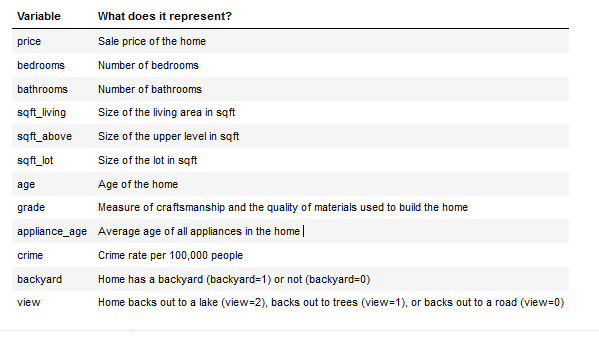
## 1. Introduction

We are analysts, and we have access to a unique set of historical data from different end times that analyzed housing data set. In the dataset, we are exploring many types of variables corresponding to the price of housing. The variable used can determine what type of regression model will be developed given to the specific scenario. We are trying to do a case study and develop a regression model.

Data analysis can better prepare for different scenarios, such as housing prices in a given area.  Data sets are essential and developing regression models. Especially nowadays, with everything collecting data around us, finding validity is more critical than ever. This regression model can help someone like a real estate agent get the upper edge over their competitors. Regression models are used, such as in case studies or anywhere is very important in general. Using interactive terms, quadratic predictors, and qualitative predictors will help us evaluate how efficient the regression model is given the response variables to determine the price of housing. Then the regression model can be developed to be exceptionally accurate. Then a regression model could be implemented the price of housing in a given area.

## 2. Data Preparation

The data sets that have been given are from housing.csv. The data set has many variables that are related to wage growth. The data set consists of 22 columns and around 2,692 rows. The columns are the particular variable, and the rows are the different values of historical data sets for wage growth, given the particular variables being compared. We were asked to evaluate the wage growth of 2,692 historical data sets entries, given their data sets, to develop a  regression model for predicting the price of housing.  They had so many variables that could be extremely important for a data analyst to develop some regression model based on the data. The most critical key is finding the linear and quadratic relationships between variables and establishing a regression model that makes sense of the data.



However, we must also consider the interaction terms and qualitative predictors, and quadratic variables used. Such variables can help find the linear and quadratic relationships with each other under certain conditions. Under these conditions, the regression model can determine a strong linear or quadratic relationship between the variables. We want to find a regression model that best fits an actual scenario data set.

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

### Correlation Analysis

*Chart, scatter chart

Description automatically generated*

When analyzing the scatter plot between the price of housing and the living area (SQFT\_Living), we can see a linear relationship. However, we can see the due to the quantity of datasets, which is hard to see.  However, we can see that there is a linear relationship going in the positive direction. A promising sign that there is a linear relationship between the price of Housing and SQFT\_living. We can see that a first-order linear regression model might fit these data set values pretty well; however, we must further analyze the data sets. However, by observing the scatterplot, I would conclude that it would be a first-order linear regression model.  We will do further analysis to determine the accuracy of which kind of regression model will fit the scatter plot the best to develop a regression model that determines the housing price based on SQFT\_living. Based on the value set, there appears to be a strong correlation between the variables. It would be a positive linear relationship between the variables, so a  strong positive correlation in terms of a linear regression model between SQFT\_living and Price growth rate merits an investigation.

*Chart, scatter chart

Description automatically generated*

When analyzing the scatter plot between the price of housing and the age of the house (age), we can see there is some kind of linear relationship. However, we can see the due to the quantity of datasets, which is hard to see.  However, we can see that there is a linear relationship going in the negative direction. There is a linear relationship between the price of housing and the age of the house. We can see that a first-order linear regression model might fit these data set values pretty well; however, we must further analyze the data sets. However, by observing the scatterplot, I would conclude that it would be a first-order linear regression model.

Further analysis to determine the accuracy of which kind of regression model will fit the scatter plot the best to develop a regression model that determines the housing price based on the house's age. Based on the value set, there appears to be a strong correlation between the variables. It would be a negative linear relationship between the variables, so a  strong negative correlation in terms of a linear regression model between the age of the house and Price growth rate merits a investigation.

### Reporting Results

|  |
| --- |
|  |
| First Order Regression Model with Quantitative and Qualitative Variables Then, The general form of this regression model is:  y = Price  *x*1 = SQFT\_living  *x*2 = Grade  x3 = Bathrooms  x4 = View |
| The Coefficients of Beta Estimates  y = Price  *x*1 = SQFT\_living  *x*2 = Grade  x3 = Bathrooms  x4 = View |

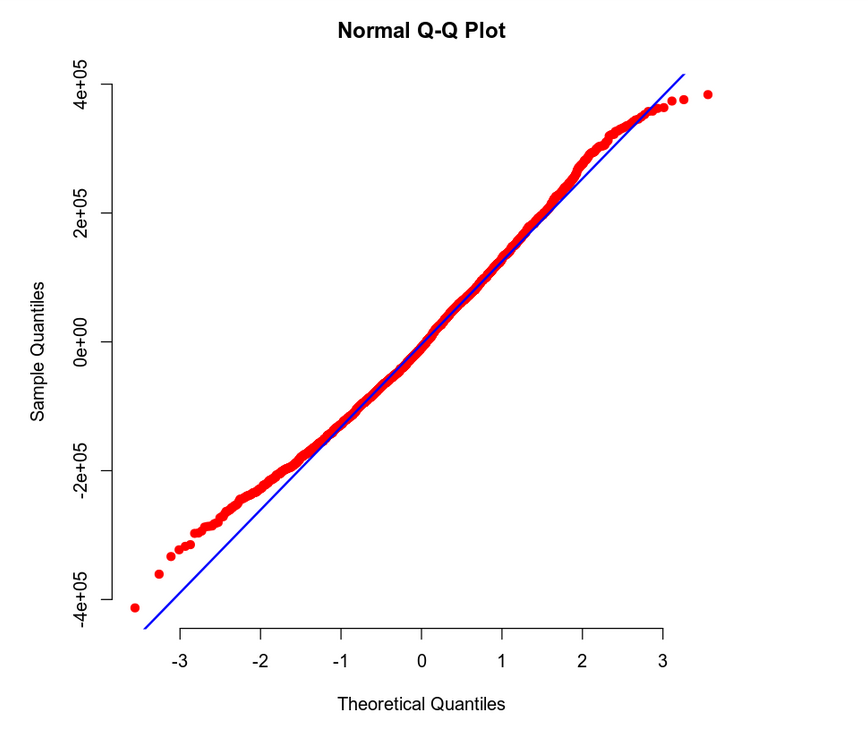
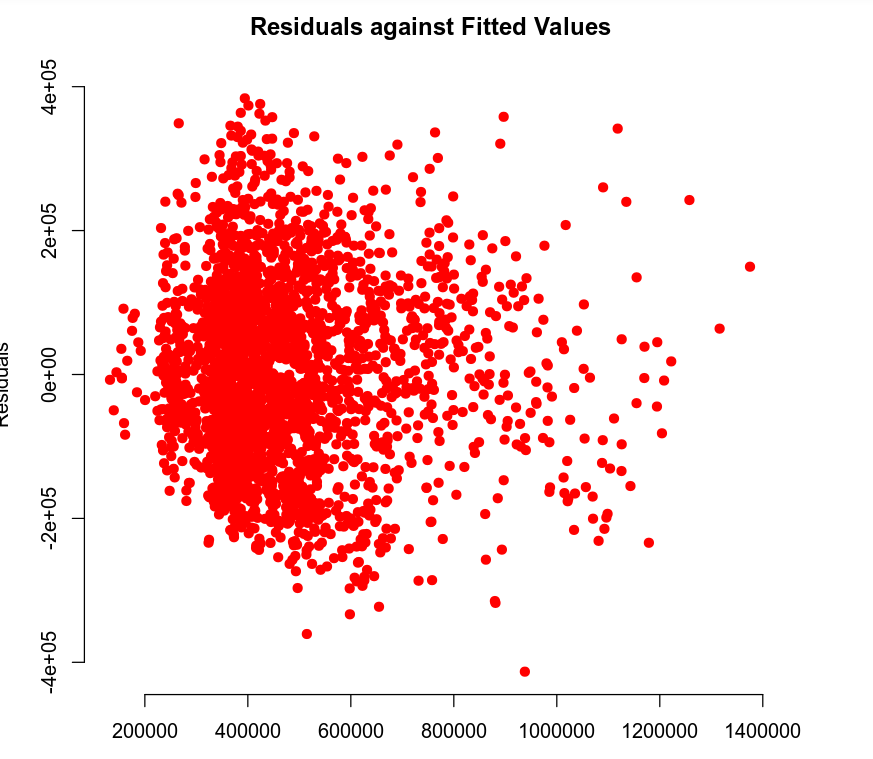
|  |
| --- |
| **The Coefficient of multiple determination and Adjusted Coefficient of multiple determination**  (R-squared) (Adjusted R-squared) |

The coefficient of multiple determination is the quotient of the fitted values variances and observed values of the dependent variable. The calculated coefficient of considerable determination for the regression model R2 is  0.645. This means that 64.5% of the housing price variation is explained by the regression model that uses SQFT\_living, Grade, Bathrooms, an interaction term between SQFT\_living:Grade SQFT\_living:Bathrooms, Bathrooms:Grade and qualitative term of view.

Based on the value of adjusted R2, the proportion of variation explained by the estimated regression line is 0.644, which means that 64.41 % of the housing price variation is explained by the regression model that uses SQFT\_living, Grade, Bathrooms, an interaction term between SQFT\_living:Grade SQFT\_living:Bathrooms, Bathrooms:Grade and qualitative trem View.

The estimated coefficient for SQFT\_living is 1.301E02, which means that on average, the housing price increases by1.301E02 for each unit increase in the price of the house. The estimated coefficient for the grade is 7.811E04, which means that on average, the price of housing increases by 7.811E04 for each unit increase in the price of the house. The estimated coefficient for bathrooms is 8.866E04. This means that, on average, the price of housing decreases by 8.866E04 for each unit increase in the price of the house. The estimated coefficient for the view is 1.234E05. This means that, on average, the price of housing increases by 1.234E05 for each unit increase in the price of the house. The estimated coefficient for SQFT\_living:grade is 6.239E00, which means that on average, the price of housing decreases by 6.239E00 for each unit increase in price of the house.

The estimated coefficient for SQFT\_living:bathroom is 5.886E00, which means that on average, the price of housing increases by 5.886E00 for each unit increase in the price of the house. The estimated coefficient for grade:bathrooms are 7.332E03, which means that on average, the price of housing increases by 7.332E03 for each unit increase in the price of the house.

* + 

We can see the linear relationship if the multiple regression model with normal probability plot with all the given housing data in housing.csv. It is essential to have the ability to evaluate the linear relationship in multiple regression models. For this reason, multiple regression models are only considered valid if certain assumptions can be made about the sample population. These assumptions can help determine if the regression model's data population sample is accurate or violates any assumptions. Violating these assumptions can conclude that the multiple regression model does not have a strong linear relationship among predicate variables.

**1. Mean of zero: The mean of each residual for each set of values for the predictor variables is zero (Chan, 2020).**

If someone views the residuals against the fitted values scatter plot, they should see an even distribution of values throughout the whole scatter plot. If the residuals for each set value of the predicated variables are evenly spread, then the mean of the variables is zero. The residual positive values should cancel the negative values if we take the sum of values. Given the predictor variables, this is a crucial indicator of a strong linear relationship throughout the regression model. In the scatter plot, there is a high concentration of data set values between $300,000 to $550,000 for the price range. However, it is evenly distributed on the positive and negative sides. Then we can say that the predictor variables have a strong linear relationship, and we see this in the residual against the fitted value scatter plot.

**2.  Independence: The residuals are independent (Chan, 2020).**

A sign of residual independence will be looking at the residual against fitted values and seeing if the values are scattered evenly throughout the scatterplot. Thus a strong indication of the independence of residual values.  No patterns in a plot of residuals versus time should exist, although our scatter plot is not the best example. However, we do see signs of independence of the residuals.

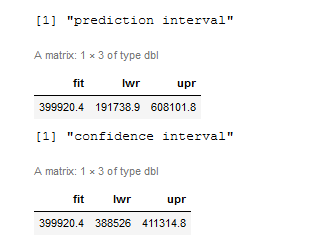
**3. Normality: The residuals of each set of values for the predictor variables form a normal distribution (Chan, 2020).**

Suppose there are residual against fitted values that the errors are normally distributed. There seem to be randomly distributed; no patterns in the scatter plot, which is a strong indication of normality distribution—looking for in a regression model with strong linear relationships among variables. In the Normal Q-Q Plot, there are normally distributed residuals. We see the data points are closely around the sloped line, indicating a strong linear relationship among the predicated variables. There is a strong linear relationship between the variables, which is an indication of normality.

**4. Constant variance: The residuals of each set of values for the predictor variables should have equal or similar variance (Chan, 2020).**  
 In the Normal Q-Q Plot, each residuals' data values are throughout the scatterplot, and the variance remains almost constant throughout the scatter plot. The common term for this condition is homoscedasticity. Moreover, this is another sign that the multiple regression model has a strong linear relationship among the residual variables.

### Evaluating Significance of Model

### Making Predictions Using the Model



The predicted housing price is $399,920.4.

95% Prediction Interval = ($191,738.9, $608,101.8)

95% Confidence Interval = ($388,526, $411,314.8)

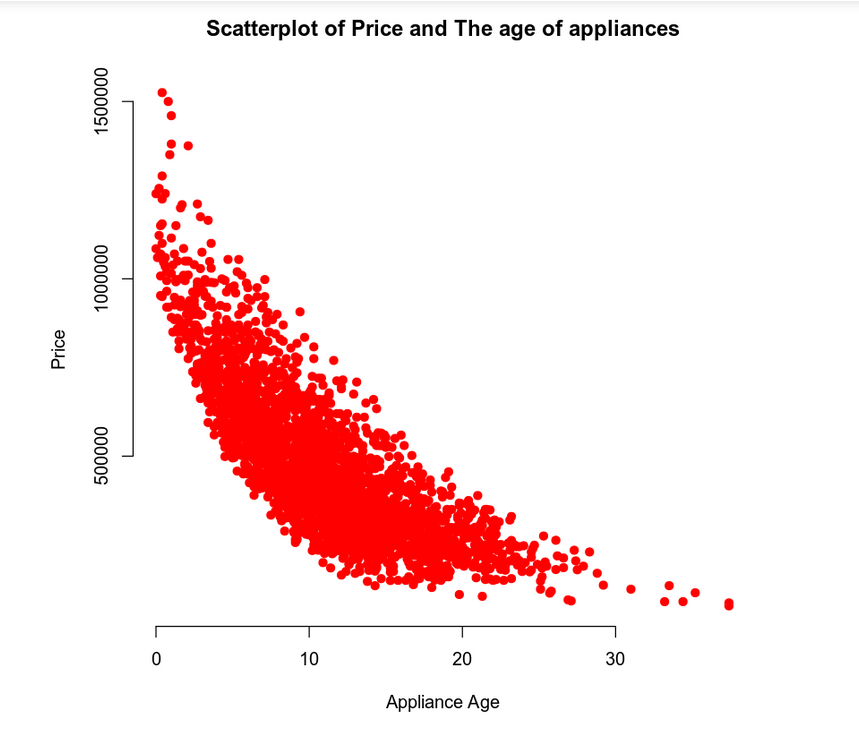
95 percent Prediction interval = ($191,738.9, $608,101.8). We are 95% confident that the predicted value for hosing price is $399,920.4 when sqft\_living is 2,150 sqft and grade is 7, and the number of bathrooms is 3, and the view is 0, and housing price is between $191,738.9 and $608,101.8.

95 percent Confidence interval = ($388,526, $411,314.8). Then we are 95% confident that the predicted value for hosing price is $399,920.4 when sqft\_living is 2,150 sqft and grade is 7, and the number of bathrooms is 3, and the view is 0, and housing price is between $191,738.9 and $608,101.8.

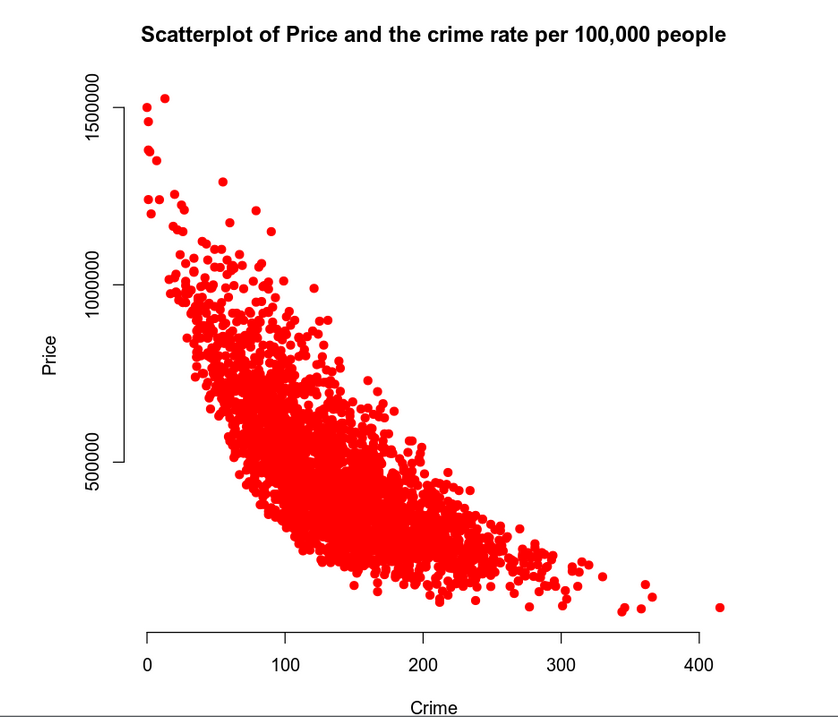
The prediction interval is wider because the prediction interval is essentially an interval around a predicted value. It is for one specific value. Then this gets plugged into the regression model equation. As a result, we get the prediction just for that one prediction using that one row, one specific set of values, which gives us the prediction interval. Now, the lower and an upper bound of the confidence interval is different, and this is the average prediction. If we had several different housing price instances, we take all those values of the data sets and the multiple values, then plug them into the equation. We can then create a lower and an upper bound for that entire sample, which would come out to be the confidence interval. The reason the prediction interval is wider is due to the added uncertainty. We are using smaller sample size, just one in the case of the confidence interval. However, the uncertainty is lower if we have a larger sample, just the confidence interval. For this reason, the prediction interval will always be wider than the confidence interval.

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

### Correlation Analysis



When analyzing the scatter plot between the price of a house and the age of appliances, we can see that directly it is not a  linear relationship. However, the curvature on the scatterplot indicates a parabola that corresponds to a quadratic situation, which means that by observation, it appears to be a second-order linear regression model. The concave-up parabola might fit our regression model; however, further, analyzation of the data sets is needed. However, by observing the scatterplot, I would conclude that it would be a second-order linear regression model.  We will do further analysis to determine the accuracy of which kind of regression model will fit the scatter plot the best to develop a regression model that determines the wage growth rate based on inflation. There appears to be a strong correlation in terms of a quadratic regression model between housing price and the age of appliances that merits more investigation.



When analyzing the scatter plot between the price of a house and the crime, that is, directly and not a linear relationship. However, the curvature on the scatterplot indicates a parabola that corresponds to a quadratic situation, which means that by observation, it appears to be a second-order linear regression model. The concave-up parabola might fit our regression model; however, we must further analyze the data sets. However, by observing the scatterplot, I would conclude that it would be a second-order linear regression model.  Further analysis is needed to determine the accuracy of which kind of regression model will fit the scatter plot. There appears to be a strong correlation in terms of a quadratic regression model between housing price and crime that merits more investigation.

### Reporting Results

|  |
| --- |
|  |
| The general form of the complete Second-Order Model (Two Quantitative Variables) Then, The general form of this regression model is:  y = Price  *x*1 = appliance\_age  x2= crime |
| The Interpretation of Beta Estimates  y = Price  *x*1 = appliance\_age  x2= crime |

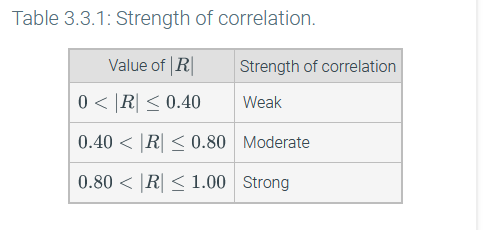


Table 1: Strength of correlation (Chan, 2020).

|  |
| --- |
| **The Coefficient of multiple determination and Adjusted Coefficient of multiple determination**  (R-squared) (Adjusted R-squared) |

The coefficient of multiple determination is the quotient of the fitted values variances and observed values of the dependent variable. The calculated coefficient of considerable determination for the regression model is  0.8088, which means that the second-order regression model explains 80.88% of the housing price variation as the response variable, appliance age, and crime rate as predictor variables, and the quadratic terms appliance2 and crime2.

Based on the value of adjusted R2, the proportion of variation explained by the estimated regression line is 0.8084, which means that 80.84% of the estimated regression equation fits or explains the second-order regression model explains the relationship for housing price as the response variable, appliance age and crime rate as predictor variables, and the quadratic terms appliance2 and crime2.

The estimated coefficient for the appliance age is x1 = -4.2456E04. However, the regression model's squared term (x1^2), the appliance variable x1, no longer characterizes a slope coefficient. The appliance age variable's estimated coefficient does not provide any helpful interpretation. For second-order regression models, data may have a nonlinear relationship. However, the quadratic relationship can be evaluated by adding the predictor variable's powers to the regression equation. Thus, a quadratic regression model includes a quantitative predictor, and this resulting scatter plot graph is nonlinear (Chan, 2020).

The estimated coefficient for the crime is x2 = -3.678E03. However, the regression model's squared term (x2^2), the crime variable x2, no longer characterizes a slope coefficient. The crime variable's estimated coefficient does not provide any useful interpretation. For second-order regression models, data may have a nonlinear relationship. However, the quadratic relationship can be evaluated by adding the predictor variable's powers to the regression equation. Thus, a quadratic regression model includes a quantitative predictor, and this resulting scatter plot graph is nonlinear (Chan, 2020).

The estimated coefficient for appliance age, the quadratic variable is x1^2 is 8.330E02. We can see that the coefficient's sign is positive and indicates a curve with an upward concavity. Quadratic variables are parabolas with either concave up or concave down characteristics, which would be a concave-up attribute in our case. Then the estimated coefficient of the quadratic variable x1^2 represents the rate of curvature for the regression model, which is what we are looking for.

The estimated coefficient for crime, the quadratic variable is x2^2 is 6.380E00. We can see that the coefficient's sign is positive and indicates a curve with an upward concavity. Quadratic variables are parabolas with either concave up or concave down characteristics, which would be a concave-up attribute in our case. Then the estimated coefficient of the quadratic variable^2 represents the rate of curvature for the regression model, which we are looking for.

The estimated coefficient for the interaction term between appliance age and crime is 1.390E01. Since the squared polynomial appliance age and age are both negative concave up, reflecting the interaction term. The significant interest is the two square dependent variables contributing to the overall regression model, which would also reflect that the interaction term captures any useful information between the two squared dependent variables. This would then have similar characteristics to a negative concave-up parabola, contributing to the graph's overall picture.

### Evaluating Model Significance

Is the model significant at a 5% level of significance?

: Model is not significant

: Model is significant

Then:

= 0

The coefficients of appliance age and crime rate as predictor variables, and the quadratic terms appliance2and crime2is 0 in the regression equation. Then under the null hypothesis, the model is not statistically significant.

The coefficients of either appliance age or crime rate as predictor variables or the quadratic terms appliance2 or crime2, or all are non-zero in the regression equation. Then under the alternative hypothesis, the model is statistically significant.

F-Statistic = 2272

P-value=2.2E-16 ≈ 0.00

P-value < = 0.05

The Null Hypothesis is rejected. Since the p-value is 0.00 and it is less than the level of significance of = 0.05, we can reject the null hypothesis and conclude that the regression model is statistically significant.

Which terms in the model are significant at a 5% level of significance? Carry out individual beta tests.

Applice Age: T-Statistic = -30.951, P-value=2E-16, This predictor is significant in the model since the p-value is less than the level of significance = 0.05.

Crime: T-Statistic = -24.827, P-value=2E-16, This predictor is significant in the model since the p-value is less than the level of significance = 0.05.

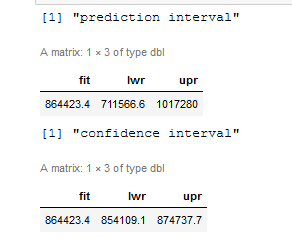
(appliance\_age)^2: T-Statistic = 10.501, P-value=2E-16, This predictor is significant in the model since the p-value is less than the level of significance = 0.05.

(crime)^2: T-Statistic = 8.782, P-value=2E-16, This predictor is significant in the model since the p-value is less than the level of significance = 0.05.

(appliance\_age)\*(crime): T-Statistic = 1.072, P-value=0.284, This predictor is not significant in the model since the p-value is greater than the level of significance = 0.05.

### Making Predictions Using the Model

What is the predicted price for a home that has one-year-old appliances and is in an area that has a crime rate of 81.02 per 100,000 individuals?



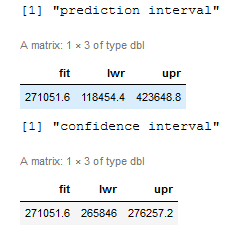
The predicted price of the house is $864,423.4.

95% Prediction Interval = ($711,566.6, $1,017,280)

95% Confidence Interval = ($864,423.4, $847,737.7)

The prediction interval is wider because the prediction interval is essentially an interval around a predicted value. It is for one specific value. Then this gets plugged into the regression model equation. As a result, we get the prediction just for that one prediction using that one row, one specific set of values, which gives us the prediction interval. Now, the lower and an upper bound of the confidence interval is different in the confidence interval, which is the average prediction. If we had several different instances, we take all those values of the data sets and the multiple values that are the same for the inflation rate, and then we plug them in the equation. We can then create a lower and an upper bound for that entire sample, which would come out to be the confidence interval. The reason the prediction interval is wider is due to added uncertainty. We are using smaller sample size, just one in the case of the confidence interval. However, the uncertainty is lower if we have a larger sample, just the confidence interval. For this reason, the prediction interval will always be wider than the confidence interval.

What is the predicted price for a home with 15-year-old appliances in an area with a crime rate of 200.50 per 100,000 individuals?



The predicted price of the house is $271,051.6.

95% Prediction Interval = ($118,454.4, $423,648.8)

95% Confidence Interval = ($256,846, $276.257.2)

The prediction interval is wider because the prediction interval is essentially an interval around a predicted value. It is for one specific value. Then this gets plugged into the regression model equation. As a result, we get the prediction just for that one prediction using that one row, one specific set of values, which gives us the prediction interval. Now, the lower and an upper bound of the confidence interval is different in the confidence interval, which is the average prediction. If we had several different instances, we take all those values of the data sets and the multiple values that are the same for the inflation rate, and then we plug them into the equation. We can then create a lower and an upper bound for that entire sample, which would come out to be the confidence interval. The reason the prediction interval is wider is due to added uncertainty. We are using a smaller sample of size, just one in the case of the confidence interval. However, the uncertainty is lower if we have a larger sample, just the confidence interval. For this reason, the prediction interval will always be wider than the confidence interval.

## 5. Nested Models F-Test

### Reporting Results

|  |
| --- |
|  |
| the general form of the complete Second-Order Model (Two Quantitative Variables) Then, The general form of this regression model is:  reduced model  y = Price  *x*1 = appliance\_age  x2= crime |
| The Interpretation of Beta Estimates    y = Price  *x*1 = appliance\_age  x2= crime |

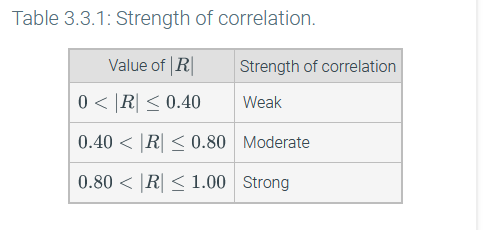


Table 1: Strength of correlation (Chan, 2020).

|  |
| --- |
| **The Coefficient of multiple determination and Adjusted Coefficient of multiple determination**  (R-squared) (Adjusted R-squared) |

The coefficient of multiple determination is the quotient of the fitted values variances and observed values of the dependent variable. The calculated coefficient of multiple determination for the regression model is  0.7995. It means that the regression model explains 79.95% of the housing price variation as the response variable, appliance age, and crime rate as predictor variables.

Based on the value of adjusted R2, the proportion of variation explained by the estimated regression line is 0.7993. This means that 79.93% of the estimated regression equation fits or explains the regression model explains the relationship for housing price as the response variable, appliance age, and crime rate as predictor variables.

### Evaluating Model Significance

Is the model significant at a 5% level of significance?

: Model is not significant

: Model is significant

Then:

= 0

The coefficients of appliance age and crime rate as predictor variables and interim variable is 0 in the regression equation. Then under the null hypothesis, the model is not statistically significant.

The coefficients of either appliance age or crime rate as predictor variables or interim variables are non-zero in the regression equation. Then under the alternative hypothesis, the model is statistically significant.

F-Statistic = 3573

P-value=2.2E-16 ≈ 0.00

P-value < = 0.05

The Null Hypothesis is rejected. Since the p-value is 0.00 and it is less than the level of significance of = 0.05, we can reject the null hypothesis and conclude that the regression model is statistically significant.

Which terms in the model are significant at a 5% level of significance? Carry out individual beta tests.

Appliance Age: T-Statistic = 130.27, P-value=2E-16, This predictor is significant in the model since the p-value is less than the level of significance = 0.05.

Crime: T-Statistic = -49.65, P-value=2E-16, This predictor is significant in the model since the p-value is less than the level of significance = 0.05.

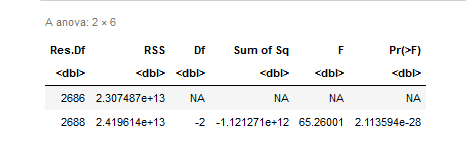
(appliance\_age)\*(crime): T-Statistic = 31.63, P-value=0.284, This predictor is significant in the model since the p-value is less than the level of significance = 0.05.

### Model Comparison

We have these two models, and we have to decide whether we should use the complete model or the reduced model to predict the response variable for the regression model. Moreover, the additional term here adds value and predicts the response variable's significance to the regression model. If it is significant, we would like to use the complete model to predict the complete regression model. However, If it is not significant, then we should use the reduced model.

|  |
| --- |
| Complete Second-Order Model with Interaction Model  Complete model  reduced model  y = Price  *x*1 = appliance\_age  x2= crime |

We have the complete second-order model with interaction terms which is considered a complete model. In the regression model, appliance age and crime are the dependent variables, and housing price is the independent variable. Can we determine if the complete or reduced models are needed for a regression model given the significance level? We need to use the F-test to determine which model to use.



Is the model significant at a 5% level of significance?

: Model is not significant

: Model is significant

Then:

0

The coefficient for the quadratic terms x1^2 and x2^2 is 0. The reduced model is sufficient

The coefficient for the quadratic terms x1^2 and x2^2 at least one is non-zero. The complete model should be used. Then under the alternative hypothesis, the model is statistically significant.

F-Statistic = 65.26001

P-value = 2.1135E-28

P-value < = 0.05

The Null Hypothesis is rejected. Since the p-value is 2.1135E-28 and it is less than the level of significance of = 0.05, we can reject the null hypothesis. Then we conclude that the x1^2 and x2^2 terms should be used in predicting the wage growth rare. Therefore, the complete model should be used.

## 

## 6. Conclusion

## When analyzing a huge amount of data, it is essential to find relationships between different variables of data value sets. Both models developed in this homework assignment are similar in the values and the graphs—the complete second-order regression model with quantitative variables. Overall, neither significantly impacted the regression model's accuracy, given the particular variables added to the model. However, for the complete second-order model (Two Quantitative Variables), the coefficient of multiple determination for the regression model is 0.8088, which means that 80.88% of the housing variation price as the response variable, appliance age, and crime rate as predictor variables. This coefficient of multiple determination was then better than the coefficient of determination for just the nested models. The other equation had, which was that 79.88% of the variation. There is only a difference of 1.00 percent. Adding two quantitative predictors is essential when someone has categorical comparisons in a regression model. A quantitative predictor is a binary value set added to the regression model, which gives conditions to the overall regression model.

However, when contributed to the overall regression model, there was no huge difference in the regression model's overall wage growth rate. However, I think more variables need to be analyzed and evaluated to find which variable will produce the highest probability of finding the regression model with the most linear relationship between variables. The key is to understand when a data set has a linear relationship or nonlinear relationship to determine what type of regression model to use. Having a second-order regression model can help the real estate agents to better price the value of a house.

## 7. Citations

Chan, C., Berrier, H., Pardoe, L., & Sturdivant, R. (2020). zyBook for Applied Statistics II for Science, Technology, Engineering, and Math (STEM).