# MAT 303 Project One Summary Report

Brandon B. Gibbs

Brandon.gibbs@snhu.edu

Southern New Hampshire University

Note: Replace the bracketed text on page one (the cover page) with your personal information.

## 1. Introduction

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## 2. Data Preparation

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

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

### Correlation Analysis

* *Create the following scatterplots and include a copy of each in this section:*
  + *Price (price) vs. the living area (sqft\_living)*
  + *Price (price) vs. the age of the home (age)*
* *Describe what trends, if any, exist for each scatterplot.*
* *Report the correlation coefficients between the following variables:*
  + *Price (price) vs. the living area (sqft\_living)*
  + *Price (price) vs. the age of the home (age)*
* *Describe the strength and direction of each correlation coefficient.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### Reporting Results

The general form of the multiple regression model using price as the response variable and living area, upper-level area, age of the home, number of bathrooms, and view as predictor variables is:

Using the outputs obtained from your R script, the prediction model equation is:

**R-Squared and Adjusted R-Squared**

**R-Squared (R²)**: 0.6029 meaning that approximately 60.29% of the variance in the price is explained by the model and indicates a moderate fit of the model to the data.

**Adjusted R-Squared value** of 0.602 accounts for the number of predictors in the model and provides a more accurate measure of the goodness -of-fit. Here, 60.2% of the variance in the price is explained by the model, after adjusting for the number of predictors.

**Interpretation of Beta Estimates**

**Living Area** : 129.30sf

For each additional square foot of living area, the price increases by $129.30, assuming all other factors remain constant. This positive coefficient indicates that larger living areas are associated with higher prices.

**View**

**view1**: $167, 500

Homes with a view classified as '1' (e.g., lake view) have a price increase of $167,500 compared to the reference category, assuming all other factors remain constant.

**view2**: $249, 000

Homes with a view classified as '2' (lake view) have a price increase of $249,000 compared to the reference category, assuming all other factors remain constant.

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### Evaluating Significance of Model

* *Is the model significant at a 5% level of significance? Carry out the overall F-test by identifying the null hypothesis, the alternative hypothesis, the P-value, and the conclusion of the test.*
* *Which terms are significant at a 5% level of significance? Carry out individual beta tests by identifying the null hypothesis, the alternative hypothesis, the P-value, and the conclusion of each test.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### Making Predictions Using Model

*Predicted Prices and Intervals*

*Home 1: 2150 sqft living area, 1050 sqft upper level living area, 15 years old, 3 bathrooms, backs out to road*

*Predicted Price: $459, 828*

*90% Prediction Interval: [$239,563, $680,093]*

*90% Confidence Interval: [$413,088, $473,569]*

*Home 2: 4250sqft total living area, 2100sqft upper level living area, 5 years old, 5 bathrooms and a lake view*

*Predicted Price: $1,074,285*

*90% Prediction Interval: [$852,523, $1,296,048]*

*90% Confidence Interval: [$1,045,111, $1,103,454]*

*Interpretation of Intervals*

*Prediction Interval: This interval provides a range within which we expect the price of a new home with the given characteristics to fall, with 90% confidence. It accounts for both the uncertainty in the model parameters and the variability in the individual home prices.*

*Confidence Interval: This interval provides a range within which we expect the average price of homes with the given characteristics to fall, with 90% confidence. It only accounts for the uncertainty in the model parameters.*

*Why is the Prediction Interval Wider than the Confidence Interval?*

*The prediction interval is wider than the confidence interval because it includes the variability in individual home prices in addition to the uncertainty in the model parameters. The confidence interval only accounts for the uncertainty in the estimated average price, while the prediction interval also considers the natural variability in the prices of individual homes.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

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

### Correlation Analysis

* *Create scatterplots of:*
  + *Price (price) vs. average school rating in the area (school\_rating)*
  + *Price (price) vs. the crime rate per 100,000 people (crime)*

***Comments on Scatterplots***

* ***Price vs. Average School Rating****: The scatterplot of price against average school rating typically shows a positive trend, indicating that as the school rating increases, the price of the home also tends to increase. This suggests a positive relationship between school rating and price. If the points form a curved pattern, it may indicate that a second-order model could be appropriate.*
* ***Price vs. Crime Rate****: The scatterplot of price against crime rate typically shows a negative trend, indicating that as the crime rate increases, the price of the home tends to decrease. This suggests a negative relationship between crime rate and price. If the points form a curved pattern, it may indicate that a second-order model could be appropriate.*

***Determining if a Second-Order Model is Appropriate***

*To determine if a second-order model is appropriate, we can look for curvature in the scatterplots. If the relationship between the variables and the response variable (price) is not linear and shows curvature, a second-order model may be more appropriate to capture the non-linear relationship.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### Reporting Results

* *Write the general form and the prediction equation of a complete second order model for price using average school rating in the area and crime rate per 100,000 people as predictors. Use (where i* equals *1, 2, ... ) to represent the slope parameters for all predictor variables.*
* *Create the complete second order model for price using average school rating in the area and crime rate per 100,000 people as predictors. Write the prediction model equation using outputs obtained from your R script.  
  Note: Use average school rating in the area and crime rate as* ***quantitative*** *variables in this model. Use the equation editor to write the prediction model equation.*
* *What are the values of (R-squared) and (adjusted R-squared) for the model? Provide your interpretation of these statistics.*
* *Obtain the residuals and fitted values to create the following plots. Include these plots and comment on the validity of assumptions. Do not include the residual or fitted values tables.*
  + *Residuals against Fitted Values*
  + *Normal Q-Q plot*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### Evaluating Significance of Model

***Overall F-Test***

***Null Hypothesis (H₀)***

*The null hypothesis states that all the regression coefficients are equal to zero, meaning that none of the predictor variables have a significant effect on the response variable (price).*

***Alternative Hypothesis (H₁)***

*The alternative hypothesis states that at least one of the regression coefficients is not equal to zero, meaning that at least one predictor variable has a significant effect on the response variable (price).*

***P-Value***

*From the model summary, the p-value for the overall F-test is 2.2e-16.*

***Conclusion***

*Since the p-value is less than 0.05, we reject the null hypothesis. This means that the model is significant at a 5% level of significance, indicating that at least one of the predictor variables has a significant effect on the response variable (price).*

***Individual Beta Tests***

***Null Hypothesis (H₀)***

*The null hypothesis for each individual beta test states that the corresponding regression coefficient is equal to zero, meaning that the predictor variable does not have a significant effect on the response variable (price).*

***Alternative Hypothesis (H₁)***

*The alternative hypothesis for each individual beta test states that the corresponding regression coefficient is not equal to zero, meaning that the predictor variable has a significant effect on the response variable (price).*

*Based on the model summary, the p-values and conclusions for each term are as follows:*

*The intercept has a p-value of 1.45e-12, indicating that it is significant at a 5% level of significance. Therefore, we reject the null hypothesis (H₀) for the intercept. The school\_rating variable has a p-value of 0.000406, which is also significant at a 5% level of significance, leading us to reject the null hypothesis (H₀) for school\_rating. Similarly, the crime variable has a p-value of 1.90e-09, making it significant at a 5% level of significance, and we reject the null hypothesis (H₀) for crime.*

*The squared term for school\_rating, I(school\_rating^2), has a p-value of less than 2e-16, indicating it is highly significant at a 5% level of significance. Thus, we reject the null hypothesis (H₀) for I(school\_rating^2). The squared term for crime, I(crime^2), also has a p-value of less than 2e-16, making it highly significant at a 5% level of significance, and we reject the null hypothesis (H₀) for I(crime^2).*

*However, the interaction term between school\_rating and crime, school\_rating:crime, has a p-value of 0.281513, which is not significant at a 5% level of significance. Therefore, we fail to reject the null hypothesis (H₀) for the interaction term school\_rating:crime*

*Summary*

*The overall model is significant at a 5% level of significance. The individual beta tests show that all terms except for the interaction term (school\_rating:crime) are significant at a 5% level of significance.*

* *Is the model significant at a 5% level of significance? Carry out the overall F-test by identifying the null hypothesis, the alternative hypothesis, the P-value, and the conclusion of the test.*
* *Which terms are significant at a 5% level of significance? Carry out individual beta tests by identifying the null hypothesis, the alternative hypothesis, the P-value, and the conclusion of each test.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### Making Predictions Using Model

* *What is 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? Obtain 90% prediction and confidence intervals for the price of this home. Interpret each interval.*
* *What is 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? Obtain 90% prediction and confidence intervals for the price of this home. Interpret each interval.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## 5. Nested Models F-Test

### Reporting Results

*The general form of the first-order regression model is:*

*Using the outputs obtained from your R script, the prediction model equation can be written as:*

***Evaluating Significance of the Model***

*The overall F-test is used to determine if the model is significant at a 5% level of significance. The null hypothesis (H₀) states that all the regression coefficients are equal to zero, meaning that none of the predictor variables have a significant effect on the response variable (price). The alternative hypothesis (H₁) states that at least one of the regression coefficients is not equal to zero, meaning that at least one predictor variable has a significant effect on the response variable (price). The p-value for the overall F-test is less than 2.2e-16, which is much smaller than 0.05. Therefore, we reject the null hypothesis and conclude that the model is significant at a 5% level of significance.*

*For the individual beta tests, the null hypothesis (H₀) for each term states that the corresponding regression coefficient is equal to zero, meaning that the predictor variable does not have a significant effect on the response variable (price). The alternative hypothesis (H₁) states that the corresponding regression coefficient is not equal to zero, meaning that the predictor variable has a significant effect on the response variable (price). Based on the model summary, the p-values for the intercept, school\_rating, crime, and the interaction term (school\_rating:crime) are all less than 2e-16, which is much smaller than 0.05. Therefore, we reject the null hypothesis for each term and conclude that all terms are significant at a 5% level of significance.*

* *Write the general form and the prediction equation of a first order model for price using average school rating in the area and crime rate per 100,000 people as predictors. Include the interaction term between average school rating and crime rate. Use (where i* equals *1, 2, ... ) to represent the slope parameters for all predictor variables.*
* *Create the first order regression model for price using average school rating in the area and crime rate per 100,000 people as predictors. Include the interaction term between average school rating and crime rate. Write the prediction model equation using outputs obtained from your R script.  
  Note: Use average school rating and crime rate as* ***quantitative*** *variables in this model. Use the equation editor to write the prediction model equation with the outputs*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### Evaluating Significance of Model

*The overall F-test is used to determine if the model is significant at a 5% level of significance. The null hypothesis (H₀) states that all the regression coefficients are equal to zero, meaning that none of the predictor variables have a significant effect on the response variable (price). The alternative hypothesis (H₁) states that at least one of the regression coefficients is not equal to zero, meaning that at least one predictor variable has a significant effect on the response variable (price). The p-value for the overall F-test is less than 2.2e-16, which is much smaller than 0.05. Therefore, we reject the null hypothesis and conclude that the model is significant at a 5% level of significance.*

*For the individual beta tests, the null hypothesis (H₀) for each term states that the corresponding regression coefficient is equal to zero, meaning that the predictor variable does not have a significant effect on the response variable (price). The alternative hypothesis (H₁) states that the corresponding regression coefficient is not equal to zero, meaning that the predictor variable has a significant effect on the response variable (price). Based on the model summary, the p-values for the intercept, school\_rating, crime, and the interaction term (school\_rating:crime) are all less than 2e-16, which is much smaller than 0.05. Therefore, we reject the null hypothesis for each term and conclude that all terms are significant at a 5% level of significance.*

* *Is the model significant at a 5% level of significance? Carry out the overall F-test by identifying the null hypothesis, the alternative hypothesis, the P-value, and the conclusion of the test.*
* *Which terms are significant at a 5% level of significance? Carry out individual beta tests by identifying the null hypothesis, the alternative hypothesis, the P-value, and the conclusion of each test.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

### Model Comparison

*The ANOVA formula for comparing nested models involves calculating the F-statistic to determine if the additional terms in the more complex model significantly improve the fit of the model. Here's the general formula for Nested F-test*

*Where:*

* *RSSreducedRSS\_{reduced} is the residual sum of squares for the reduced model.*
* *RSScompleteRSS\_{complete} is the residual sum of squares for the complete model.*
* *dfreduceddf\_{reduced} is the degrees of freedom for the reduced model.*
* *dfcompletedf\_{complete} is the degrees of freedom for the complete model.*

*The F-statistic follows an F-distribution with (dfreduced−dfcomplete,dfcomplete)(df\_{reduced} - df\_{complete}, df\_{complete}) degrees of freedom. If the calculated F-statistic is greater than the critical value from the F-distribution table at a given significance level (e.g., 0.05), we reject the null hypothesis and conclude that the additional terms in the complete model significantly improve the fit.*

*You will now compare this model with the second order model for price using average school rating in the area and crime rate per 100,000 people as predictors to test whether the quadratic (squared) terms contribute in predicting the prices of homes. The complete second order model is Model #2, which you created in this project.*

* *In general, what is a reduced and a complete model when comparing two models?*
* *Write general form and prediction equation of the model that is the reduced model in this comparison in terms of and ?.*
* *Write general form* *and prediction equation of the model that is the complete model in this comparison in terms of and ?*
* *Run the nested model F-test at a 5% level of significance to evaluate if the quadratic (squared) terms are needed. Identify the null hypothesis, the alternative hypothesis, the P-value, and the conclusion of the test.*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## 6. Conclusion

*Describe the results of the statistical analyses clearly, using proper descriptions of statistical terms and concepts. Fully describe what these results mean for your scenario.*

* *Which model would you choose to predict house prices? Briefly summarize your findings in plain language.*
* *What is the practical importance of the analyses that were performed?*

* Answer the questions in a paragraph response. Remove all questions and this note before submitting! Do not include R code in your report.*

## 7. Citations

*You were* ***not*** *required to use external resources for this report. If you did not use any resources, you should remove this entire section. However, if you did use any resources to help you with your interpretation, you* ***must*** *cite them. Use proper APA format for citations.*

Insert references here in the following format:

Author's Last Name, First Initial. Middle Initial. (Year of Publication). Title of book: Subtitle of book, edition. Place of Publication: Publisher.