect on the price.

The alternative hypothesis () is: At least one regression coefficient is not equivalent to zero. This implies that at least one of the predictors has an effect on the price.

The p-value for the F-statistic is less than 2.2e-16, which is significantly smaller than the 5% level of significance. Therefore, we reject the null hypothesis.

In conclusion, the model is significant at a 5% level of significance. This means that at least one of the predictors (sqft\_living, sqft\_above, age, bathrooms, view1, view2) has a significant effect on the price of the house.

The significance of individual terms can be evaluated using individual beta tests. A p-value less than 0.05 would lead to the rejection of the null hypothesis at a 5% level of significance.

0 for some ὶ = 1,2,….n

: ≠ 0

* sqft\_living:
* Null Hypothesis (): The coefficient of sqft\_living is equal to zero. = 0
* Alternative Hypothesis (): The coefficient of sqft\_living is not equal to zero.
* P-value: < 2e-16
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The sqft\_living term is significant at a 5% level of significance.
* sqft\_above:
* Null Hypothesis (): The coefficient of sqft\_above is equal to zero. = 0
* Alternative Hypothesis (): The coefficient of sqft\_above is not equal to zero.
* P-value: 0.00894
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The sqft\_above term is significant at a 5% level of significance.
* age:
* Null Hypothesis (): The coefficient of age is equal to zero. = 0
* Alternative Hypothesis (): The coefficient of age is not equal to zero.
* P-value: < 2e-16
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The age term is significant at a 5% level of significance.
* bathrooms:
* Null Hypothesis (): The coefficient of bathrooms is equal to zero. = 0
* Alternative Hypothesis (): The coefficient of bathrooms is not equal to zero.
* P-value: 9.13e-13
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The bathrooms term is significant at a 5% level of significance.
* view1:
* Null Hypothesis (): The coefficient of view1 is equal to zero. = 0
* Alternative Hypothesis (): The coefficient of view1 is not equal to zero.
* P-value: < 2e-16
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The view1 term is significant at a 5% level of significance.
* view2:
* Null Hypothesis (): The coefficient of view2 is equal to zero. = 0
* Alternative Hypothesis (): The coefficient of view2 is not equal to zero.
* P-value: < 2e-16
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The view2 term is significant at a 5% level of significance.

In conclusion, all the terms in the model (sqft\_living, sqft\_above, age, bathrooms, view1, view2) are significant at a 5% level of significance. This means that all these predictors have a significant effect on the price of the house.

### Making Predictions Using Model

**Scenario 1**

The predicted price for a home with 2150 sqft living area, 1050 sqft upper level living area, 15 years old, 3 bathrooms, and backs out to road is as follows:

The equation is:

price=7709+129.3×sqft\_living+19.51×sqft\_above+1451×age+43970×bathrooms+167500×view1+249000×view2

Substituting the given values into the equation, I get:

price=7709+129.3×2150+19.51×1050+1451×15+43970×3+167500×0+249000×0

Solving this equation will give me the predicted price for the home. Note that I have used 0 for view1 and view2 as the home backs out to a road and not to a lake.

Price = 7709 +2747945+20485+21765+131910+0+0

price=$459,814

So, the predicted price for a home with 2150 sqft living area, 1050 sqft upper level living area, 15 years old, 3 bathrooms, and backs out to road is approximately $459,814.

Prediction Interval: The prediction interval provides a range for the predicted price of a specific home with the given characteristics. The 90% prediction interval for the price of this home is ($239,563, $680,093.4). This means that we are 90% confident that the price of a home with these specific characteristics will fall within this range.

Confidence Interval: The confidence interval provides a range for the average price of all homes with the given characteristics. The 90% confidence interval for the price of this home is from ($446,087.9, $473,568.5). This means that we are 90% confident that the average price of all homes with these specific characteristics will fall within this range.

In summary, while the confidence interval gives me an estimate of where I expect the average home price to be, the prediction interval gives me an estimate of where I expect the price of a specific home to be. The prediction interval is wider than the confidence interval, reflecting the additional uncertainty when predicting a single observation as opposed to the mean.

**Scenario 2**

The predicted price for a home with 4250 sqft living area, 2100 sqft upper level living area, 5 years old, 5 bathrooms, and backs out to a lake is as follows:

The equation is:

price=7709+129.3×sqft\_living+19.51×sqft\_above+1451×age+43970×bathrooms+167500×view1+249000×view2

Substituting the given values into the equation, I get:

price=7709+129.3×4250+19.51×2100+1451×5+43970×5+167500×1+249000×0

Solving this equation:

price=7709+549525+40971+7255+219850+167500+0

price=$1,074,810

So, the predicted price for a home with 4250 sqft living area, 2100 sqft upper level living area, 5 years old, 5 bathrooms, and backs out to a lake is approximately $1,074,810.

Prediction Interval: The prediction interval provides a range for the predicted price of a specific home with the given characteristics. The 90% prediction interval for the price of this home is from $852,522.6 to $1,296,048. This means that we are 90% confident that the price of a home with these specific characteristics will fall within this range.

Confidence Interval: The confidence interval provides a range for the average price of all homes with the given characteristics. The 90% confidence interval for the price of this home is from $1,045,117 to $1,103,454. This means that we are 90% confident that the average price of all homes with these specific characteristics will fall within this range.

In summary, while the confidence interval gives me an estimate of where I expect the average home price to be, the prediction interval gives me an estimate of where I expect the price of a specific home to be. The prediction interval is wider than the confidence interval, reflecting the additional uncertainty when predicting a single observation as opposed to the mean.

When I make predictions, I use two types of intervals: prediction intervals and confidence intervals. Both give me a range of values, but they are used in slightly different ways.

A confidence interval is like a net that captures where I expect the average home price to fall based on certain home features. It is pretty reliable, but it assumes my model is perfect and does not account for all errors.

A prediction interval, on the other hand, is like a wider net that captures where I expect the price of a specific home to fall. It is wider because predicting the price of a single home is trickier and has more uncertainty than predicting an average price.

In short, if you are looking at the average price of homes with certain features, the confidence interval is my go-to. But if you are trying to predict the price of a specific home, the prediction interval will give you a more realistic range. That is why the prediction interval is wider—it accounts for more uncertainty.

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

### Correlation Analysis

The scatterplot below shows a relationship between the price of properties and the average school rating in the area. The plot is densely populated with points that form a curve, suggesting a positive correlation between the two variables: as the average school rating increases, so does the price. Given the curvature in the scatterplot, showing the prices rise more steeply at higher school ratings, indicating a possible quadratic relationship. Therefore, a second order (quadratic) model would provide a better fit for this data accounting for the rate of change increasing as school ratings improve.

A graph of a scatter plot

Description automatically generated

The scatterplot below illustrates an inverse relationship between property prices and crime rates per 100,000 residents. The data points form a curve that begins high for areas with low crime rates, then sharply declines as crime rates rise, eventually plateauing. This pattern suggests a non-linear association, hinting that a quadratic model may be more effective in capturing this trend.

*A graph with red dots

Description automatically generated*

### Reporting Results

The general form of a complete second order model for price using average school rating in the area and crime rate per 100,000 people as predictors is:

E(Y)= ​++ ​ + ​+ ​ +

Where:

E(Y) is the expected value of the dependent variable we want to predict, which is the price.

 and​ are the predictor variables, which are the average school rating in the area and the crime rate per 100,000 people, respectively.

(where ὶ = 0, 1, 2,…., 5) are the slope parameters for the predictor variables and the intercept.

The prediction equation would be similar:

Price^=+​(school\_rating)+​(crime)+​( school\_rating)²+​(crime)²+​( school\_rating \* crime)

Here, ^ denotes estimated or predicted value of Price based on the model.. So, β̂ᵢ are the estimated slope parameters obtained from the data. This equation would be used to predict the price given values for average school rating in the area and crime rate per 100,000.

Based on the output from your R script, the complete second order model for price using average school rating in the area and crime rate per 100,000 people as predictors can be written as follows:

Price^= 733900 − 73750 \* school\_rating − 3155 \* crime + 11650 \* school\_rating² + 6.377 \* crime² − 52.27 \* school\_rating \* crime\

Where:

Price is the dependent variable I want to predict.

school\_rating and crime are the predictor variables.

The coefficients (733900, -73750, -3155, 11650, 6.377, -52.27) are the estimated parameters obtained from the data.

The model’s R-squared (R²) value is 0.8088, and the adjusted R-squared () value is 0.8084. R², also known as the coefficient of determination, indicates that about 81% of the variability in property prices can be accounted for by our model that uses average school rating and crime rate as predictors.

In this case, the adjusted R-squared is slightly lower than the R-squared, but it still suggests that our model explains approximately 81% of the price variability.

The interpretation of the Residuals against Fitted Values Plot is as follows:

The residuals are indeed scattered around zero with no discernible pattern or trend. This confirms that the assumptions of linearity and homoscedasticity, which is the constant variance of the errors, are met. In a valid model, it is expected that the residuals will be randomly dispersed around zero across the range of fitted values. This is exactly what we observe here. Therefore, we can confidently state that our model meets these key assumptions.

A diagram of red dots

Description automatically generated

The interpretation of the Normal Q-Q Plot is as follows:

In the below Q-Q plot, the majority of points align closely with the reference line, particularly in the middle quantiles. This alignment confirms that the residuals are normally distributed. However, there are noticeable deviations from the reference line at both ends of the plot, especially in the upper right corner. These deviations could signify the presence of potential outliers yet while this may be the case the Q-Q plot definitively indicates that the assumption of normality for the residuals is met for this model.

*A graph with a red line

Description automatically generated*

### Evaluating Significance of Model

The model’s significance can be evaluated using an overall F-test. The null and alternative hypotheses are as follows:

: ≠ 0 for ὶ = 1,2,….,5

Null Hypothesis ((The model is not significant)

Alternative Hypothesis (: ≠ 0 for ὶ = 1,2,….,5 (The model is significant)

The p-value for the F-statistic is given as < 2.2e-16, which is practically zero. This is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that at least one of the predictors’ coefficients is not zero. This means that the model is statistically significant at a 5% level of significance. The predictors in the model (average school rating, crime rate, their interaction, and their squared terms) collectively have a significant effect on the price.

The individual beta tests are used to test the null hypothesis that each regression coefficient is equal to zero versus the alternative hypothesis that it is not zero.

0 for some ὶ = 1,2,….n

: ≠ 0

* Intercept :
* Null Hypothesis (): = 0 (The intercept is not significant)
* Alternative Hypothesis (): B0 ≠ 0 (The intercept is significant)
* P-value: 1.45e-12 (which is practically zero)
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The intercept is statistically significant at a 5% level of significance.
* school\_rating:
* Null Hypothesis (): = 0 (The school rating is not significant)
* Alternative Hypothesis (): ≠ 0 (The school rating is significant)
* P-value: 0.000406
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The school rating is statistically significant at a 5% level of significance.
* crime:
* Null Hypothesis (): = 0 (The crime rate is not significant)
* Alternative Hypothesis (): ≠ 0 (The crime rate is significant)
* P-value: 1.90e-09 (which is practically zero)
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The crime rate is statistically significant at a 5% level of significance.
* school\_rating squared:
* Null Hypothesis (): = 0 (The square of the school rating is not significant)
* Alternative Hypothesis (): ≠ 0 (The square of the school rating is significant)
* P-value: < 2e-16 (which is practically zero)
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The square of the school rating is statistically significant at a 5% level of significance.
* crime Squared:
* Null Hypothesis ): = 0 (The square of the crime rate is not significant)
* Alternative Hypothesis (): ≠ 0 (The square of the crime rate is significant)
* P-value: < 2e-16 (which is practically zero)
* Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis. The square of the crime rate is statistically significant at a 5% level of significance.
* Interaction Term (school\_rating: crime):
* Null Hypothesis (): = 0 (The interaction term is not significant)
* Alternative Hypothesis (): ≠ 0 (The interaction term is significant)
* P-value: 0.281513
* Conclusion: Since the p-value is greater than 0.05, we fail to reject the null hypothesis. The interaction term is not statistically significant at a 5% level of significance.

In summary, the regression model was evaluated for statistical significance using individual beta tests for each coefficient. The intercept, school rating, crime rate, square of school rating, and square of crime rate were found to be statistically significant at a 5% level of significance, as their p-values were less than 0.05. This means these predictors have a significant effect on the price. However, the interaction term between school rating and crime rate was not statistically significant at a 5% level of significance, as its p-value was greater than 0.05.

### Making Predictions Using Model

**Scenario 1**

The predicted price for a home in an area with average school rating of 9.80 and a crime rate of 81.02 per 100,000 individuals is as follows:

Price^= 733900 − 73750 \* school\_rating − 3155 \* crime + 11650 \* school\_rating² + 6.377 \* crime² − 52.27 \* school\_rating \* crime

Price^= 733900 – 73750 \* 9.80 -3155 \* 81.02 +11650 \* (9.80)² + 6.377 \* (81.02)² - 52.27 \* 9.80 \*81.02

The predicted price for a home in an area with an average school rating of 9.80 and a crime rate of 81.02 per 100,000 individuals is approximately $874,497.

The 90% prediction interval for this price is ($721,606.2, $1,027,388). This interval provides a range of values that would contain the future observed price of a home with a similar school rating and crime rate, with a confidence level of 90%.

The 90% confidence interval for this price is ($863,681.4, $885,312.7). This interval provides a range of values that would contain the true average price of homes with a similar school rating and crime rate, with a confidence level of 90%.

In other words, I am 90% confident that the true average price of homes with a school rating of 9.80 and a crime rate of 81.02 per 100,000 individuals lies between $863,681.4 and $885,312.7. Similarly, if I were to predict the price of a new home with the same school rating and crime rate, I am 90% confident that its price would fall between $721,606.2 and $1,027,388.

**Scenario 2**

The predicted price for a home in an area with average school rating of 4.28 and a crime rate of 215.50 per 100,000 individuals is as follows:

Price^= 733900 − 73750 \* school\_rating − 3155 \* crime + 11650 \* school\_rating² + 6.377 \* crime² − 52.27 \* school\_rating \* crime

Price^= 733900 – 73750 \* 4.28 -3155 \* 215.50 +11650 \* (4.28)² + 6.377 \* (215.50)² - 52.27 \*

4.28 \* 215.50

The predicted price for a home in an area with an average school rating of 4.28 and a crime rate of 215.50 per 100,000 individuals is approximately $199,706.7.

The 90% prediction interval for this price is ($46,991.65, $352,421.7). This interval provides a range of values that would contain the future observed price of a home with a similar school rating and crime rate, with a confidence level of 90%.

The 90% confidence interval for this price is ($191,753.5, $207,659.9). This interval provides a range of values that would contain the true average price of homes with a similar school rating and crime rate, with a confidence level of 90%.

In other words, we are 90% confident that the true average price of homes with a school rating of 4.28 and a crime rate of 215.50 per 100,000 individuals lies between $191,753.5 and $207,659.9. Similarly, if we were to predict the price of a new home with the same school rating and crime rate, we are 90% confident that its price would fall between $46,991.65 and $352,421.7.

## 5. Nested Models F-Test

### Reporting Results

The general form of a first order model for price using average school rating in the area and crime rate per 100,000 people as predictors is:

E(Y)= ​++ ​ +

E(Y)= ​++ ​ +

The prediction equation would be :

Price^=+​(school\_rating)+ ​(crime)+ ​( school\_rating \* crime)

Price^= − 410233.37 + 155559.97 \* school\_rating + 2230.07 \* crime − 564.85(school\_rating \* crime)

In this equation:

() (slope parameter) = -410233.37

() (school\_rating) = 155559.97

(​) (crime) = 2230.07

() (interaction term) = -564.85

This model includes the main effects of school rating and crime rate, as well as their interaction effect, to predict house prices.



### Evaluating Significance of Model

To determine if the model is significant at a 5% level of significance, we need to carry out the overall F-test.

: ≠ 0

Null Hypothesis

The null hypothesis states that all the regression coefficients are equal to zero, meaning that the predictors do not explain the variability in the response variable (house prices).

Alternative Hypothesis ():

The alternative hypothesis states that at least one of the regression coefficients is not equal to zero, meaning that the predictors do explain some of the variability in the response variable.

P-value:

The p-value for the overall F-test is given as (< 2.2 \times 10^{-16}).

Conclusion:

Since the p-value is much smaller than the significance level of 0.05, we reject the null hypothesis. This means that the model is significant at the 5% level of significance, indicating that the predictors (school rating, crime rate, and their interaction) do explain a significant portion of the variability in house prices.

In summary, the overall F-test shows that the model is significant, and the predictors are useful in explaining the variability in house prices.

To determine which terms are significant at a 5% level of significance, we need to carry out individual beta tests for each predictor.

* Intercept
* Null Hypothesis (): = 0
* Alternative Hypothesis (): B0 ≠ 0
* P-value: (< 2 \times 10^{-16})
* Conclusion: Since the p-value is much smaller than 0.05, we reject the null hypothesis. The intercept is significant.
* school\_rating ( )):
* Null Hypothesis (): = 0
* Alternative Hypothesis (): ≠ 0
* P-value: (< 2 \times 10^{-16})
* Conclusion: Since the p-value is much smaller than 0.05, we reject the null hypothesis. The school rating is significant.
* crime ):
* Null Hypothesis (): = 0
* Alternative Hypothesis (): ≠ 0
* P-value: (< 2 \times 10^{-16})
* Conclusion: Since the p-value is much smaller than 0.05, we reject the null hypothesis. The crime rate is significant.
* Interaction Term ():
* Null Hypothesis (): = 0
* Alternative Hypothesis (): ≠ 0
* P-value: (< 2 \times 10^{-16})
* Conclusion: Since the p-value is much smaller than 0.05, we reject the null hypothesis. The interaction term is significant.

In summary all terms (intercept, school rating, crime rate, and their interaction) are significant at a 5% level of significance.

### Model Comparison

When comparing two models, the reduced model is a simpler version that includes fewer predictors, while the complete model is a more complex version that includes additional predictors or interaction terms. In this case, the reduced model is a linear regression model with school rating and crime rate as predictors, and their interaction term. The general form and prediction equation of the reduced model are:

General form:

E(Y)= ​++ ​ +

E(Y)= ​++ ​ +

Prediction equation:

= +​(school\_rating)+​(crime)+​( school\_rating \* crime)

The complete model is a second order model with school rating and crime rate as predictors, their interaction term, and their squared terms. The general form and prediction equation of the complete model are:

General form:

E(Y)= ​++ ​ + ​+ ​ +

Prediction equation:

=+​(school\_rating)+​(crime)+​( school\_rating)²+​(crime)²+​( school\_rating \* crime)

The nested model F-test is used to determine if the quadratic (squared) terms are needed in the model. The null hypothesis is that the additional terms in the complete model do not significantly improve the fit of the model, while the alternative hypothesis is that they do.

In this case, the null and alternative hypotheses are:

: ≠ 0 for ὶ = k,….,n

Null Hypothesis (): The additional predictors in the complete model do not significantly improve the fit of the model. Where ὶ is the number of predictors in the reduced model, and n is the number of additional predictors in the complete model.

Alternative Hypothesis (): At least one of the additional predictors in the complete model significantly improves the fit of the model

The p-value for the F-statistic is given as 2.22716e-28, which is practically zero and less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the quadratic terms significantly improve the fit of the model at a 5% level of significance. This means that the complete model (Model #2) is a better fit for the data compared to the reduced model.

## 6. Conclusion

To predict house prices, the complete second order regression model is recommended. This model incorporates average school rating in the area and crime rate per 100,000 people as predictors, along with the interaction between school rating and crime, school rating squared, and crime squared. The statistical analyses shown includes a first order regression model, a complete second order regression model, a reduced model test, and a nested F-Test (ANOVA) to evaluate whether the reduced model is sufficient or if the complete second order model should be used. The complete second order model was selected due to its highest R-squared value, indicating that it explains the greatest percentage of variance. This model validates a strong correlation concerning these predictors and the response variable of home price.

The practical significance of these analyses lies in their ability to provide insights into the factors influencing house prices, aiding in informed decision-making for real estate investment and policy planning. For instance, they can assist potential homebuyers or real estate investors in understanding the potential cost of a house based on the school rating and crime rate in the area. Additionally, they can guide lawmakers on the impact of improving school ratings or reducing crime rates on housing prices within a community.

## 7. Citations

zyBooks. (n.d.). Learn.zybooks.com. Retrieved July 26, 2024, from <https://learn.zybooks.com/zybook/MAT-303-16699.202456-1/chapter/2/section/1>