Property Value, Community Services, Crime, and Demographics

DATA 603 Project

Adam Hachmi

Zhenyu (Mike) Li

Robin (Scott) Pellegrino

Yue (Flora) Yang

Introduction

Having plenty of nutritious food, clean drinking water, and adequate shelter are basic needs every human being requires to survive. It is no wonder then that many Canadians want to live in and own their very own home, however, buying one can be costly and Canadians can spend up to 80% of their income on housing and utilities combined (RBC, 2017). This means buying a house is a major financial decision that cannot be taken lightly; not only does it give one a place to live, but it can also act as an investment for uncertain financial futures ahead. What does all this mean? Buying a home at a great price and at the right time and right place is something all prospective buyers can appreciate.

This report aims to explore the average property value for communities in the City of Calgary and how it is potentially influenced by factors as follows: crime rate (by type of Crime committed), the number of community services (such as clinics, community centres, hospitals, etc.), community sector (north, east, south, west, northwest, etc.) and community demographics (by age and total population). When analysis was started for this project it was expected that a model with linearity, homoscedasticity, normality, and multicollinearity assumptions would be met and have crime, number of community services, and demographics as strong predictors. More specifically, it was expected that property values would be lowered by the presence of higher crime rates and increased in communities with a higher number of services and a larger population.

Datasets

To conduct this study, datasets detailing crime, demographics, community services and property assessments within Calgary were needed. Through the Calgary Open Data source, the following datasets were retrieved.

- *Crime and Disorder Statistics*
- *Census By Community* – 2019
- *Community Services*
- *Property Assessment*

As stated, each of these datasets were publicly available from the city of Calgary Open Data source and free to use as desired (City of Calgary, 2020). To appropriately utilize the information within each dataset, each was cleaned as follows.

Crime and Disorder Statistics (City of Calgary, 2018a):

Entries occurring outside of the year 2019 were removed. Many of the original variables were not relevant to this study and were therefore removed. The variables which were kept were *Community name*, *crime category*, and *Crime count*. This data was then sorted and grouped by community, producing a

table which consisted of crime types as columns and community names as rows, with each element being the number of times a specific crime was recorded in that community.

Census by Community- 2019 (City of Calgary, 2019):

As multiple years of information were available, the year 2019 was chosen to allow the study to hold more relevance. When determining which demographic criteria to include, it was found that gender proportions between communities did not vary significantly and therefore only age was included. Ages were grouped into ranges of $0-24$, $25-64$, and $65+$ years old. Once grouped, the data was assigned to a new table where the columns were the defined age groups, the rows were communities and each element was the count of persons in that age range for a community. Additionally, the community classification (Residential, industrial etc.) and sector (North, Northwest, Central, etc.) were also maintained.

Community Services (City of Calgary, 2016):

This dataset contained information of each community service facility within the City. This included community centres, court houses, clinics, etc. Due to the relatively low number of facilities, it was decided that the number of facilities per community would be used, as opposed to the different types. Therefore, information from this dataset was extracted into a new table containing the number of services per community.

Property Assessments (City of Calgary, 2018b):

Within this dataset was every property assessment completed by the City for the years 2005 to 2020. As only 2019 values were of interest, all other assessments were omitted. Of the remaining data, the community which the property was located, and the assessed value were the only parameters of interest. This data was extracted into a new table and reduced so that each entry was the average property value for a given community.

With each of these datasets managed for the desired varaibles, they were combined into a single table so that each row contained the necessary information for a single community. As a summary, the following lists all variables used in the modelling process.

Response Variable

[1]$Property\ Value,\ (Canadian\ Dollars\ (CAD))$

Independent Variables

[2] – $Number\ of\ Services\ (Count)$

[3] – $Community\ Classification\ (Class)$

[4] – $Community\ Sector\ (Sector)$

[5] – $Street\ Robbery,\ Instances\ of\ Street\ Robbery\ (Count)$

[6] – $Theft\ of\ Vehicle,\ Instances\ of\ Theft\ of\ Vehicle\ (Count)$

[7] – $Theft\ from\ Vehicle,\ Instances\ of\ Theft\ from\ Vehicle\ (Count)$

[8] – $Commercial\ Break\ \&\ Enter,\ Instances\ of\ Commercial\ Break\ \&\ Enter\ (Count)$

[9] – $Residential\ Break\ \&\ Enter,\ Instances\ of\ Residential\ Break\ \&\ Enter\ (Count)$

[10] – $Assault\ (Non-Domestic),\ Instances\ of\ Assault\ (Count)$

[11] – *Violence Other* (*Non – Domestic*), *Instances of Violence* (*Count*)

[12] – *Commercial Robbery*, *Instances of Commercial Robbery* (*Count*)

[13] – *Physical Disorder*, *Instances of Physical Disorder* (*Count*)

[14] – *Social Disorder*, *Instances of Social Disorder* (*Count*)

[15] – *Age* $0 - 24$, *Population Between Ages* 0 *to* 24 (*Count*)

[16] – *Age* $25 - 64$, *Population Between Ages* 25 *to* 64 (*Count*)

[17] – *Age* > 65, *Population Above the Age of* 65 (*Count*)

[18] – *Population*, *Total Community Population* (*Count*)

Methodology

Data for the model variables was cleaned and aggregated using R Studio from the datasets as outlined above into the described combined dataset.Once the relevant variables for the model were loaded available for manipulation, the best fit model was developed and used to make predictions regarding the average property value within a City of Calgary community for the year 2019.

Modeling Plan

The model for this project was developed using the methods outlined throughout DATA 603 (Ngamkham, 2020) using data from the cleaned dataset *combined.csv*. The data was first tested for outliers. Outlier detection was completed using Cook's distance and leverage points, with final determination using a Cook's distance, D_i, greater than 1. Prior to elimination, outliers were investigated to determine their validity as well as significance to the future model. Once removed, multiple linear regression methods were employed to fit the remaining data.

As there was significant potential for variables to be related, multicollinearity was tested by computing the variance inflation factor (VIF) and assessing those with high VIF scores (> 5.0). This method determines the strength of correlation between independent variables and results in a large value when the coefficient of determination between variables,$R^2_{X_j|X_j}$, is large.This is shown within the VIF equation as follows.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_j}}$$

For the variables which exhibited VIF values larger than 5, it was determined that combining them would reduce the significance of the variable and therefore selective variables were removed for future model estimation.

Once multicollinearity was managed, the *Stepwise Regression*, *Backward Elimination*, *Forward Selection*, and *All Possible Regression* procedures were used to collectively determine a best fit first order model. This reduced model was then compared to the full model using a partial F-test (ANOVA table).

As it was likely that each of the above methods would result if a different model estimation, it was decided to use the results of the *All Possible Regression* procedure, and base the model decision off of the Mallow C_p (C_p) criterion and Akaike's Information Criterion (AIC). Ideally, the selected model would produce a C_p value of $p + 1$, where p is the number of first order predictor variables, as well a low AIC value, relative to other models. Additionally, the other regression procedures were used to support the chosen model, that is, if a specific variable appeared in all models, it was likely a significant predictor.

With a reduced model containing only statistically significant predictors, the presence of interaction effects and higher order variables was investigated. To verify the inclusion of possible terms, individual t-tests, paired scatter plots, and partial F-tests were conducted.

Once the best fit model was found, the remaining assumptions made to produce the model were tested. Again, these assumptions were:

1.$VariableIndependence$

2.$Linearity$

3.$Data\ \ Normality$

4.$Equal\ \ Varience\ \ (Homoscedasticity)$

To determine if the assumptions of Variable independence and linearity hold, a plot of residuals vs fitted data for the best fit model was analyzed. For the data to meet these assumptions, the plot must not show definitive patterns, but also be distributed rougly evenly about the $y = 0$ line.

When investigating data normality, there are several possible methods. In this study, the distribution of residual values, $Q-Q$ plots and the Shapiro-Wilk test were utilized. The distribution or residuals and $Q-Q$ plots provide visual representations from which normality can be qualitatively assessed. The Shapiro-Wilk test uses the residuals of the model to quantitativly determine if the data is significantly normally disributed.

The final assumption to be investigated is the assumption of constant variance, or *homoscedasticity*. This was tested both visually and empirically, using the plot of residuals vs fitted, as well a the Breusch-Pagan test. When analyzing the plot, patterns, groupings, or conical trends indicate the data may not satisfy this assumption. As this only provides a qualitative assessment of the assumption, the Breusch-Pagan test was used to determine, quantitatively, the degree of homoscedasticity within the data.

Likely, some of these assumptions do not hold for the data. Therefore, should it be necessary, a model transformation will be used. As the assumptions of normailty and homoscedasticty are to be the most violated, the Box-Cox transformation was determined to be the appropriate transform method. This method transforms the response variable by an exponential factor. This factor, λ is estimated based on the method of maximum liklihood, and transforms the response variable as follows:

$$Y_i^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda} = \beta_0 + \beta_1 X_1 + ...$$

Once transformed, each assumption will be re-assessed and the overall validity of the model re-evaluated.

All modeling code can be found within Appendix A.

Results & Discussion

Prior to investigating the significance of each variable on assessed property value, the data was assessed for outliers. As stated previously, a Cooks distance cut-off of 1 was used. Three observations were detected and removed from the data, as illustrated within Figure 1 below.

Next, each variable was assessed for the presence of multicollinearity. This was completed using the VIF method. Investigating the results indicated that there was strong multicollinearity between the three age ranges (age_0, age_25,age_65) and population (pop). This was expected as population was determined by summing each of the age ranges. Removing population as a contributing variable reduced the multicollinearity within the data. Further multicollinearity was observed between age_0 and age_25. As age_25 was believed to be more significant in the model, age_0 was also removed as a contributing variable. With this, no further multicollinearity was observed. The linearity between these 4 variables can be seen illustrated within Figure 2, below

With the remaining variables, a full first order model was produced. This model include all variables, with the exception of *pop* and age_0, as described above.

As mentioned, several methods were used to estimate which variables were significant in predicting assessed property value. Initially, individual t-test values and a partial F-test method was used to estimate a model.

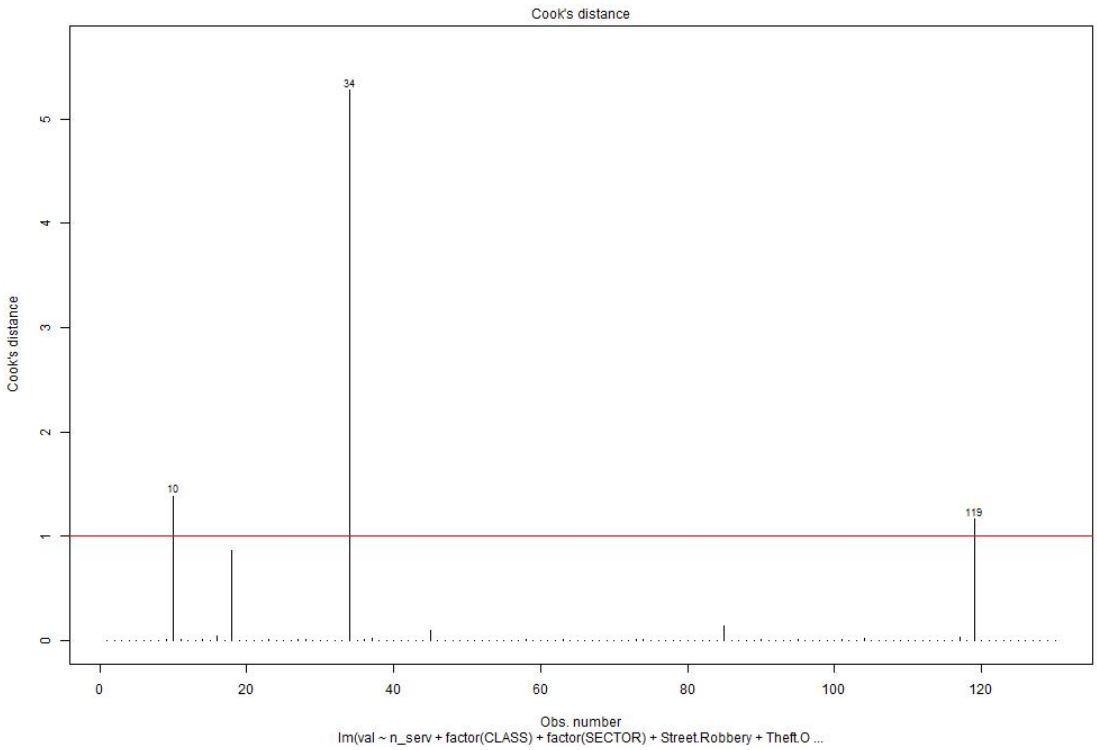


Figure 1: Cook's Distance plot identifying outliers.

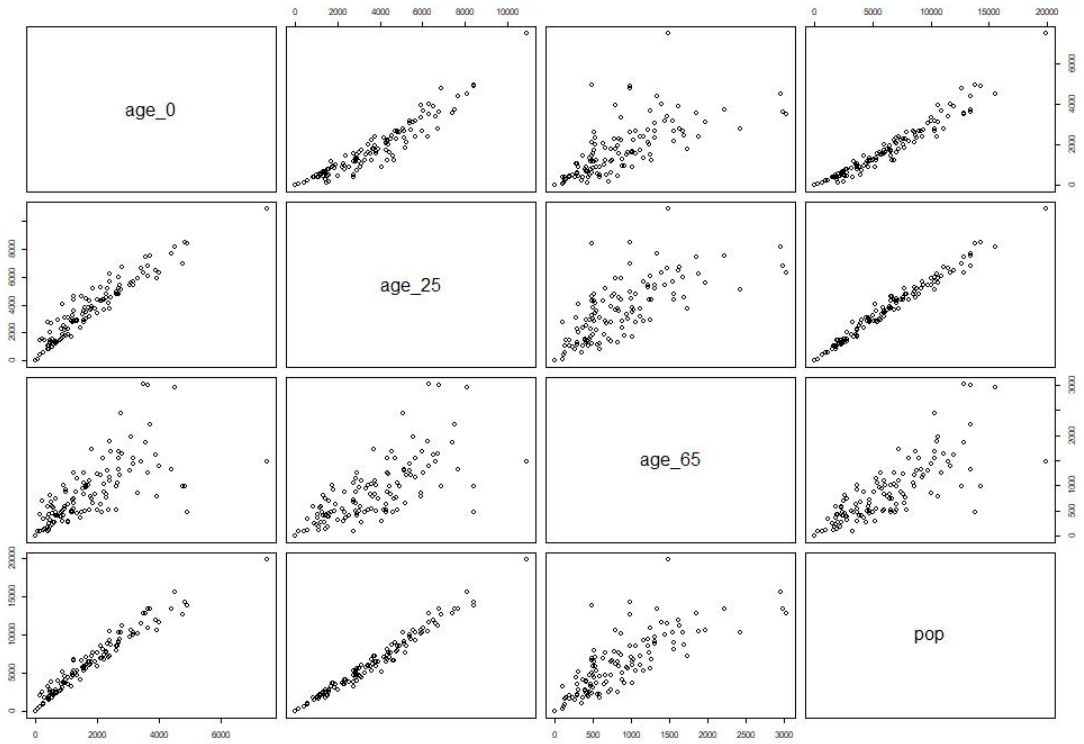


Figure 2: Pairs plot of Variables with Multicollinearity.

Using an α value of 0.05, it was found that $Class$,$Assault_{NonDomestic}$,Age_{25-65}, and $Social\ \ Disorder$ were the only significant variables. Although some initially exceed α within the full model, when variables were removed, their significance became relevant. To validate the removal of variables, the Partial F-test was used, producing the following hypotheses.

$$H_0 : The\ removed\ variables\ are\ not\ significant\ in\ the\ model\ estimation,\ \ \beta_i = \beta_j = 0$$

$$H_A : At\ least\ one\ of\ the\ removed\ variables\ is\ significant\ in\ the\ model\ estimation,\ at\ least\ one\ \beta_i \neq 0$$

Completing the Partial F-test resulted in an F-statistic of 0.5375 and a p-value of 0.9277. This indicates that the null hypothesis is to be accepted, supporting that the removed variables are not likely to be significant when estimating property value. This model produced an adjusted coefficient of determination (R^2_{adj}) of 0.6558, and and error (RMSE) of $\$1,122,000$. Given the data, this R^2_{adj} was somewhat larger that expected, and indicated that a measurable relationship may be estimated should assumptions hold. However; the magnitude of the RMSE value was unsettling as it is almost three times as large as some homes. This provided motivation to determine alternate improved models.

The next regression method utilized were the stepwise, forwards, and backwards regression methods. As the names suggests, these methods step through each variable to determine the variables which meet criteria and are to be kept. First, the results of the stepwise method determined that only $Social\ \ Disorder$, $Violence_{Other}$, $Class$, and $Age_{(25-65)}$ were significant estimators. This model produced an R^2_{adj} of 0.6559, and an RMSE of $\$1,122,000$. Similarly, the forward regression model determined that the variables $Social\ \ Disorder$, $Violence_{Other}$, $Class$, and $Age_{(25-65)}$ were significant. Therefore, this model produced the same R^2_{adj} and RMSE values. The backwards regression found that the variables $Social\ \ Disorder$, $Assault_{NonDomestic}$, $Class$, and $Age_{(25-65)}$ were significant when estimating property value. This model produced R^2_{adj} and RMSE values of 0.6558 and $\$1,122,000$ respectively.

The final regression method utilized was the *all possible subsets* method. Figure 3, below, illustrates the change of Mallows C_p criterion, Akaike's Information Criterion (AIC), and both adjusted and non-adjusted coefficients of determination, R^2, with models of an increasing number of variables. It was decided that the model consisting of the 4 best fitting variables would be utilized for further fitting. This model was chosen based on the relatively low C_p value of -3.8849, and AIC value of 3906.6. Due to the negative C_p value, more weight was given to the AIC value when deciding the best fit model, as the available smaller C_p values would introduce more bias due to overfitting. It was also found that this model had the largest R^2_{adj} value of 0.6559, further supporting the decision to move forward with it as the best fit model. This model determined that the variables $Class$, $Social\ \ Disorder$, $violence_{Other}$, and $Age_{(25-65)}$ were significant. As this is the same model as the stepwise and forward regression model, the R^2_{adj} and RMSE values were, again, 0.6559 and $\$1,122,000$ respectively. Verifying these results with a partial F-test, the same hypotheses as above were tested, finding a p-value of 0.9285, supporting the appropriateness of determined model as the null hypothesis, H_0, is to be accepted. Therefore, the first order regression model is as follows.

$$\hat{Value} = 97,204 \times Violence_{Other} - 1,436,189 \times Social\ \ Disorder - 164 \times Age_{(25-65)} + \phi$$

where

$$\phi = \begin{cases} 20,231,316 & if\ \ Class = Industrial \\ 22,385,987 & if\ \ Class = Major\ \ Park \\ 18,000,327 & if\ \ Class = Residential \end{cases}$$

As the model chosen had a C_p value significantly lower than the ideal $p+1$ value, -3.9 vs. 5, it is likely that the model introduces a measurable amount of bias into the property value estimation. Additionally, due to the large number of t-tests conducted in this method, there is a likelihood that type I error was introduced into the model.

The regression coefficients in this model are somewhat unexpected. It was thought that an increase of any crime type would decrease the property value, however it can be seen that an increase of Violence appears to also increase property value, if all else is held constant. Additionally, a larger number of working aged

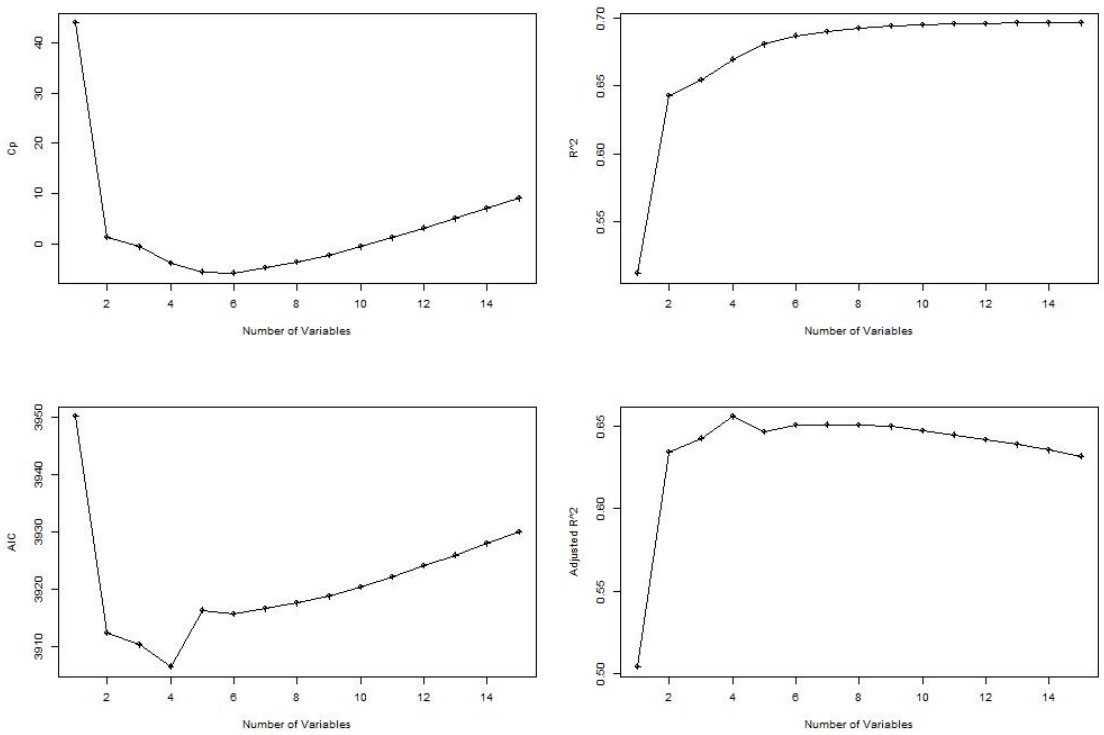


Figure 3: C_p, AIC, R^2 and R^2_{Adj} values of all possible regression.

persons (Aged 25-65) appears to also decrease the property value, which is somewhat unexpected. A trend which was expected was that properties within areas classified as parks had a higher value than industrial and residential. Further, a residential classification will have a lower value than both industrial and major park, if all else is held constant.

In attempts to improve the fit of the model further, the interaction effects between each variable were investigated. Again, using an α value of 0.05, it was found that $Class$ and $Social\ \ Disorder$, and $Class$ and $Violence_{NonDomestic}$ had statistically significant interaction effects. Verifying the results against a model containing all possible interaction terms, an F-statisic and p-value of 0.379 and 0.6854, respectively, were determined. These values suggest that we fail to reject the following null hypothesis, H_0, indicating that the retained interaction terms are appropriate.

$$H_0 : The\ \ excluded\ \ interaction\ \ terms\ \ are\ \ not\ \ significant\ \ in\ \ the\ \ model\ \ estimation,\ \ \beta_i = \beta_j = 0$$

$$H_A : At\ least\ one\ of\ the\ excluded\ intercation\ terms\ is\ significant\ in\ the\ model\ estimation,\ at\ least\ one\ \beta_i \neq 0$$

With these interaction terms, the updated regression model becomes:

$$\hat{Value} = 601,057 \times Violence_{Other} + 575,389 \times Social\ \ Disorder - 141 \times Age_{(25-65)} + \phi$$

where:

$$\phi = \begin{cases} -3,993,516 & if\ \ Class = Industrial \\ 30,019,518 - 2,701,952 \times Social\ \ Disorder - 3,383,610 \times Violence_{other} & if\ \ Class = Major\ \ Park \\ -153,792 - 479,214 \times Social\ \ Disorder - 565,410 \times Violence_{Other} & if\ \ Class = Residential \end{cases}$$

Investigating higher order terms, using the "*pairs*" plots, no concavity was observed, as is shown in Figure 4 below. Therefore, no higher order terms were included in the model estimation. As a result, the above estimation was determined to be the best fit model for the data. This model produced an R^2_{adj} of 0.8494, and an RMSE of $\$742,200$. Compared to the first order model, including the interaction terms produced an evident improvement. Examining the regression coefficients, the expected influence appears to be more appropriate. It can be seen that occurrences of crimes have less of an impact on industrial classified areas. This may be a result of city zoning and the limited number of areas industrial facilities may be constructed. It can also be seen that crime occurrences have the most influence on the value of major park classifications. This seems rational as it is believed that the value of parks would be tied to how likely individuals are to frequent it, and fewer people are going to go spend time where many crimes are committed. However, this is purely speculation.

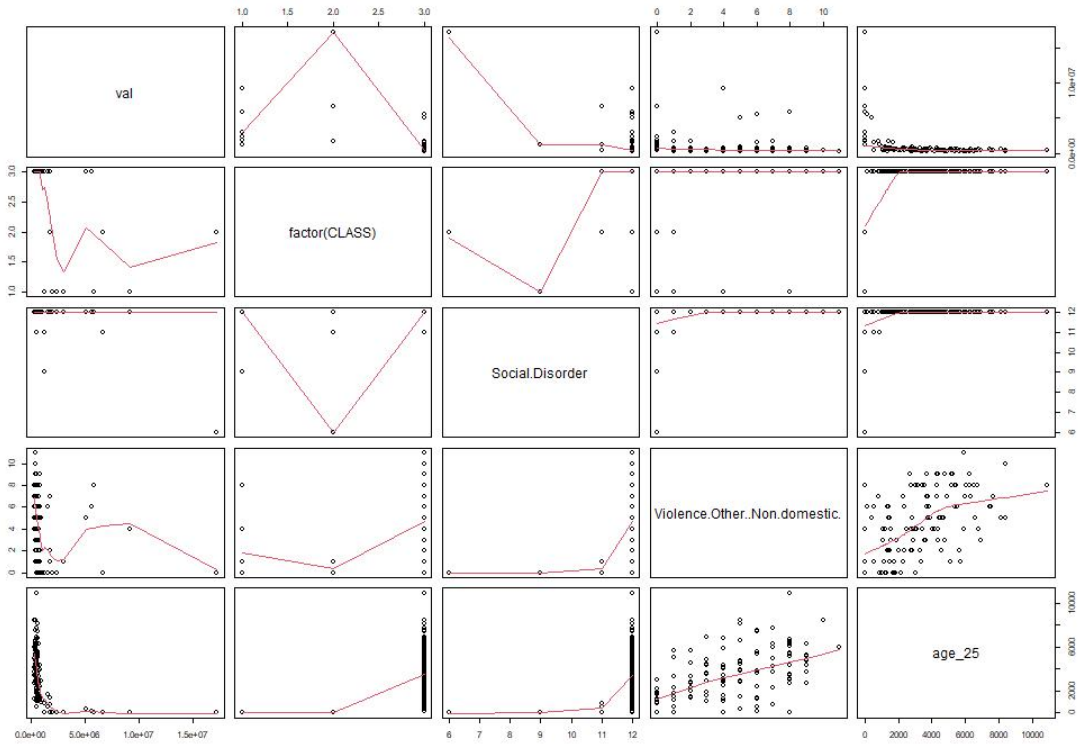


Figure 4: Paired scatter plots used to determine if higher order terms are necessary.

For the model estimation to hold any significance, the assumptions made when producing the model must be assessed. As stated, these assumptions are *Variable Independence*, *Linearity*, *Normality*, *No Multicollinearity*, and *Homoscedasticity*. As values within the data are not time dependent, each value corresponds to the entire year, it can be assumed that the variables are independent. As has been shown previously, multicollinearity has been addressed and managed. This leaves the assumptions of *Linearity*, *Normality* and *Homoscedasticity* to be investigated. As can be seen in the residuals vs fitted plot, shown in Figure 5, the data appears to rest around the $y = 0$ line. The red line illustrates that the data appears to be sufficiently linear. To investigate the assumption of normality, a histogram, Q-Q plot, and the Shapiro-Wilk test were used. First observing the distribution of the residuals, shown as Figure 6, it can be seen that the values appear mostly normal, although there are points which are seen to fall at extreme values outside of the majority. The $Q - Q$ plot, shown as Figure 7, further supports these observations. The majority of points follow the diagonal linear, with some "extreme" points on each end. Finally, the Shaprio-Wilk test statistically quantifies the normality of residuals within the data using the following hypotheses.

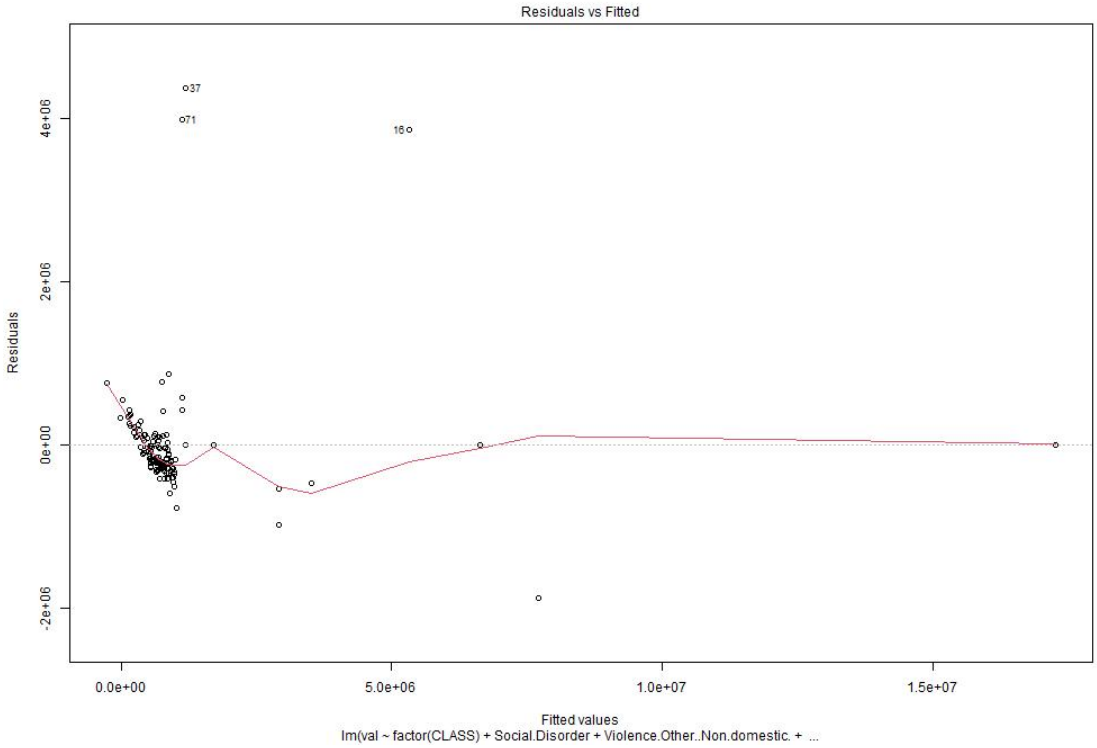


Figure 5: Plot of residuals vs fitted data to visualize linearity and Homoscedasticity.

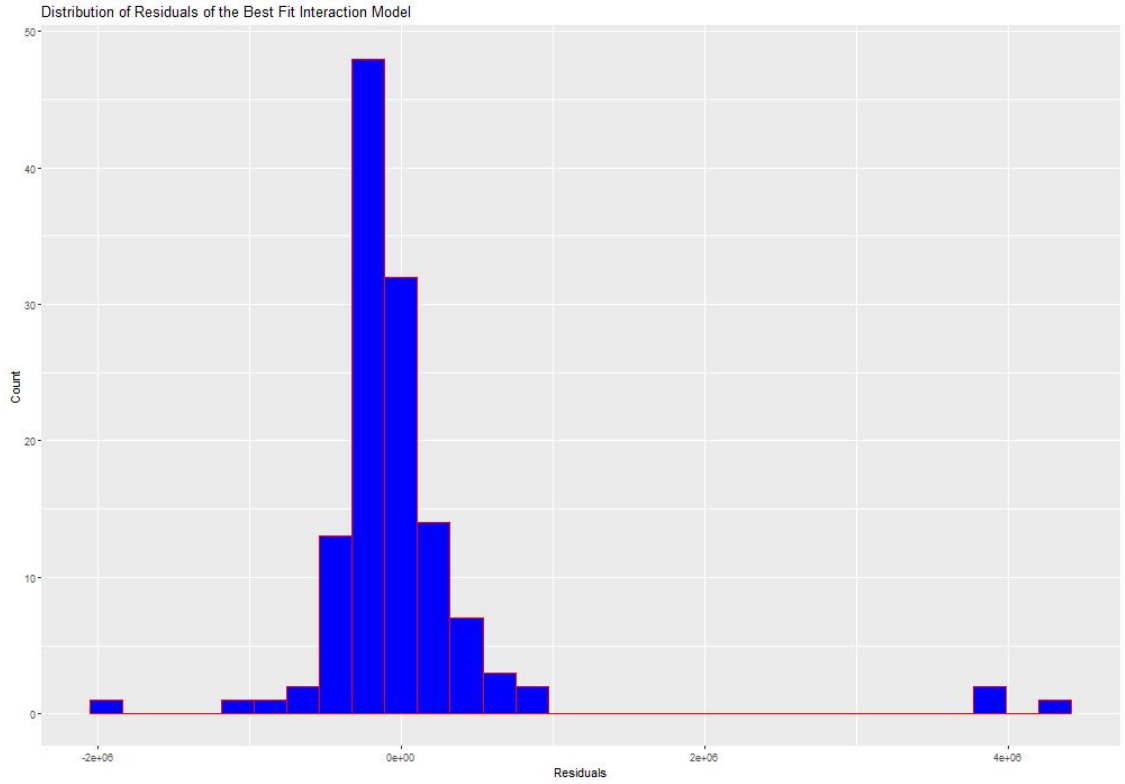


Figure 6: Histogram illustrating the distribution of the residuals.

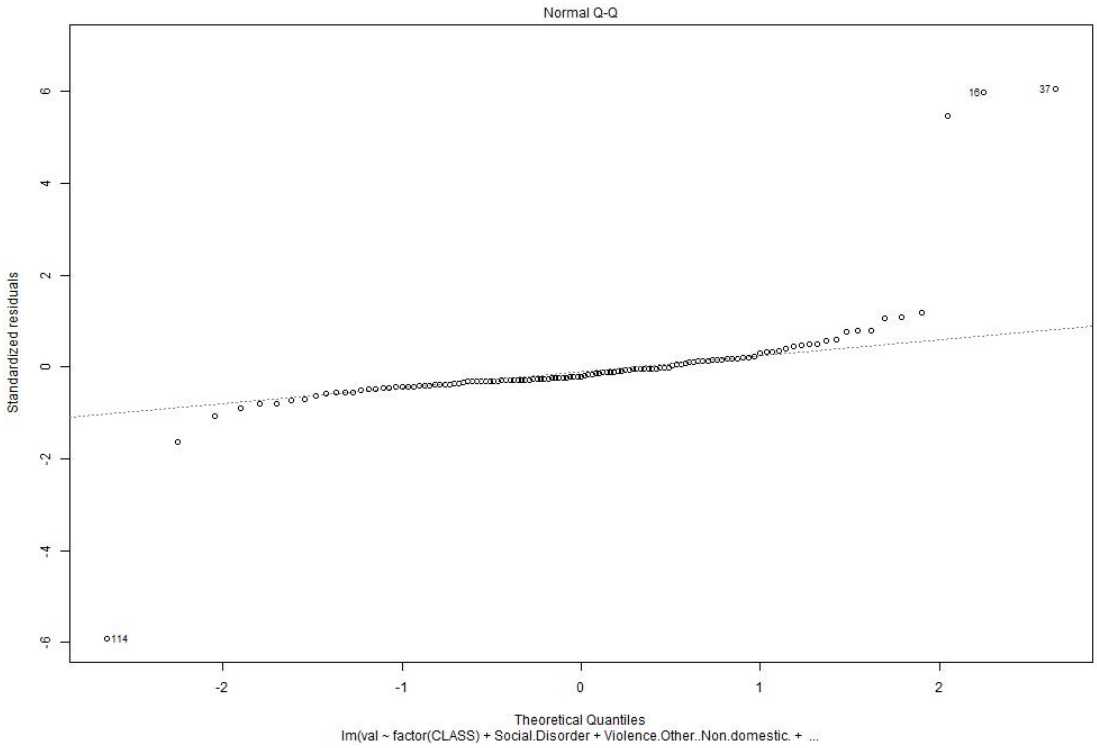


Figure 7: Q-Q plot illustrating normality within the data.

$$H_0 : The\ data\ is\ significantly\ normally\ distributed.$$

$$H_A : The\ data\ is\ \mathbf{NOT}\ significantly\ normally\ distributed.$$

From conducting the Shapiro-Wilk test, a p-value of $< 2.2e-16$ was found. From this value, the null hypothesis, H_0 is to be rejected in favor of the alternative, H_A. This suggests that the data is not significantly normally distributed, and the normality assumption does not hold.

Turning to the assumption of homoscedasticity, the variance of the error terms is assumed to have constant variance. Returning to the residuals plot, Figure 5, no discernible pattern, nor "cone" is visible. Additionally, the point appear to be distributed fairly evenly about $y = 0$. To test for homoscedasticity, or the presence of heteroscedasticty, the Breusch-Pagan test was utilized. This test utilized statistical parameters of the model ($\chi^2 test$) to determine if homoscedasticity is present. The results of the Breusch-Pagan test are used to evaluate the following hypotheses.

$$H_0 : Heteroscedasticiy\ is\ not\ present\ in\ the\ model,\ the\ data\ is\ homoscedastic,\ \sigma_i^2 = \sigma_j^2 = \sigma_k^2$$

$$H_A : Heteroscedasticiy\ is\ present\ in\ the\ model,\ the\ data\ is\ \mathbf{NOT}\ homoscedastic,\ at\ least\ one\ \sigma_i^2 \neq \sigma_j^2$$

From conducting the test on the best fit model, a p-value of 0.01197 was determined. As this is less than the pre-defined α value of 0.05, the null hypothesis, H_0, is rejected in favor of the alternative, H_A, implying that the data is not homoscedastic, and the assumption does not hold.

Based on these results, it is not possible to claim the the estimated best fit model indeed fits the data. In attempts to remedy these assumption shortcomings, a transformation of the response variable was conducted. As the assumptions of homoscedasticy and normality are the most significantly broken, the Box-Cox

transformation was employed. It is reasonable to utilize this method as all values of the response variable (assessed value) are greater than 0. It was found that a transformation power, λ, of -0.73232 was most suitable for the fitted data. The plot illustrating maximum likelihood is show as Figure 8. Applying the Box-Cox transformation, the transformed regression model was determined to be:

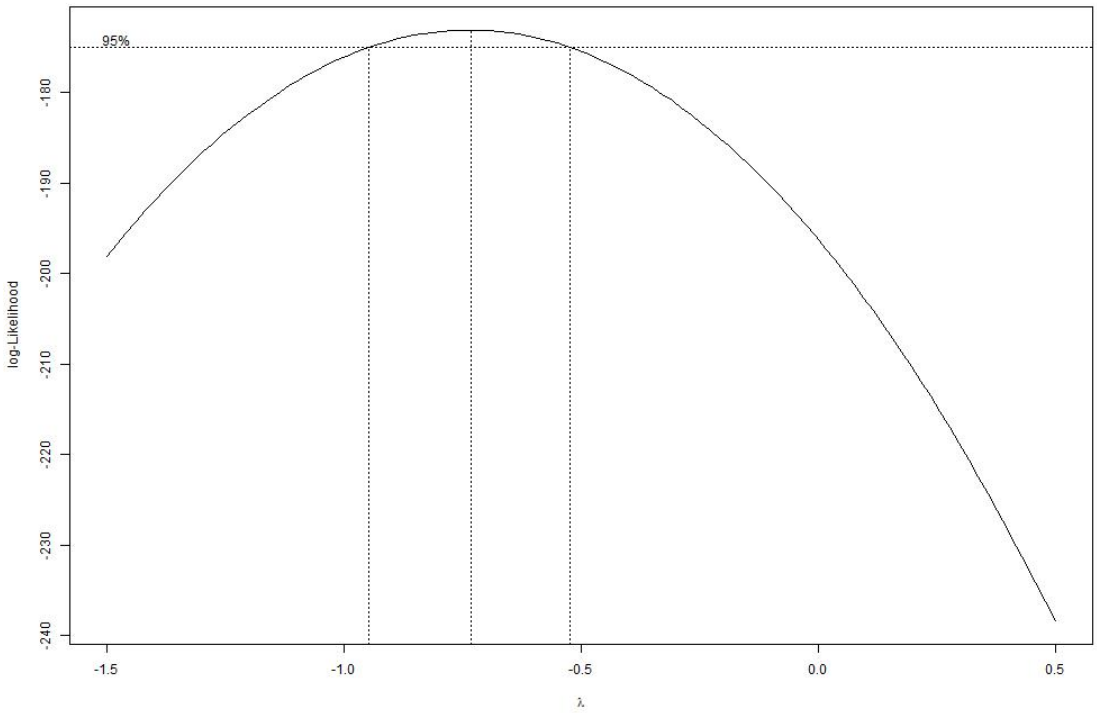


Figure 8: Box-Cox Maximum likelihood λ value estimation.

$$\hat{Value}^{-0.73232} = 2.266e-6 \times Violence_{Other} + 6.677e-6 \times Social\ \ Disorder - 5.290e-9 \times Age_{(25-65)} + \phi$$

where:

$$\phi = \begin{cases} 1.365 & if\ \ Class = Industrial \\ 1.365 - 8.067e-6 \times Social\ \ Disorder - 2.416e-5 \times Violence_{other} & if\ \ Class = Major\ \ Park \\ 1.365 - 4.578e-6 \times Social\ \ Disorder - 2.717e-6 \times Violence_{Other} & if\ \ Class = Residential \end{cases}$$

This model produced an R^2_{adj} of 0.4693, and RMSE of $2.094e-5$. Compared to the non-transformed model, this R^2_{adj} appears inferior. To determine if the decreased R^2_{adj} value has worth, the assumptions must be re-evaluated. As the Box-Cox transform is intended to primarily alter the status of homoscedasticity and normalily, these two conditions were re-investigated. As other changes may occur within the data during the transform, linearity was also checked. From the residuals plot presented in Figure 9, it can be seen that the residuals are more dispersed, compared to the non-transformed model. From the plot, it is difficult to determine if the change in linearity is an improvement over the non-transformed model.

Inspecting the residual distribution and $Q-Q$ plot, presented as Figures 10 and 11, there appears to be a noticeable improvement in terms of normality. The Shapiro-Wilk results, however, return a p-value of 0.04604. With the same normality hypotheses of

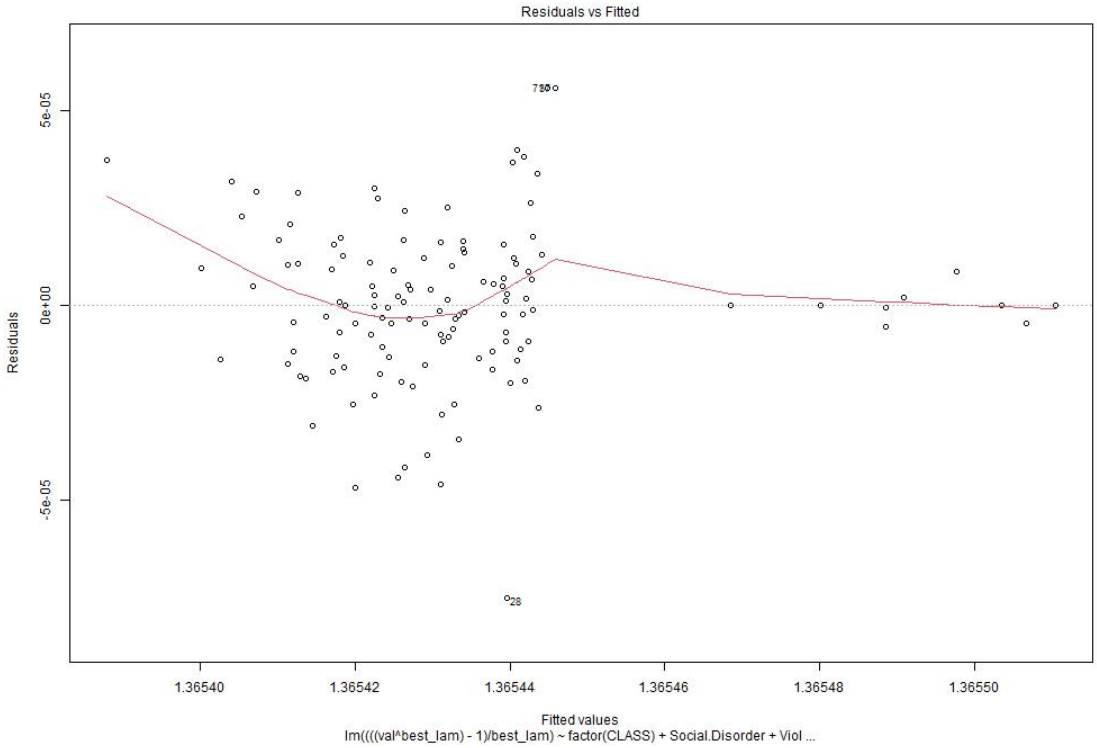


Figure 9: Plot of transformed residuals vs fitted data to visualize linearity and Homoscedasticity.

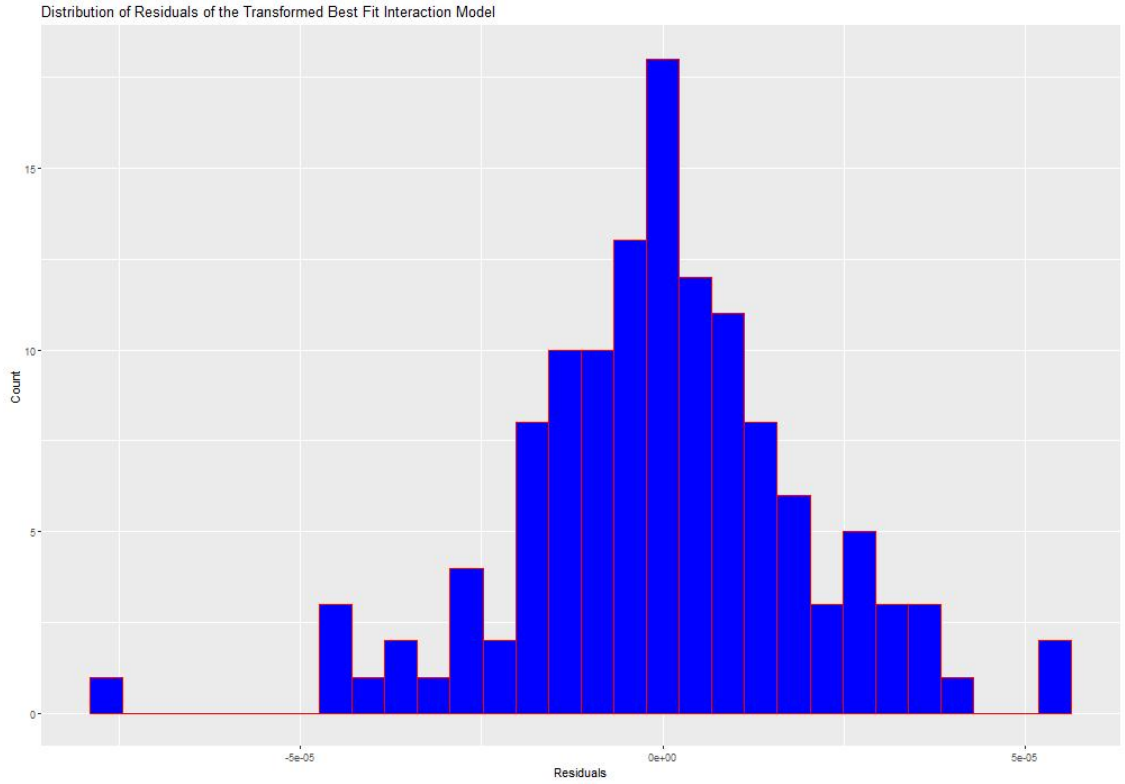


Figure 10: Histogramillustrating the distribution of transformed residuals.

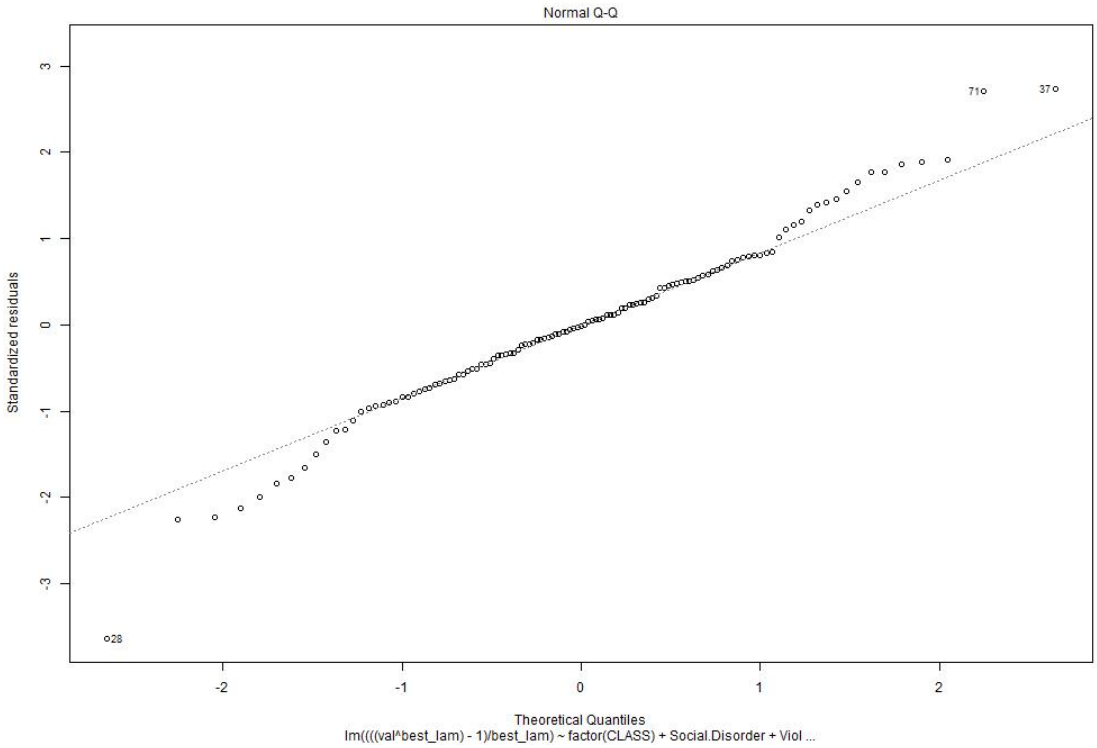


Figure 11: Q-Q plot illustrating normality within the transformed data.

$$H_0 : The\ data\ is\ significantly\ normally\ distributed.$$

$$H_A : The\ data\ is\ \textbf{NOT}\ significantly\ normally\ distributed.$$

it is determined that the Alternate hypothesis, H_A, is, again, favorable over the null, H_0. While the change is improved, the normality of the transformed data remains slightly below the level of significance.

The final assumption to verify is the homoscedasticty. Using the same method as previously described, the Breusch-Pagan test produced a p-value of 0.4267. Again, using the same hypotheses, stated as:

$$H_0 : Heteroscedasticiy\ is\ not\ present\ in\ the\ model,\ the\ data\ is\ homoscedastic,\ \sigma_i^2 = \sigma_j^2 = \sigma_k^2$$

$$H_A : Heteroscedasticiy\ is\ present\ in\ the\ model,\ the\ data\ is\ \textbf{NOT}\ homoscedastic,\ at\ least\ one\ \sigma_i^2 \neq \sigma_j^2$$

it is possible to claim that the data is homoscedastic. Due to the p-value surpassing the α of 0.05, the null hypothesis is to be accepted. With this, the assumption of homoscedasticity holds for the transformed data.

It is clear that this transform method aids in satisfying some assumptions. While the R^2_{adj} for the transformed model is considerably lower than the original best fit model, It is believed to be more accurate for the data as assumptions are improved. While the normality is not definitively changed, the homoscedasticty is. Additionally, it was not expected to find a model with a perfect fit. It is believed that the variety of assessed property types contributed to the data irregularity. To improve the model, conducting additional outlier analyses may be useful. One potential drawback with this is it may be difficult to determine where to stop, potentially removing points which should not be removed. To improve the relevance of the model, only certain property or community types could be analyzed. For example, the majority of homes in Calgary

are within the range of $300 - 500$ thousand dollars. By using only homes, or residential properties, many of the significantly more expensive facilities, such as hospitals, would not interfere, proving a potentially more accurate model at predicting a house price.

Conclusion

This study found that of the available predictor variables, only the number of social disorders, non-domestic violent crimes, persons aged 25 to 65, and community classification were significant when estimating assessed property value. An all possible regression method was used to determine the best estimate model. This model was selected primarily on the AIC value, and secondly on C_p value. It was fount that interaction effects were present between community class and the number of social disorders, and community class and the number of non-domestic violent crimes. No higher order terms were found to be significant. Therfore the best fit model was determined to be:

$$\hat{Value} = 601,057 \times Violence_{Other} + 575,389 \times Social\ \ Disorder - 141 \times Age_{(25-65)} + \phi$$

where:

$$\phi = \begin{cases} -3,993,516 & if\ \ Class = Industrial \\ 30,019,518 - 2,701,952 \times Social\ \ Disorder - 3,383,610 \times Violence_{other} & if\ \ Class = Major\ \ Park \\ -153,792 - 479,214 \times Social\ \ Disorder - 565,410 \times Violence_{Other} & if\ \ Class = Residential \end{cases}$$

This determined best fit model produced an R^2_{adj} value of 0.8494, and an RMSE of $\$742,200$. When verifying the regression assumptions, it was found that the assumptions of normality and homoscedasticity were not satisfied. Performing a Box-Cox transformation on the response variable was completed in attempts to remedy these violations. After transformation, the assumption of homoscedasticty became satisfied, however, the assumption of normality remained un-satisfied. This transformed model was determined to be:

$$\hat{Value}^{-0.73232} = 2.266e-6 \times Violence_{Other} + 6.677e-6 \times Social\ \ Disorder - 5.290e-9 \times Age_{(25-65)} + \phi$$

where:

$$\phi = \begin{cases} 1.365 & if\ \ Class = Industrial \\ 1.365 - 8.067e-6 \times Social\ \ Disorder - 2.416e-5 \times Violence_{other} & if\ \ Class = Major\ \ Park \\ 1.365 - 4.578e-6 \times Social\ \ Disorder - 2.717e-6 \times Violence_{Other} & if\ \ Class = Residential \end{cases}$$

Compared to the best fit model, the transformed model produced a lower R^2_{adj} value of 0.4693. Although the best fit model produced a larger R^2_{adj}, due to the assumption violations, it is not justifiable to state that it is a superior model. Generally, both these models indicate that an increase in crime rates will reduce property value, with the exception of communities classified as industrial. Ultimately, both models are not likely to produce overly meaningful results. Further investigation should be conducted, potentially isolating property type, and conducting a more rigorous outlier detection.

References

Royal Bank of Canada (RBC), 2017, *Housing Trends and Affordability*, RBC Economics | Research, Copyright: Royal Bank of Canada, retrieved from: http://www.rbc.com/economics/economic-reports/pdf/canadian-housing/house-jun2017.pdf

City of Calgary, 2020, Open Government Licence - City of Calgary, accessed on 03-11-2020, available at:https://data.calgary.ca/stories/s/Open-Calgary-Terms-of-Use/u45n-7awa

City of Calgary, 2018a, Community Crime and Disorder Statistics(to be archived), accessed on 24-11-2020, available at: https://data.calgary.ca/Health-and-Safety/Community-Crime-and-Disorder-Statistics-to-be-arch/848s-4m4zhttps://data.calgary.ca/Health-and-Safety/Community-Crime-and-Disorder-Statistics-to-be-arch/848s-4m4z

City of Calgary, 2019, census by community 2019, accessed on 3-11-2020, available at: https://data.calgary.ca/Demographics/Census-by-Community-2019/rkfr-buzb

City of Calgary, 2016, Community services, accessed on 3-11-2020, available at: https://data.calgary.ca/Services-and-Amenities/Community-Services/x34e-bcjz

City of Calgary, 2018b, Property Assessments, accessed on 2-11-2020, available at: https://data.calgary.ca/dataset/Property-Assessments/6zp6-pxei

T. Ngamkham, 2020, DATA 603 Course Notes, *Mutliple Linear Regression Part 1, 2, 3, 4*, Accessed from D2L content boards

Appendix A

R Code

```
data = read.csv('G:/DATA 603/Project/Data/combined.csv')
head(data,4)
```

```
##                           NAME comm n_serv      val       CLASS    SECTOR
## 1                    ABBEYDALE  ABB      1 322768.9 Residential NORTHEAST
## 2                       ACADIA  ACA      2 482854.0 Residential     SOUTH
## 3 ALBERT PARK/RADISSON HEIGHTS  ALB      1 418044.4 Residential      EAST
## 4                     ALTADORE  ALT      1 767906.4 Residential    CENTRE
##   Street.Robbery Theft.OF.Vehicle Theft.FROM.Vehicle Commercial.Break...Enter
## 1              2               12                 12                        2
## 2              3               12                 12                       11
## 3              6               12                 12                       11
## 4              1               11                 12                       12
##   Social.Disorder Assault..Non.domestic. Residential.Break...Enter
## 1              12                     9                       10
## 2              12                    11                       12
## 3              12                    12                       12
## 4              12                     8                       12
##   Physical.Disorder Violence.Other..Non.domestic. Commercial.Robbery age_0
## 1                 9                             4                 5  2006
## 2                12                             8                 5  2389
## 3                12                             7                 6  2090
## 4                11                             8                 0  2144
##   age_25 age_65   pop
## 1   3423    522  5951
## 2   6247   1883 10519
## 3   4274    623  6987
## 4   4308    477  6929
```

```
w=600
h=400
```

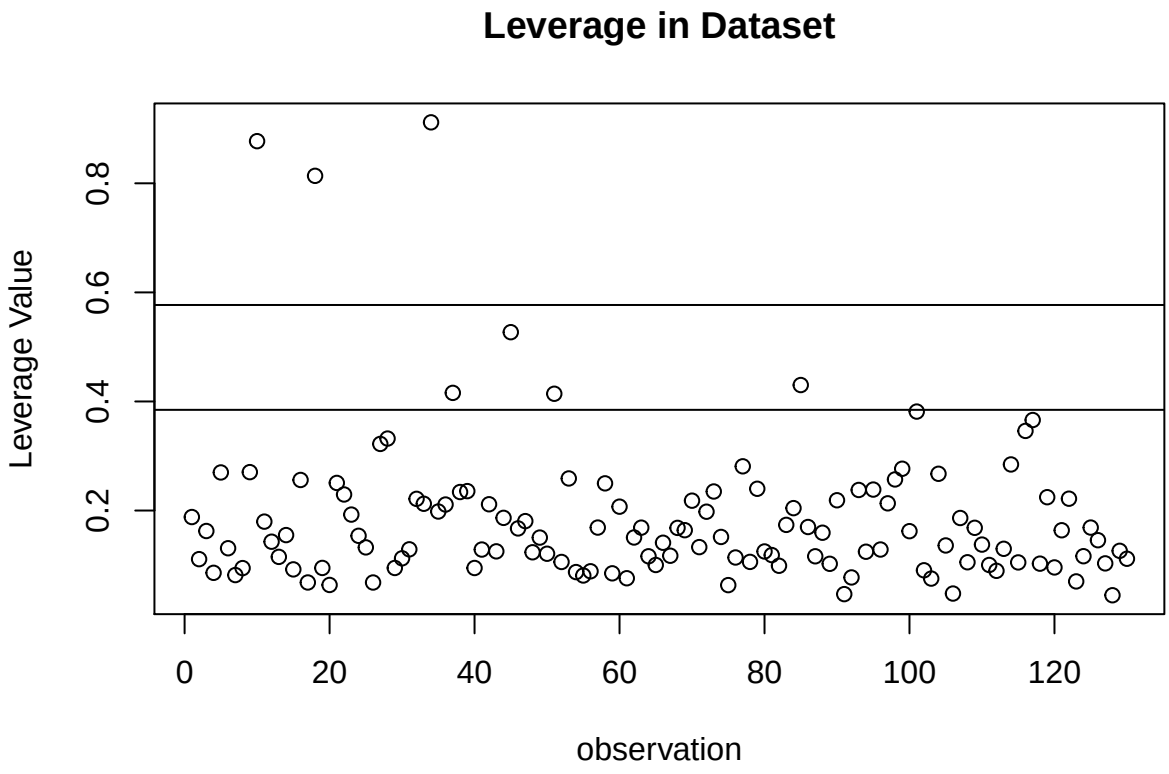
```
model = lm(val~n_serv+factor(CLASS)+factor(SECTOR)+Street.Robbery+
             Theft.OF.Vehicle+Theft.FROM.Vehicle+Commercial.Break...Enter+
             Social.Disorder+Assault..Non.domestic.+Residential.Break...Enter+
             Physical.Disorder+Violence.Other..Non.domestic.+Commercial.Robbery+
             age_0+age_25+age_65+pop,data)
```

Remove outliers

```
lev=hatvalues(model)
p = length(coef(model))
n = nrow(data)
outlier = lev[lev>(3*p/n)]
print(outlier)
```

```
##        10        18        34
## 0.8773014 0.8137214 0.9117880
```

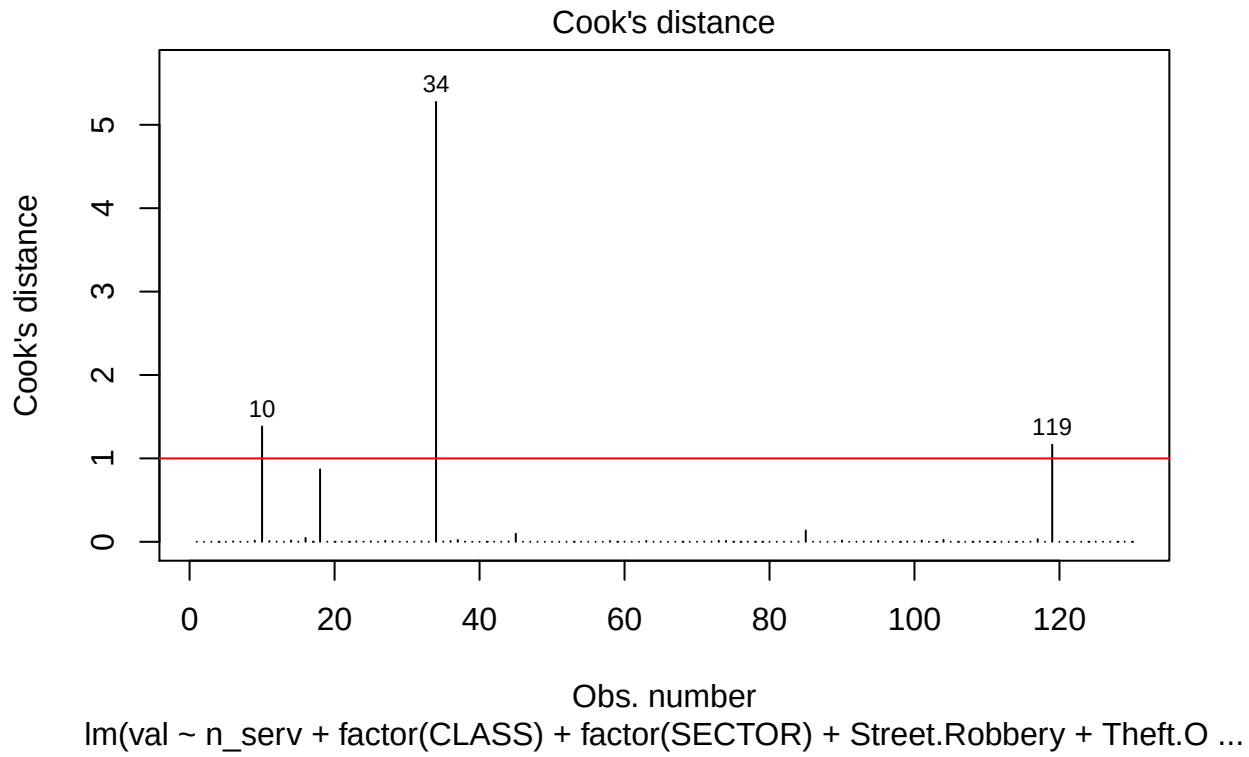
```
plot(rownames(data),lev, main = "Leverage in Dataset", xlab="observation",
     ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
abline(h = 3 *p/n, lty = 1)
```



```
jpeg('Cooks.jpg',width =w,height=h)
plot(model, which=4)
abline(h=1,col='red')
dev.off()
```

```
## pdf
##   2
```

```
plot(model, which=4)
abline(h=1,col='red')
```




```
out_data = data[cooks.distance(model)<1,]

model = lm(val~n_serv+factor(CLASS)+factor(SECTOR)+Street.Robbery+Theft.OF.Vehicle+
             Theft.FROM.Vehicle+Commercial.Break...Enter+Social.Disorder+
             Assault..Non.domestic.+Residential.Break...Enter+
             Physical.Disorder+Violence.Other..Non.domestic.+Commercial.Robbery+
             age_0+age_25+age_65+pop,out_data)
summary(model)
```

```
##
## Call:
## lm(formula = val ~ n_serv + factor(CLASS) + factor(SECTOR) +
##     Street.Robbery + Theft.OF.Vehicle + Theft.FROM.Vehicle +
```

```
##     Commercial.Break...Enter + Social.Disorder + Assault..Non.domestic. +
##     Residential.Break...Enter + Physical.Disorder + Violence.Other..Non.domestic. +
##     Commercial.Robbery + age_0 + age_25 + age_65 + pop, data = out_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5610280  -375110   -42938   239872  5748060
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.173e+07  2.774e+06   7.831 4.51e-12 ***
## n_serv                       -1.310e+05  1.817e+05  -0.721  0.47245
## factor(CLASS)Major Park       1.982e+06  1.144e+06   1.733  0.08614 .
## factor(CLASS)Residential     -2.640e+06  8.089e+05  -3.264  0.00149 **
## factor(SECTOR)EAST           -7.567e+05  5.687e+05  -1.331  0.18626
## factor(SECTOR)NORTH          -2.882e+05  5.903e+05  -0.488  0.62644
## factor(SECTOR)NORTHEAST      -5.893e+05  5.520e+05  -1.068  0.28821
## factor(SECTOR)NORTHWEST      -2.088e+05  4.956e+05  -0.421  0.67438
## factor(SECTOR)SOUTH          -2.155e+05  3.746e+05  -0.575  0.56634
## factor(SECTOR)SOUTHEAST       2.595e+05  6.342e+05   0.409  0.68328
## factor(SECTOR)WEST            2.816e+05  4.286e+05   0.657  0.51270
## Street.Robbery               -3.385e+04  8.011e+04  -0.423  0.67349
## Theft.OF.Vehicle             -1.893e+04  7.928e+04  -0.239  0.81177
## Theft.FROM.Vehicle            2.703e+04  1.039e+05   0.260  0.79532
## Commercial.Break...Enter     -5.884e+04  4.890e+04  -1.203  0.23162
## Social.Disorder              -1.570e+06  2.443e+05  -6.426 4.13e-09 ***
## Assault..Non.domestic.        1.005e+05  5.665e+04   1.773  0.07912 .
## Residential.Break...Enter     5.533e+03  5.774e+04   0.096  0.92385
## Physical.Disorder             6.680e+04  8.103e+04   0.824  0.41166
## Violence.Other..Non.domestic. 6.149e+04  6.235e+04   0.986  0.32633
## Commercial.Robbery            4.301e+04  7.976e+04   0.539  0.59090
## age_0                         3.946e+02  3.406e+02   1.159  0.24923
## age_25                       -4.034e+02  2.248e+02  -1.795  0.07562 .
## age_65                        2.689e+01  3.004e+02   0.090  0.92883
## pop                                  NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1158000 on 103 degrees of freedom
## Multiple R-squared:  0.7001, Adjusted R-squared:  0.6331
## F-statistic: 10.45 on 23 and 103 DF,  p-value: < 2.2e-16
```

check multicollinearity:

```
imcdiag(model,method = 'VIF')
```

```
## Warning in summary.lm(lm(x[, i] ~ x[, -i])): essentially perfect fit: summary
## may be unreliable

## Warning in summary.lm(lm(x[, i] ~ x[, -i])): essentially perfect fit: summary
## may be unreliable
```

```
## Warning in summary.lm(lm(x[, i] ~ x[, -i])): essentially perfect fit: summary
## may be unreliable

## Warning in summary.lm(lm(x[, i] ~ x[, -i])): essentially perfect fit: summary
## may be unreliable

## Warning in summary.lm(lm(x[, i] ~ x[, -i])): essentially perfect fit: summary
## may be unreliable

## Warning in summary.lm(lm(x[, i] ~ x[, -i])): essentially perfect fit: summary
## may be unreliable


##
## Call:
## imcdiag(mod = model, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                                VIF detection
## n_serv                      1.2845         0
## factor(CLASS)Major Park     2.8555         0
## factor(CLASS)Residential    4.0771         0
## factor(SECTOR)EAST          1.5940         0
## factor(SECTOR)NORTH         1.4844         0
## factor(SECTOR)NORTHEAST     3.0039         0
## factor(SECTOR)NORTHWEST     2.1361         0
## factor(SECTOR)SOUTH         2.1616         0
## factor(SECTOR)SOUTHEAST     1.7135         0
## factor(SECTOR)WEST          1.8111         0
## Street.Robbery              3.5696         0
## Theft.OF.Vehicle            5.9567         0
## Theft.FROM.Vehicle          3.6775         0
## Commercial.Break...Enter    2.9597         0
## Social.Disorder             2.0840         0
## Assault..Non.domestic.      4.3002         0
## Residential.Break...Enter   4.0481         0
## Physical.Disorder           4.8027         0
## Violence.Other..Non.domestic. 3.0869       0
## Commercial.Robbery          2.3505         0
## age_0                          Inf         1
## age_25                         Inf         1
## age_65                         Inf         1
## pop                            Inf         1
##
## Multicollinearity may be due to age_0 age_25 age_65 pop regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ====================================
```

```
model = lm(val~n_serv+factor(CLASS)+factor(SECTOR)+Street.Robbery+Theft.OF.Vehicle+
             Theft.FROM.Vehicle+Commercial.Break...Enter+Social.Disorder+
             Assault..Non.domestic.+Residential.Break...Enter+Physical.Disorder+
             Violence.Other..Non.domestic.+Commercial.Robbery+age_0+age_25+
             age_65,out_data)
imcdiag(model,method = 'VIF')
```

```
##
## Call:
## imcdiag(mod = model, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                                 VIF detection
## n_serv                       1.2845         0
## factor(CLASS)Major Park      2.8555         0
## factor(CLASS)Residential     4.0771         0
## factor(SECTOR)EAST           1.5940         0
## factor(SECTOR)NORTH          1.4844         0
## factor(SECTOR)NORTHEAST      3.0039         0
## factor(SECTOR)NORTHWEST      2.1361         0
## factor(SECTOR)SOUTH          2.1616         0
## factor(SECTOR)SOUTHEAST      1.7135         0
## factor(SECTOR)WEST           1.8111         0
## Street.Robbery               3.5696         0
## Theft.OF.Vehicle             5.9567         0
## Theft.FROM.Vehicle           3.6775         0
## Commercial.Break...Enter     2.9597         0
## Social.Disorder              2.0840         0
## Assault..Non.domestic.       4.3002         0
## Residential.Break...Enter    4.0481         0
## Physical.Disorder            4.8027         0
## Violence.Other..Non.domestic. 3.0869         0
## Commercial.Robbery           2.3505         0
## age_0                       19.3490         1
## age_25                      23.5518         1
## age_65                       3.2541         0
##
## Multicollinearity may be due to age_0 age_25 regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```
model = lm(val~n_serv+factor(CLASS)+factor(SECTOR)+Street.Robbery+Theft.OF.Vehicle+
             Theft.FROM.Vehicle+Commercial.Break...Enter+Social.Disorder+
             Assault..Non.domestic.+Residential.Break...Enter+Physical.Disorder+
             Violence.Other..Non.domestic.+Commercial.Robbery+age_25+
             age_65,out_data)
imcdiag(model,method = 'VIF')
```

```
##
## Call:
## imcdiag(mod = model, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                                 VIF detection
## n_serv                       1.2765         0
## factor(CLASS)Major Park      2.8481         0
## factor(CLASS)Residential     4.0614         0
## factor(SECTOR)EAST           1.5529         0
## factor(SECTOR)NORTH          1.4812         0
## factor(SECTOR)NORTHEAST      2.8102         0
## factor(SECTOR)NORTHWEST      2.0225         0
## factor(SECTOR)SOUTH          2.1503         0
## factor(SECTOR)SOUTHEAST      1.7059         0
## factor(SECTOR)WEST           1.7857         0
## Street.Robbery               3.5681         0
## Theft.OF.Vehicle             5.9122         0
## Theft.FROM.Vehicle           3.6618         0
## Commercial.Break...Enter     2.7593         0
## Social.Disorder              2.0769         0
## Assault..Non.domestic.       4.2428         0
## Residential.Break...Enter    4.0313         0
## Physical.Disorder            4.7714         0
## Violence.Other..Non.domestic. 3.0060        0
## Commercial.Robbery           2.3375         0
## age_25                       4.9488         0
## age_65                       3.2417         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ==================================
```

```
summary(model)
```

```
##
## Call:
## lm(formula = val ~ n_serv + factor(CLASS) + factor(SECTOR) +
##     Street.Robbery + Theft.OF.Vehicle + Theft.FROM.Vehicle +
##     Commercial.Break...Enter + Social.Disorder + Assault..Non.domestic. +
##     Residential.Break...Enter + Physical.Disorder + Violence.Other..Non.domestic. +
##     Commercial.Robbery + age_25 + age_65, data = out_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5738079  -327266   -30308   188839  5649239
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                    2.163e+07  2.778e+06   7.788 5.33e-12 ***
## n_serv                        -1.477e+05  1.814e+05  -0.814  0.41744
## factor(CLASS)Major Park        1.914e+06  1.144e+06   1.673  0.09729 .
## factor(CLASS)Residential      -2.699e+06  8.087e+05  -3.337  0.00118 **
## factor(SECTOR)EAST            -6.509e+05  5.622e+05  -1.158  0.24961
## factor(SECTOR)NORTH           -2.564e+05  5.906e+05  -0.434  0.66508
## factor(SECTOR)NORTHEAST       -4.269e+05  5.348e+05  -0.798  0.42657
## factor(SECTOR)NORTHWEST       -7.635e+04  4.831e+05  -0.158  0.87471
## factor(SECTOR)SOUTH           -1.840e+05  3.742e+05  -0.492  0.62386
## factor(SECTOR)SOUTHEAST        2.107e+05  6.339e+05   0.332  0.74029
## factor(SECTOR)WEST             3.404e+05  4.263e+05   0.799  0.42637
## Street.Robbery                -3.578e+04  8.023e+04  -0.446  0.65653
## Theft.OF.Vehicle              -2.687e+04  7.911e+04  -0.340  0.73481
## Theft.FROM.Vehicle             3.491e+04  1.039e+05   0.336  0.73753
## Commercial.Break...Enter      -7.359e+04  4.730e+04  -1.556  0.12276
## Social.Disorder               -1.553e+06  2.443e+05  -6.359 5.51e-09 ***
## Assault..Non.domestic.         9.288e+04  5.637e+04   1.648  0.10240
## Residential.Break...Enter      1.230e+03  5.772e+04   0.021  0.98304
## Physical.Disorder              5.922e+04  8.090e+04   0.732  0.46580
## Violence.Other..Non.domestic.  7.318e+04  6.162e+04   1.188  0.23770
## Commercial.Robbery             4.988e+04  7.967e+04   0.626  0.53265
## age_25                        -1.719e+02  1.032e+02  -1.666  0.09872 .
## age_65                         5.397e+00  3.003e+02   0.018  0.98569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1160000 on 104 degrees of freedom
## Multiple R-squared:  0.6962, Adjusted R-squared:  0.6319
## F-statistic: 10.83 on 22 and 104 DF,  p-value: < 2.2e-16
```

```
red_model = lm(val~factor(CLASS)+Social.Disorder+Assault..Non.domestic.+age_25,out_data)
summary(red_model)
```

```
##
## Call:
## lm(formula = val ~ factor(CLASS) + Social.Disorder + Assault..Non.domestic. +
##     age_25, data = out_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6058207  -342528   -74656   176792  5840011
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.036e+07  2.404e+06   8.469 6.89e-14 ***
## factor(CLASS)Major Park    2.219e+06  8.803e+05   2.520   0.0130 *
## factor(CLASS)Residential -2.288e+06  5.119e+05  -4.471 1.77e-05 ***
## Social.Disorder          -1.457e+06  2.064e+05  -7.059 1.14e-10 ***
## Assault..Non.domestic.    6.735e+04  2.845e+04   2.367   0.0195 *
## age_25                   -1.227e+02  5.039e+01  -2.435   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1122000 on 121 degrees of freedom
```

```
## Multiple R-squared:  0.6695, Adjusted R-squared:  0.6558
## F-statistic: 49.02 on 5 and 121 DF,  p-value: < 2.2e-16
```

```
anova(red_model, model)
```

```
## Analysis of Variance Table
##
## Model 1: val ~ factor(CLASS) + Social.Disorder + Assault..Non.domestic. +
##     age_25
## Model 2: val ~ n_serv + factor(CLASS) + factor(SECTOR) + Street.Robbery +
##     Theft.OF.Vehicle + Theft.FROM.Vehicle + Commercial.Break...Enter +
##     Social.Disorder + Assault..Non.domestic. + Residential.Break...Enter +
##     Physical.Disorder + Violence.Other..Non.domestic. + Commercial.Robbery +
##     age_25 + age_65
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1    121 1.5233e+14
## 2    104 1.4003e+14 17 1.2302e+13 0.5375 0.9277
```

```
jpeg('age_pairs.jpg',width =w, height = h)
pairs(~age_0+age_25+age_65+pop,out_data)
dev.off()
```

```
## pdf
##   2
```

```
step_model = ols_step_both_p(model,pent=0.05, prem = 0.1)
summary(step_model$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6120628  -309753   -78868   176373  5797232
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.023e+07  2.400e+06   8.429 8.56e-14 ***
## Social.Disorder              -1.436e+06  2.054e+05  -6.991 1.61e-10 ***
## factor(CLASS)Major Park       2.155e+06  8.785e+05   2.453   0.0156 *
## factor(CLASS)Residential     -2.231e+06  5.113e+05  -4.364 2.71e-05 ***
## age_25                       -1.640e+02  5.621e+01  -2.918   0.0042 **
## Violence.Other..Non.domestic.  9.720e+04  4.097e+04   2.373   0.0192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1122000 on 121 degrees of freedom
## Multiple R-squared:  0.6696, Adjusted R-squared:  0.6559
## F-statistic: 49.04 on 5 and 121 DF,  p-value: < 2.2e-16
```

```r
step_forw = ols_step_forward_p(model, pent=0.1)
summary(step_forw$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6120628  -309753   -78868   176373  5797232
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.023e+07  2.400e+06   8.429 8.56e-14 ***
## Social.Disorder              -1.436e+06  2.054e+05  -6.991 1.61e-10 ***
## factor(CLASS)Major Park       2.155e+06  8.785e+05   2.453   0.0156 *
## factor(CLASS)Residential     -2.231e+06  5.113e+05  -4.364 2.71e-05 ***
## age_25                       -1.640e+02  5.621e+01  -2.918   0.0042 **
## Violence.Other..Non.domestic.  9.720e+04  4.097e+04   2.373   0.0192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1122000 on 121 degrees of freedom
## Multiple R-squared:  0.6696, Adjusted R-squared:  0.6559
## F-statistic: 49.04 on 5 and 121 DF,  p-value: < 2.2e-16
```

```r
step_back = ols_step_backward_p(model, prem=0.1)
summary(step_back$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6058207  -342528   -74656   176792  5840011
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.036e+07  2.404e+06   8.469 6.89e-14 ***
## factor(CLASS)Major Park   2.219e+06  8.803e+05   2.520   0.0130 *
## factor(CLASS)Residential -2.288e+06  5.119e+05  -4.471 1.77e-05 ***
## Social.Disorder          -1.457e+06  2.064e+05  -7.059 1.14e-10 ***
## Assault..Non.domestic.    6.735e+04  2.845e+04   2.367   0.0195 *
## age_25                   -1.227e+02  5.039e+01  -2.435   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1122000 on 121 degrees of freedom
## Multiple R-squared:  0.6695, Adjusted R-squared:  0.6558
## F-statistic: 49.02 on 5 and 121 DF,  p-value: < 2.2e-16
```

```
ks1 = ols_step_best_subset(model, details= T)
ks1
```

```
##
## ---------------------------------------------------------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------------------------------------------------------
##      1         factor(CLASS)
##      2         factor(CLASS) Social.Disorder
##      3         factor(CLASS) Social.Disorder age_25
##      4         factor(CLASS) Social.Disorder Violence.Other..Non.domestic. age_25
##      5         factor(CLASS) factor(SECTOR) Social.Disorder Assault..Non.domestic. age_25
##      6         factor(CLASS) factor(SECTOR) Commercial.Break...Enter Social.Disorder Assault..Non.dor
##      7         factor(CLASS) factor(SECTOR) Commercial.Break...Enter Social.Disorder Assault..Non.dor
##      8         factor(CLASS) factor(SECTOR) Commercial.Break...Enter Social.Disorder Assault..Non.dor
##      9         n_serv factor(CLASS) factor(SECTOR) Commercial.Break...Enter Social.Disorder Assault.
##     10         n_serv factor(CLASS) factor(SECTOR) Commercial.Break...Enter Social.Disorder Assault.
##     11         n_serv factor(CLASS) factor(SECTOR) Street.Robbery Commercial.Break...Enter Social.Dis
##     12         n_serv factor(CLASS) factor(SECTOR) Street.Robbery Theft.OF.Vehicle Commercial.Break.
##     13         n_serv factor(CLASS) factor(SECTOR) Street.Robbery Theft.OF.Vehicle Theft.FROM.Vehicle
##     14         n_serv factor(CLASS) factor(SECTOR) Street.Robbery Theft.OF.Vehicle Theft.FROM.Vehicle
##     15         n_serv factor(CLASS) factor(SECTOR) Street.Robbery Theft.OF.Vehicle Theft.FROM.Vehicle
## ---------------------------------------------------------------------------------------------------------
##
##                                                    Subsets Regression Summary
## ---------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)         AIC         SBIC          SBC           MS
## ---------------------------------------------------------------------------------------------------------
##   1        0.5124      0.5045      0.1241    43.9167    3950.0339    3586.5526    3961.4107    2.2833
##   2        0.6428      0.6341     -0.1588     1.2689    3912.5024    3550.3212    3926.7233    1.6861
##   3        0.6542      0.6428     -0.1803    -0.6219    3910.3953    3548.5570    3927.4605    1.6458
##   4        0.6696      0.6559     -0.1323    -3.8849    3906.6197    3545.4066    3926.5290    1.5857
##   5        0.6804      0.6468     -0.1591    -5.5928    3916.3869    3543.8544    3956.2055    1.5465
##   6        0.6869      0.6509     -0.1738    -5.8225    3915.7719    3543.8824    3958.4347    1.5277
##   7        0.6898      0.6510     -0.1724    -4.7965    3916.6124    3545.2584    3962.1194    1.5266
##   8        0.6922      0.6506     -0.1913    -3.6379    3917.6023    3546.8042    3965.9535    1.5275
##   9        0.6942      0.6498     -0.1892    -2.3302    3918.7650    3548.5304    3969.9604    1.5305
##  10        0.6950      0.6474      -0.197    -0.5957    3920.4425    3550.6998    3974.4821    1.5399
##  11        0.6956      0.6449     -0.2079     1.1878    3922.1789    3552.9275    3979.0626    1.5501
##  12        0.6959      0.6418     -0.2397     3.1149    3924.0900    3555.2980    3983.8180    1.5628
##  13        0.6962      0.6389     -0.2961     5.0008    3925.9508    3557.6346    3988.5229      1.57
##  14        0.6962      0.6354     -0.3135     7.0003    3927.9502    3560.0765    3993.3665    1.5891
##  15        0.6962      0.6319     -0.3175     9.0000    3929.9498    3562.5185    3998.2103    1.6037
## ---------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
jpeg('cp_AIC.jpg',heigh = h,width=w)
par(mfrow=c(2,2))
plot(ks1$cp,type = "o",pch=10, xlab="Number of Variables",ylab= "Cp")
plot(ks1$rsq,type = "o",pch=10, xlab="Number of Variables",ylab= "R^2")
plot(ks1$aic,type = "o",pch=10, xlab="Number of Variables",ylab= "AIC")
plot(ks1$adjr,type = "o",pch=10, xlab="Number of Variables",ylab= "Adjusted R^2")
dev.off()
```

```
## pdf
##   2
```

Best Model

```
best_first_model = lm(val~factor(CLASS)+Social.Disorder+Violence.Other..Non.domestic.+
                        age_25,out_data)
summary(best_first_model)
```

```
##
## Call:
## lm(formula = val ~ factor(CLASS) + Social.Disorder + Violence.Other..Non.domestic. +
##     age_25, data = out_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6120628  -309753   -78868   176373  5797232
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.023e+07  2.400e+06   8.429 8.56e-14 ***
## factor(CLASS)Major Park       2.155e+06  8.785e+05   2.453   0.0156 *
## factor(CLASS)Residential     -2.231e+06  5.113e+05  -4.364 2.71e-05 ***
## Social.Disorder              -1.436e+06  2.054e+05  -6.991 1.61e-10 ***
## Violence.Other..Non.domestic.  9.720e+04  4.097e+04   2.373   0.0192 *
## age_25                       -1.640e+02  5.621e+01  -2.918   0.0042 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1122000 on 121 degrees of freedom
## Multiple R-squared:  0.6696, Adjusted R-squared:  0.6559
## F-statistic: 49.04 on 5 and 121 DF,  p-value: < 2.2e-16
```

anova test

```
anova(best_first_model,model)
```

```
## Analysis of Variance Table
##
## Model 1: val ~ factor(CLASS) + Social.Disorder + Violence.Other..Non.domestic. +
##     age_25
```

```
## Model 2: val ~ n_serv + factor(CLASS) + factor(SECTOR) + Street.Robbery +
##     Theft.OF.Vehicle + Theft.FROM.Vehicle + Commercial.Break...Enter +
##     Social.Disorder + Assault..Non.domestic. + Residential.Break...Enter +
##     Physical.Disorder + Violence.Other..Non.domestic. + Commercial.Robbery +
##     age_25 + age_65
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1    121 1.5230e+14
## 2    104 1.4003e+14 17 1.2273e+13 0.5362 0.9285
```

interaction

```
int_model = lm(val~(factor(CLASS)+Social.Disorder+Violence.Other..Non.domestic.+
                      age_25)^2,out_data)
summary(int_model)
```

```
##
## Call:
## lm(formula = val ~ (factor(CLASS) + Social.Disorder + Violence.Other..Non.domestic. +
##     age_25)^2, data = out_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1875600  -245295  -126167    69531  4347824
##
## Coefficients: (3 not defined because of singularities)
##                                                      Estimate Std. Error
## (Intercept)                                        -3.994e+06  3.260e+06
## factor(CLASS)Major Park                             3.401e+07  3.758e+06
## factor(CLASS)Residential                            9.472e+06  9.813e+06
## Social.Disorder                                     5.754e+05  2.883e+05
## Violence.Other..Non.domestic.                      -9.441e+06  1.267e+07
## age_25                                             -1.144e+02  8.295e+01
## factor(CLASS)Major Park:Social.Disorder            -2.702e+06  3.573e+05
## factor(CLASS)Residential:Social.Disorder           -9.557e+05  8.290e+05
## factor(CLASS)Major Park:Violence.Other..Non.domestic. -3.384e+06  1.180e+06
## factor(CLASS)Residential:Violence.Other..Non.domestic. -5.466e+05  1.222e+05
## factor(CLASS)Major Park:age_25                             NA         NA
## factor(CLASS)Residential:age_25                            NA         NA
## Social.Disorder:Violence.Other..Non.domestic.       8.369e+05  1.056e+06
## Social.Disorder:age_25                                     NA         NA
## Violence.Other..Non.domestic.:age_25                -5.132e+00  1.369e+01
##                                                    t value Pr(>|t|)
## (Intercept)                                         -1.225  0.22310
## factor(CLASS)Major Park                              9.050 4.24e-15 ***
## factor(CLASS)Residential                             0.965  0.33645
## Social.Disorder                                      1.996  0.04830 *
## Violence.Other..Non.domestic.                       -0.745  0.45779
## age_25                                              -1.379  0.17042
## factor(CLASS)Major Park:Social.Disorder             -7.563 1.05e-11 ***
## factor(CLASS)Residential:Social.Disorder            -1.153  0.25141
## factor(CLASS)Major Park:Violence.Other..Non.domestic. -2.867  0.00493 **
## factor(CLASS)Residential:Violence.Other..Non.domestic. -4.472 1.83e-05 ***
```

```
## factor(CLASS)Major Park:age_25                                  NA       NA
## factor(CLASS)Residential:age_25                                 NA       NA
## Social.Disorder:Violence.Other..Non.domestic.                0.792  0.42972
## Social.Disorder:age_25                                          NA       NA
## Violence.Other..Non.domestic.:age_25                         -0.375  0.70836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 746200 on 115 degrees of freedom
## Multiple R-squared:  0.8611, Adjusted R-squared:  0.8478
## F-statistic:  64.8 on 11 and 115 DF,  p-value: < 2.2e-16
```

```
best_int = lm(val~factor(CLASS)+Social.Disorder+Violence.Other..Non.domestic.+
                age_25+factor(CLASS)*Social.Disorder+
                factor(CLASS)*Violence.Other..Non.domestic.,out_data)
summary(best_int)
```

```
##
## Call:
## lm(formula = val ~ factor(CLASS) + Social.Disorder + Violence.Other..Non.domestic. +
##     age_25 + factor(CLASS) * Social.Disorder + factor(CLASS) *
##     Violence.Other..Non.domestic., data = out_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1875600  -240626  -123285    85022  4374894
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                      -3993515.6  3242850.0
## factor(CLASS)Major Park                          34013034.5  3738410.5
## factor(CLASS)Residential                          3839724.5  7243911.3
## Social.Disorder                                    575389.3   286739.7
## Violence.Other..Non.domestic.                      601057.1   108035.0
## age_25                                               -141.4       37.7
## factor(CLASS)Major Park:Social.Disorder          -2701952.0   355375.1
## factor(CLASS)Residential:Social.Disorder          -479214.9   614060.7
## factor(CLASS)Major Park:Violence.Other..Non.domestic. -3383610.3  1173838.9
## factor(CLASS)Residential:Violence.Other..Non.domestic.  -565410.1   111662.6
##                                                  t value Pr(>|t|)
## (Intercept)                                       -1.231 0.220612
## factor(CLASS)Major Park                            9.098 2.88e-15 ***
## factor(CLASS)Residential                           0.530 0.597073
## Social.Disorder                                    2.007 0.047092 *
## Violence.Other..Non.domestic.                      5.564 1.70e-07 ***
## age_25                                            -3.751 0.000276 ***
## factor(CLASS)Major Park:Social.Disorder           -7.603 7.93e-12 ***
## factor(CLASS)Residential:Social.Disorder          -0.780 0.436730
## factor(CLASS)Major Park:Violence.Other..Non.domestic.  -2.883 0.004696 **
## factor(CLASS)Residential:Violence.Other..Non.domestic.  -5.064 1.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 742200 on 117 degrees of freedom
```

```
## Multiple R-squared:  0.8602, Adjusted R-squared:  0.8494
## F-statistic: 79.96 on 9 and 117 DF,  p-value: < 2.2e-16
```

```
anova(int_model, best_int)
```

```
## Analysis of Variance Table
##
## Model 1: val ~ (factor(CLASS) + Social.Disorder + Violence.Other..Non.domestic. +
##     age_25)^2
## Model 2: val ~ factor(CLASS) + Social.Disorder + Violence.Other..Non.domestic. +
##     age_25 + factor(CLASS) * Social.Disorder + factor(CLASS) *
##     Violence.Other..Non.domestic.
##   Res.Df        RSS Df   Sum of Sq     F Pr(>F)
## 1    115 6.4033e+13
## 2    117 6.4455e+13 -2 -4.2212e+11 0.379 0.6854
```

Higher Order

```
jpeg('ho_pairs.jpg',height =h,width=w)
pairs(val~factor(CLASS)+Social.Disorder+Violence.Other..Non.domestic.+
        age_25,out_data,panel=panel.smooth)
dev.off()
```

```
## pdf
##   2
```

```
# nothing significant, no higher orders

higher_order = lm(val~factor(CLASS)+Social.Disorder+Violence.Other..Non.domestic.+
                    age_25+factor(CLASS)*Social.Disorder+
                    factor(CLASS)*Violence.Other..Non.domestic.+
                    I(Violence.Other..Non.domestic.)^2,out_data)
summary(higher_order)
```

```
##
## Call:
## lm(formula = val ~ factor(CLASS) + Social.Disorder + Violence.Other..Non.domestic. +
##     age_25 + factor(CLASS) * Social.Disorder + factor(CLASS) *
##     Violence.Other..Non.domestic. + I(Violence.Other..Non.domestic.)^2,
##     data = out_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1875600  -240626  -123285    85022  4374894
##
## Coefficients: (1 not defined because of singularities)
##                                                  Estimate Std. Error
## (Intercept)                                   -3993515.6  3242850.0
## factor(CLASS)Major Park                       34013034.5  3738410.5
## factor(CLASS)Residential                       3839724.5  7243911.3
```

```
## Social.Disorder                                          575389.3   286739.7
## Violence.Other..Non.domestic.                            601057.1   108035.0
## age_25                                                     -141.4       37.7
## I(Violence.Other..Non.domestic.)                               NA         NA
## factor(CLASS)Major Park:Social.Disorder                -2701952.0   355375.1
## factor(CLASS)Residential:Social.Disorder                 -479214.9   614060.7
## factor(CLASS)Major Park:Violence.Other..Non.domestic.  -3383610.3  1173838.9
## factor(CLASS)Residential:Violence.Other..Non.domestic.  -565410.1   111662.6
##                                                        t value Pr(>|t|)
## (Intercept)                                             -1.231 0.220612
## factor(CLASS)Major Park                                  9.098 2.88e-15 ***
## factor(CLASS)Residential                                 0.530 0.597073
## Social.Disorder                                          2.007 0.047092 *
## Violence.Other..Non.domestic.                            5.564 1.70e-07 ***
## age_25                                                  -3.751 0.000276 ***
## I(Violence.Other..Non.domestic.)                            NA       NA
## factor(CLASS)Major Park:Social.Disorder                 -7.603 7.93e-12 ***
## factor(CLASS)Residential:Social.Disorder                -0.780 0.436730
## factor(CLASS)Major Park:Violence.Other..Non.domestic.   -2.883 0.004696 **
## factor(CLASS)Residential:Violence.Other..Non.domestic.  -5.064 1.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 742200 on 117 degrees of freedom
## Multiple R-squared:  0.8602, Adjusted R-squared:  0.8494
## F-statistic: 79.96 on 9 and 117 DF,  p-value: < 2.2e-16
```

```
anova(best_int,higher_order)
```

```
## Analysis of Variance Table
##
## Model 1: val ~ factor(CLASS) + Social.Disorder + Violence.Other..Non.domestic. +
##     age_25 + factor(CLASS) * Social.Disorder + factor(CLASS) *
##     Violence.Other..Non.domestic.
## Model 2: val ~ factor(CLASS) + Social.Disorder + Violence.Other..Non.domestic. +
##     age_25 + factor(CLASS) * Social.Disorder + factor(CLASS) *
##     Violence.Other..Non.domestic. + I(Violence.Other..Non.domestic.)^2
##   Res.Df        RSS Df Sum of Sq F Pr(>F)
## 1    117 6.4455e+13
## 2    117 6.4455e+13  0         0
```

Check assumptions

```
# linearity:
jpeg('resid_plot.jpg',height =h, width =w)
plot(best_int, which = 1)
dev.off()
```

```
## pdf
##   2
```

```
#Normality
jpeg('resid_hist.jpg',height =h, width =w)
ggplot(out_data,aes(residuals(best_int)))+geom_histogram(col='red',fill='blue')+
  ggtitle('Distribution of Residuals of the Best Fit Interaction Model')+labs(x='Residuals',y='Count')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
dev.off()
```

```
## pdf
##   2
```

```
jpeg('qq_plot.jpg',height =h, width =w)
plot(best_int, which=2)
```

```
## Warning: not plotting observations with leverage one:
##   17, 43, 49, 83
```

```
dev.off()
```

```
## pdf
##   2
```

```
shapiro.test(residuals(best_int))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(best_int)
## W = 0.52773, p-value < 2.2e-16
```

```
#heteroscedasticity
bptest(best_int)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  best_int
## BP = 21.158, df = 9, p-value = 0.01197
```

Transform

```
jpeg('box_cox.jpg',height = h, width =w)
bc = boxcox(best_int,lambda = seq(-1.5,0.5))
dev.off()
```

```
## pdf
##   2
```

```r
best_lam = bc$x[which(bc$y==max(bc$y))]
best_lam
```

```
## [1] -0.7323232
```

```r
#best_lam = -0.73232

bc_best_model = lm((((val^best_lam)-1)/best_lam)~factor(CLASS)+Social.Disorder+
                     Violence.Other..Non.domestic.+age_25+factor(CLASS)*Social.Disorder+
                     factor(CLASS)*Violence.Other..Non.domestic.,out_data)
summary(bc_best_model)
```

```
##
## Call:
## lm(formula = (((val^best_lam) - 1)/best_lam) ~ factor(CLASS) +
##     Social.Disorder + Violence.Other..Non.domestic. + age_25 +
##     factor(CLASS) * Social.Disorder + factor(CLASS) * Violence.Other..Non.domestic.,
##     data = out_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -7.502e-05 -1.147e-05  0.000e+00  1.086e-05  5.573e-05
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       1.365e+00  9.151e-05
## factor(CLASS)Major Park                           1.103e-04  1.055e-04
## factor(CLASS)Residential                          1.561e-05  2.044e-04
## Social.Disorder                                   6.677e-06  8.091e-06
## Violence.Other..Non.domestic.                     2.266e-06  3.049e-06
## age_25                                           -5.290e-09  1.064e-09
## factor(CLASS)Major Park:Social.Disorder          -8.067e-06  1.003e-05
## factor(CLASS)Residential:Social.Disorder         -4.578e-06  1.733e-05
## factor(CLASS)Major Park:Violence.Other..Non.domestic.  -2.416e-05  3.312e-05
## factor(CLASS)Residential:Violence.Other..Non.domestic. -2.717e-06  3.151e-06
##                                                    t value Pr(>|t|)
## (Intercept)                                       14921.086  < 2e-16 ***
## factor(CLASS)Major Park                               1.045    0.298
## factor(CLASS)Residential                              0.076    0.939
## Social.Disorder                                       0.825    0.411
## Violence.Other..Non.domestic.                         0.743    0.459
## age_25                                               -4.972 2.29e-06 ***
## factor(CLASS)Major Park:Social.Disorder              -0.804    0.423
## factor(CLASS)Residential:Social.Disorder             -0.264    0.792
## factor(CLASS)Major Park:Violence.Other..Non.domestic.    -0.729    0.467
## factor(CLASS)Residential:Violence.Other..Non.domestic.   -0.862    0.390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.094e-05 on 117 degrees of freedom
## Multiple R-squared:  0.5072, Adjusted R-squared:  0.4693
## F-statistic: 13.38 on 9 and 117 DF,  p-value: 1.555e-14
```

Re-check assumptions

```
# linearity:
jpeg('resid_trans.jpg',height=h,width=w)
plot(bc_best_model, which = 1)
dev.off()
```

```
## pdf
##   2
```

```
#Normality
shapiro.test(residuals(bc_best_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bc_best_model)
## W = 0.97904, p-value = 0.04604
```

```
jpeg('resid_hist_trans.jpg',height=h,width=w)
ggplot(out_data,aes(residuals(bc_best_model)))+geom_histogram(col='red',fill='blue')+
  ggtitle('Distribution of Residuals of the Transformed Best Fit Interaction Model')+
  labs(x='Residuals',y='Count')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
dev.off()
```

```
## pdf
##   2
```

```
jpeg('qq_trans.jpg',height=h,width=w)
plot(bc_best_model, which=2)
```

```
## Warning: not plotting observations with leverage one:
##   17, 43, 49, 83
```

```
dev.off()
```

```
## pdf
##   2
```

```
#heteroscedasticity
bptest(bc_best_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bc_best_model
## BP = 9.1155, df = 9, p-value = 0.4267
```