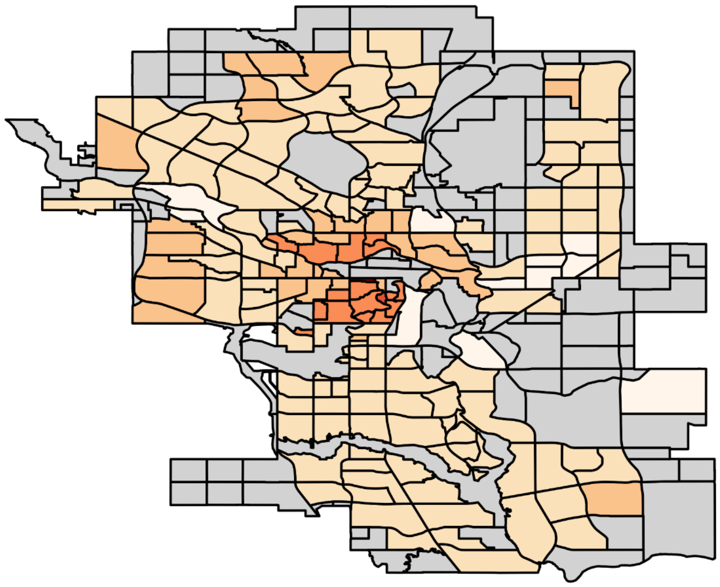


Uncovering Factors Affecting the Property Prices in Calgary: A Community-Level Analysis

Teamwork by: Alan Li 30237497

Li Chen 30153709

Xinzheng Tang 30236661

Ze Yu 30240412

Table of Content

1 Introduction

1.1 Motivation

Canada, known for being an immigration-friendly country, has experienced a significant population increase over the years, primarily due to the growth of international immigration[1]. This influx of people has contributed to a housing crisis across the nation, with the city of Calgary witnessing a surge in property prices in recent decades[2]. By the end of 2023, the price of a townhouse is expected to increase by 17.2% year-over-year, reaching $449,000[3]. It is widely acknowledged that supply and demand dynamics play a crucial role in determining long-term property prices[4]. The supply side is influenced by factors such as the number of homes for sale and new buildings available, while the demand side is affected by the number of households, economic growth, mortgage availability, interest rates, and more. Additionally, community-specific factors, including crime rates, public services, and demographic features, can also impact property prices at specific time points[5].

Whether we are international immigrants or local residents, the need to rent or purchase a home is inevitable at some point in our lives. Making this decision requires considering various factors at a specific time. This realization inspired us to conduct a community-level analysis of property prices in Calgary, focusing on data from 2019. In this study, we examine the factors affecting property prices from three perspectives: safety factors, public services, and demographic factors.

1.2 Objectives

After the data screening and further exploration, we define our guiding questions based on data availability and consistency:

1. How does safety, in terms of crime rate and disorder, impact property prices in the different communities?

2. What role do public services, such as education, attraction, community center and medical facilities, play in determining property prices in Calgary communities?

3. How do demographic factors, including gender, population, and language, influence property prices across communities in Calgary?

Through this investigation, we aim to provide valuable insights into the factors behind property prices in Calgary, thus helping both local residents and international immigrants make informed decisions when it comes to renting or purchasing a home. By shedding light on these community-level factors, we hope to contribute to a better understanding of the housing market in Calgary and address the challenges faced by individuals and families in finding suitable and affordable housing options.

2 Data and Methodology

2.1 Datasets

We collected our data in a CSV file from the City of Calgary open data source. The data was then loaded into RStudio, where all the data wrangling and data visualization was performed. For the regression analysis, we started with property unit price as a dependent variable related to 10 independent variables, including: English speaking ratio, crime rate, crime disorders, schools per person, community centers, school, clinic, attraction, library, and hospital. First of all, we read the csv file into R studio with 'read.csv' function.

2.2 Variable explanations and data Assumption

All datasets utilized in this study are publicly accessible from the Open Calgary Dataset website and the use of these datasets is permitted. All of these datasets are licensed under the following URL: https://data.calgary.ca/d/Open-Data-Terms/u45n-7awa.

Historical Property Assessments[6] (2019) (by team member: Xinzheng Tang):

This dataset comprises historical assessed values of residential, non-residential, and farm land properties in Calgary. It contains over 500,000 rows with key fields such as assessed value, community code, community name, land size, property type, and multipolygon. To ensure data accuracy, we conducted initial cleaning and transformation by removing rows with missing values in key fields. We also delineated residential properties from non-residential ones to enhance the precision of our analysis.

Census by Community[7] (2019) (by team member: Alan Li): This demographic dataset is derived from the 2019 community census available at the Open Calgary. It encompasses crucial information such as community names, gender distribution, age demographics, and languages spoken at home. These demographic factors, including population density and language distribution, are vital in understanding property pricing dynamics.

Schools in Communities and Health Clinics and Hospitals Community Services[8,9] (by Li Chen): Utilizing the datasets "Community Services" and "Schools in Communities" available on the Open Calgary website, we gathered information on public services and schools within communities. These datasets contain geographic point information, necessitating mapping to community codes for compatibility with our analysis. Key features include the presence of hospitals, libraries, community centers, and schools, all of which contribute to community attractiveness and potentially impact property prices.

Community Disorder Statistics and Community Crime Statistics[10,11] (by team member: Ze Yu): Safety data, including disorder and crime statistics, were sourced from separate datasets and combined to provide comprehensive information on community safety. By aggregating data on disorder and crime counts for each community in 2019, we obtained insights into community safety profiles, which can influence property values.

The variables utilized in our modeling process, reported annually at a community level, include:

1. Property Unit Price: Dependent variable representing the property unit price in the City of Calgary ($).
2. Eng_ratio: Independent variable indicating the percentage of English speakers at home.
3. Crime: Independent variable representing the crime rate in the City of Calgary (percentage).
4. Disorder: Independent variable indicating the number of crime disorder incidents in the City of Calgary.
5. Schools_per_person: Independent variable representing the number of schools per person in the City of Calgary.
6. Commu_center: Independent variable indicating the number of community centers in the City of Calgary.
7. Has_social_ctr: Binary independent variable indicating whether the community has social development centers.
8. Has_phs_clinic: Binary independent variable indicating whether the community has clinics.
9. Has_attraction: Binary independent variable indicating whether the community has libraries.
10. Has_hospital: Binary independent variable indicating whether the community has hospitals.

By incorporating these variables into our analysis, we aim to comprehensively explore the factors influencing property prices in Calgary communities.

2.3 Approach

In approaching this project, we adopted the methodologies acquired in Data 603. Initially, we employed a comprehensive linear regression model incorporating all predictors, followed by a rigorous assessment of variables for multicollinearity. Upon eliminating non-significant variables, we proceeded to employ pairwise regression techniques to propose a model comprising main effects. The selection of the optimal linear model was guided by the adjusted R squared value and Residual Standard Error (RSE).

Upon achieving satisfaction with our main effects model, we further scrutinized potential interactions and higher-order terms utilizing individual t-tests. Subsequently, we subjected the higher-order terms and interactions to F-tests to ascertain their significance. Any identified significant higher-order terms or interactions were subsequently incorporated into our main effects model, culminating in our final model.

Our final model underwent rigorous testing to ensure adherence to six fundamental assumptions:

- Linearity assumption - Plot of residuals versus fitted values
- Independence assumption - residual correlation
- Equal variance assumption - heteroscedasticity
- Normality assumption - normally distribution
- Multicollinearity - VIF
- Outliers - Cook's distance

2.4 Workflow

1. Data Collection:
    a. Obtain data from the City of Calgary open data source and save it in a CSV file format.
2. Data Loading and Exploration:
    a. Import the CSV file into RStudio using the 'read.csv' function.
    b. Explore the structure of the dataset using functions like 'str()' and 'summary()' to understand its variables, dimensions, and basic statistics.
    c. Identify any missing values, outliers, or data inconsistencies.
3. Data Preprocessing:
    a. Filter out rows with missing values in the response variable (property unit price) and any key predictor variables.
    a. Handle missing data using methods such as imputation or removal, ensuring data integrity.
4. Model Construction:
    a. Define the response variable (dependent variable) and predictor variables (independent variables) for the regression analysis.
    b. Initiate a full model incorporating all predictor variables to establish a baseline for comparison.
5. Model Refinement:

   a. Implement a pairwise selection method (e.g., stepwise regression) to iteratively evaluate the significance and inclusion of each predictor variable, refining the model.
   b. Assess potential interactions between variables by introducing interaction terms and examining their impact on the model's performance.
   c. Assess higher-orders for each variable and examine their impact on the model's performance.
6. Model Evaluation:
   a. Evaluate the performance of each model iteration using metrics such as adjusted R squared, residual standard error, and significance of predictors.
   b. Select the best-fitted model based on predefined criteria, considering both predictive accuracy and model complexity.
7. Assumption Testing:
   a. Validate the regression model by testing for adherence to fundamental assumptions:
      - Linearity: Examine residual plots to verify linear relationships between predictors and the response variable.
      - Independence: Assess residual correlation to ensure independence of observations.
      - Equal Variance (Homoscedasticity): Check for consistent variance of residuals across predictor values.
      - Normality: Verify the normality of residuals using diagnostic plots or statistical tests.
      - Multicollinearity: Calculate variance inflation factors (VIF) to detect multicollinearity among predictor variables.
      - Outliers: Identify influential data points using diagnostics such as Cook's distance.
8. Model Interpretation:
   a. Interpret the coefficients of the final regression model to understand the direction and magnitude of the relationships between predictors and the response variable.
   b. Visualize key relationships using plots such as scatterplots, regression lines, and residual plots to enhance understanding and communication of results.
9. Reporting and Documentation:
   a. Summarize the regression analysis findings, including the final model, key results, interpretations, and implications.

Challenges:

One of the most challenging aspects we encounter is identifying the pertinent variables for model fitting. Upon downloading data from the open Calgary website and initially employing the raw information for model fitting, we observed an exceedingly low adjusted R squared value. Through meticulous examination of the dataset, we opted to normalize the data by applying natural logarithms or per-person variables. This transformation enabled us to render the data more interpretable and to establish more pertinent factors related to housing prices.

2.5 Workload Distribution

- Searching for data and compelling R markdown file - All
- Introduction - All
- Methodology - Alan
- Main results of the analysis - Xinzheng
- Interpreting coefficients - Li
- Discussion - Ze
- Summary - All

3 Main Results of Analysis

3.1 Variable Selection Procedures

3.1.1 First order model and its hypothesis

We first manually input all independent variables to conduct a full linear regression and pick up those significant predictors (P-value < 0.05) and drop insignificant predictors (P-value > 0.05). Specifically, the predictors that should be kept are Crime, Disorder, Schools per person, Community center per person, and attraction (category predictor). We conducted the linear regression again using these independent variables and got every predictors significant (P-value < 0.05). The p-values for each predictor in this two-step process are listed in Table 1.

Table 1 P-values for each predictors in the first-order model

	Eng_r	crime	disorder	schools	comm_ctr	social_ctr	phs_clinic	attraction	library	hospital
Step1	0.3048	0.0001	2e-05	0.0230	0.0004	0.6425	0.8218	0.0002	0.2022	0.4866
Step2		2e-05	3e-06	0.045	1e-04			5e-05		

It’s worth noting that we use three all-possible-regression selection procedures to further confirm our best first-order model. Table 2 lists predictors three methods to keep and their adjusted R-squared and residual standard errors (RSE). Stepwise selection method only keeps two predictors, comm_ctr and attraction, and it has an adjusted R-squared = 0.2503, and RSE = 357.3. Forward selection method keeps four predictors, comm_ctr, attraction, Eng_r, and schools, with a higher adjusted R-squared of 0.2901, and a lower RSE of 347.7. Backward elimination method gives the result same as our manual model, which keeps predictors crime, disorder, schools, comm_ctr, and attraction, with the highest adjusted R-squared of 0.4005 and the lowest RSE of 319.5. Therefore, we choose the result from the backward elimination method as our best first-order model.

Table 2 Predictors, adjusted R-squared, and RSE of three procedure results

Selection methods	Predictors	Adjusted R-squared	RSE
Stepwise selection	comm_ctr, attraction	0.2503	357.3
Forward selection	comm_ctr, attraction, Eng_r, schools	0.2901	347.7
Backward elimination	crime, disorder, schools, comm_ctr, attraction	0.4005	319.5

Hypothesis statements for individual T-tests:

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$
$$i = crime, disorder, schools, comm\_ctr, attraction$$

Main effects individual T-test:

$$crime: t = 4.458, p - value = 2.2e - 05$$
$$disorder: t = -4.934, p - value = 3.34e - 06$$
$$schools: t = -2.031, p - value = 0.045$$
$$comm\_ctr: t = 4.067, p - value = 9.68e - 05$$
$$attraction: t = 4.248, p - value = 4.95e - 05$$

The significance level we used in our predictor selection is set as 0.05. From the results of these tests, we would reject the null hypothesis against the alternative. This suggests that variables crime, disorder, schools per person, community center, and attraction are significant predictors of residential property price on their own.

Our best first-order model is shown below:

$$\widehat{y_{RE}} = \beta_0 + \beta_1 X_{crime} + \beta_2 X_{disorder} + \beta_3 X_{schools} + \beta_4 X_{comm\_ctr} + \beta_5 X_{attraction}$$

3.1.2 Interactive terms selection and its hypothesis

Based on the best first-order we get above, we manually add all potential interactive terms into this regression model to select any significant interactions (P-value < 0.05). After the first attempt, we keep three significant interactive terms in the model, which are crime:disorder, crime:schools_per_person, and schools_per_person:commu_center, and drop those insignificant interactive terms. The second attempt to conduct regression modeling provides two significant interactive terms, crime:disorder and schools_per_person:commu_center. After keeping these two interactive terms in the model, we only have the interactive term crime:disorder significant (P-value = 0.004).

The result of the backward elimination method confirmed our interactive term selection, which improved our model's performance indicated by a higher adjusted R-square of 0.426 and a lower RSE of 312.7 compared with our first-order model.

Hypothesis statements for individual T-tests (Interactive term):

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$
$$i = crime * disorder$$

Interaction term T-tests:

$$crime * disorder: t = 2.936, p - value = 0.00415$$

Since this interactive term is a significant predictor of residential property price, we add it to our model. This also makes a practical sense the crime rate is related to disorder reported in a community.

Hypothesis statements for ANOVA Test:

$$H_0: \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0: \textit{interactive terms are not significant}$$
$$H_1: \textit{at least one } \beta_i \neq 0 \textit{ at least one interactive term is significant}$$

We conducted an ANOVA test to ensure this interactive term is significant in the presence of the first-order terms. To do this, we compared our first-order model with the interactive model (first-order + interaction). From the result of the ANOVA (F = 6.8939, p-value = 0.01007), we have sufficient evidence to reject the null hypothesis. This indicates that the interactive term significantly predicts residential property price. As a result, it is left in our model. Table 3 summarizes the results of the partial F-test.

Table 3 ANOVA table for interactive terms

Source of variation	Df	Sum of squares	Mean squares	F-statistic	P-value
Regression	1	663568	663568	6.8939	0.010
Residual	96	9240461	96258		
Total	97	9904029			

Best fitted model including interaction effects:

$$\widehat{y_{RE}} = \beta_0 + \beta_1 X_{crime} + \beta_2 X_{disorder} + \beta_3 X_{schools} + \beta_4 X_{comm\_ctr} + \beta_5 X_{attraction} + \beta_6 X_{crime} X_{disorder}$$

3.1.3 Higher-order terms selection and its hypothesis

When checking high-order terms, we plot the pair curves between the response variable and quantitative predictors to find out potential quadratic or even higher order relationships. The output of pair plots (Fig. 1) indicate that crime, schools per person and community center per person may have a higher order relationship with residential property price. Here we add a quadratic term of these variables one by one. The quadratic terms of crime and schools per person are not significant to residential property price (P-values > 0.05), while the quadratic term of community center per person is significant (P-value = 0.0205). However, when we add a three order term of community center per person to the model, all community per person terms are insignificant. Thus, the quadratic term of community center per person should be incorporated in this high-order model.

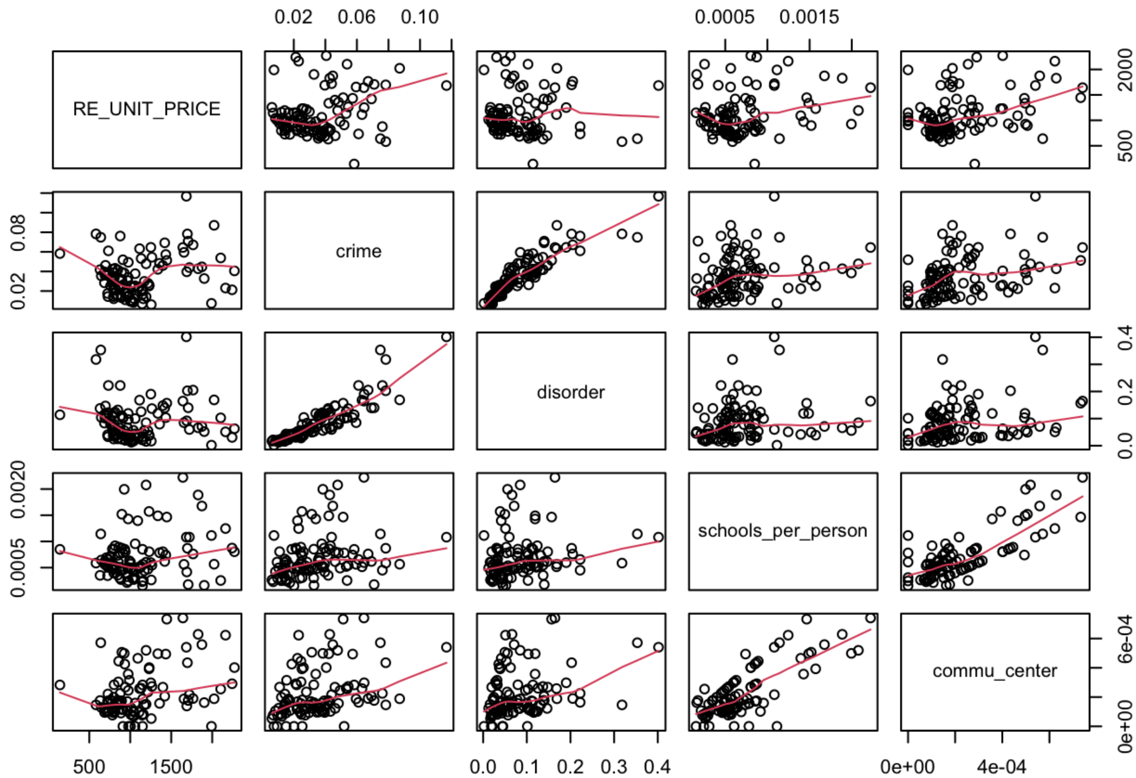


Fig. 1 Pair plots of dependent variable and predictors

Hypothesis statement for individual T-test (High-order terms):

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$
$$i = comm\_ctr^2$$

High-order individual T-test:

$$comm\_ctr^2: t = 2.357, p - value = 0.0205$$

After adding the high-order term of comm_ctr to our model, the interactive term is still significant.

Hypothesis statements for ANOVA Test:

$$H_0: \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0: \textit{high order terms are not significant}$$
$$H_1: \textit{at least one } \beta_i \neq 0 \textit{ at least one high order term is significant}$$

We conducted an ANOVA test to ensure this high term is significant in the presence of the first-order and interactive terms. To do this, we compared our high order model (first-order + interaction + high-order) with the interactive model (first-order + interaction). From the result of the ANOVA (F = 5.5544, p-value = 0.02049), we have sufficient evidence to reject the null hypothesis. This indicates that the high-order term significantly predicts residential property price. As a result, it is left in our model. Table 4 summarizes the results of the partial F-test.

Table 4 ANOVA table for high-order terms

Source of variation	Df	Sum of squares	Mean squares	F-statistic	P-value
Regression	1	510418	510418	5.5544	0.020
Residual	95	8730042	91895		
Total	96	9240461			

Best fitted model including interaction effects and high-order terms with Adjusted R-squared of 0.4604 and Residual standard error of 303.1:

$$\widehat{y_{RE}} = \beta_0 + \beta_1 X_{crime} + \beta_2 X_{disorder} + \beta_3 X_{schools} + \beta_4 X_{comm\_ctr} + \beta_5 X_{attraction} + \beta_6 X_{crime} X_{disorder} + \beta_7 X_{comm\_ctr}^{2}$$

3.2 Assumptions check

Statistical tests and models rely on assumptions of the data. In this section we tested our model to meet six assumptions associated with running multiple linear regression.

3.2.1 Linearity assumption

The multiple linear regression model assumes that there is a straight-line relationship between the response and independent variables. We plot the residuals versus predicted (fitted) values to check if any discernible pattern is presented. Fig. 2 shows that there appears to be no pattern of the residuals, indicating it met the linearity assumption . Though a less obvious downward pattern displays on the right side of the figure, it may be potentially caused by the outlier data. We will further check it by removing one outlier (observation 11) and plot the residuals versus predicted (Fig. 3). Clearly, we do not see any pattern in this plot so we conclude that our model meets the linearity assumption. The decision of whether and how to drop outliers will be discussed in the outlier assumption section 3.2.6.

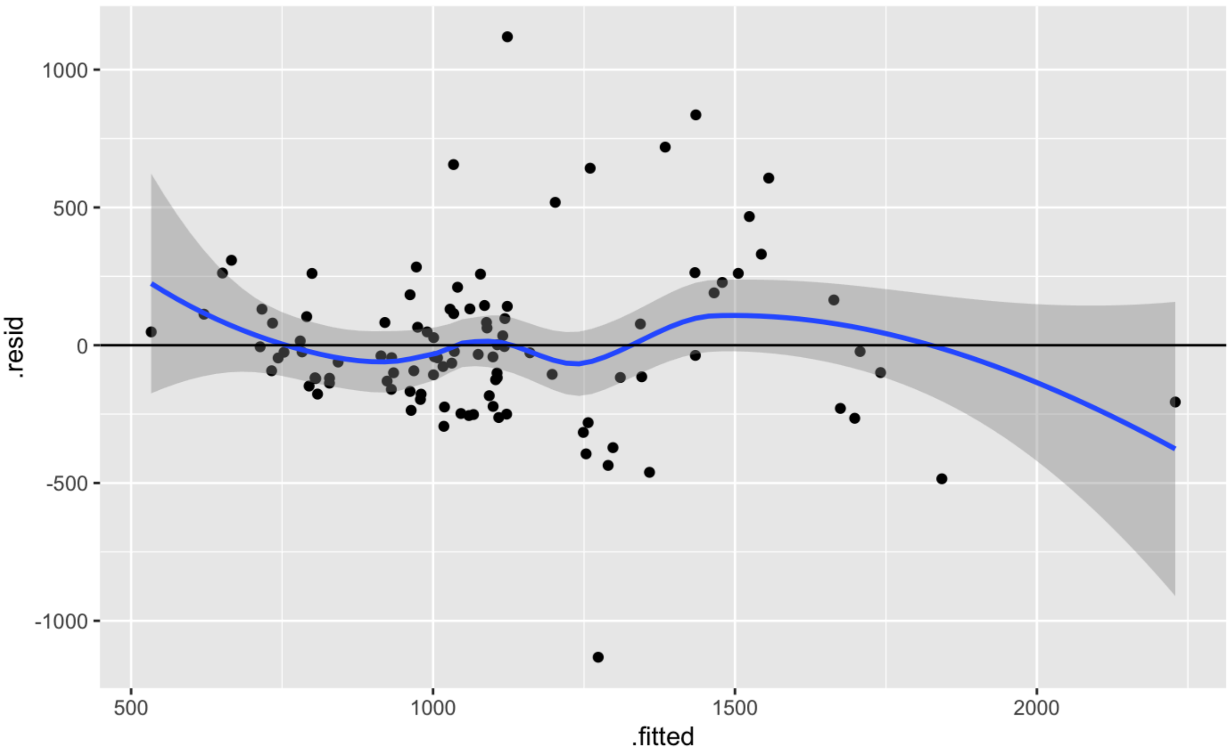


Fig. 2 Plot of residuals versus fitted values

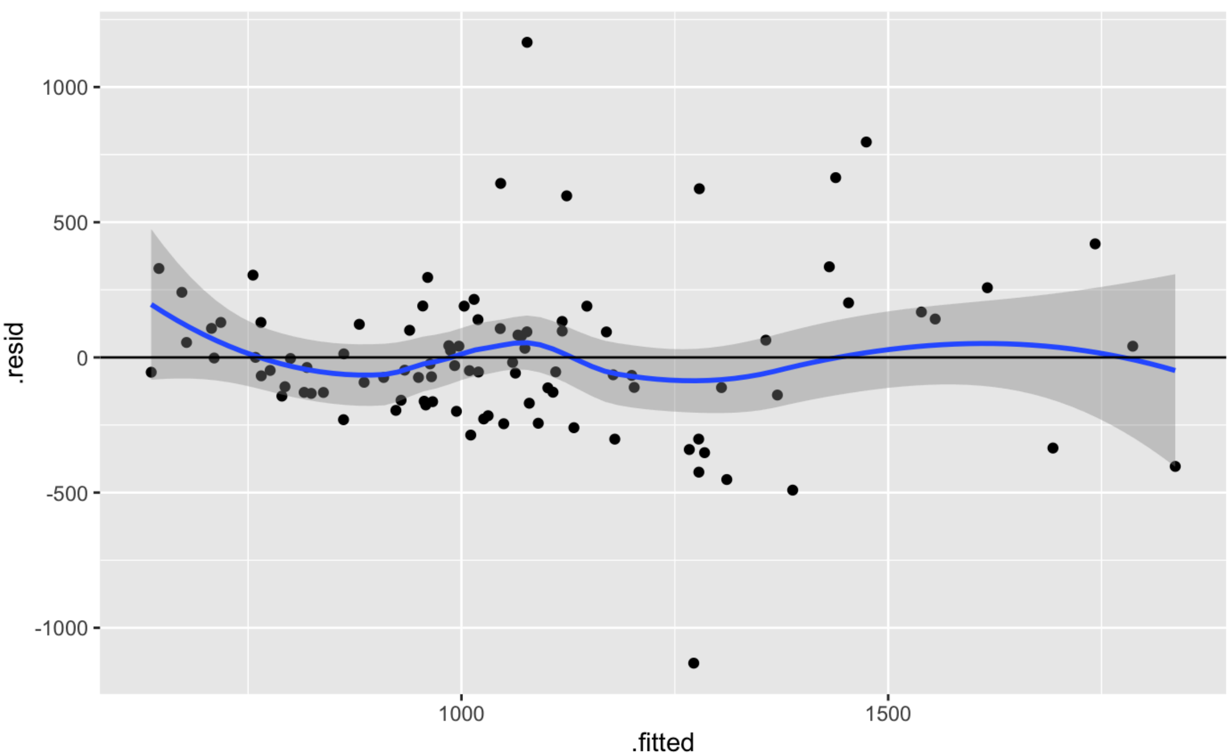


Fig. 3 Plot of residuals versus fitted values with one outlier removed

3.2.2 Independence assumption

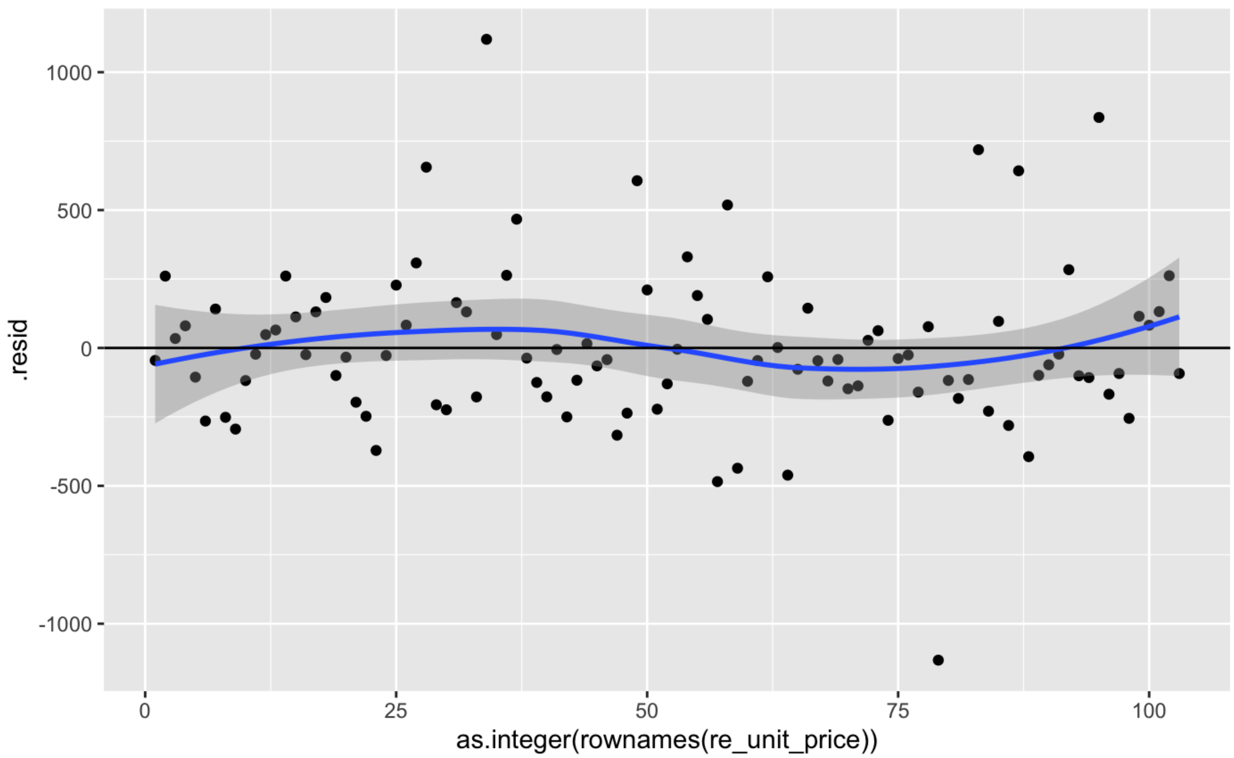


Fig. 4 Plot of residuals versus spatial order

Since all variables we used in this study were collected in 2019, our model would not have serial correlation. We then check the potential spatial association by plotting Residuals vs spatial order (Fig. 4). We can see that the plot is quite horizontal, and there is not any funneling in the residual plot. Thus, our model meets the independence assumption.

3.2.3 Equal variance assumption

Hypothesis statements for equal variance assumption:

H0: Heteroscedascity is not present

Ha: Heteroscedascity is present

Homoscedasticity is another important assumption that our multiple linear regression model should meet that means the error terms of our model should have a constant variance. We utilized the residual vs fitted plot and the scale-location plot as well as the Breusch-Pagan test to identify any non-constant variances in the errors, or heteroscedasticity. From Fig. 5 and 6, we can see a horizontal line with equally and randomly spread points, which indicates that our model meets the equal variance assumption.

From the results of the Breusch-Pagan test (BP=12.779, P-value = 0.07769 > 0.05), we do not have sufficient evidence to reject the null hypothesis, suggesting that our model succeeded to be homoscedastic.

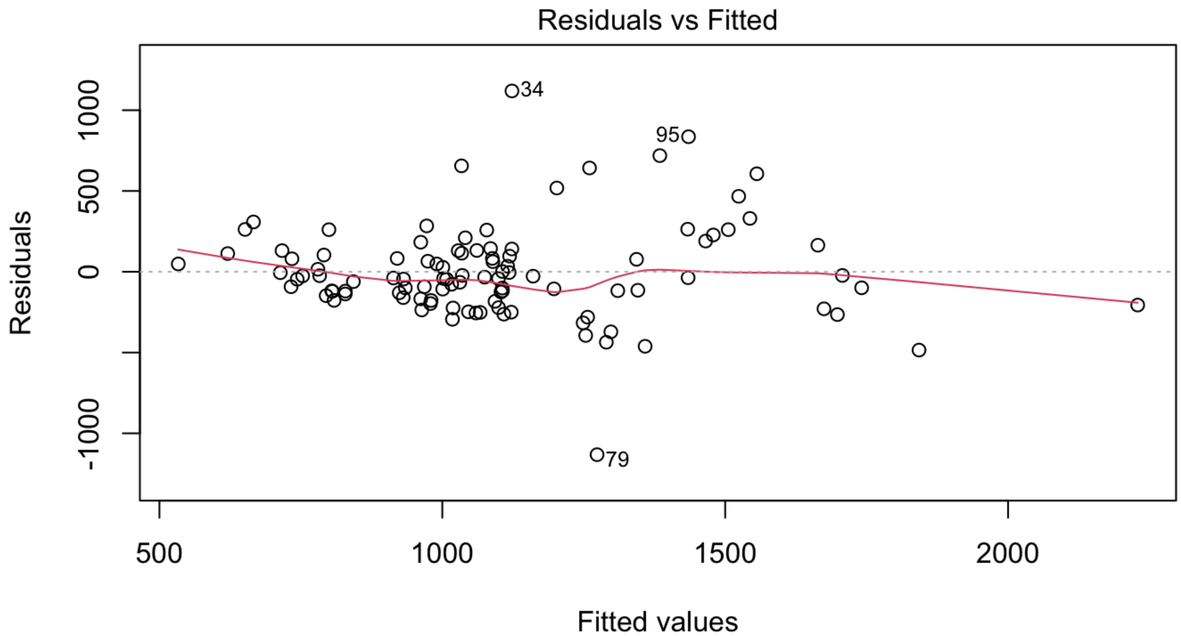


Fig. 5 Plot of residuals vs fitted

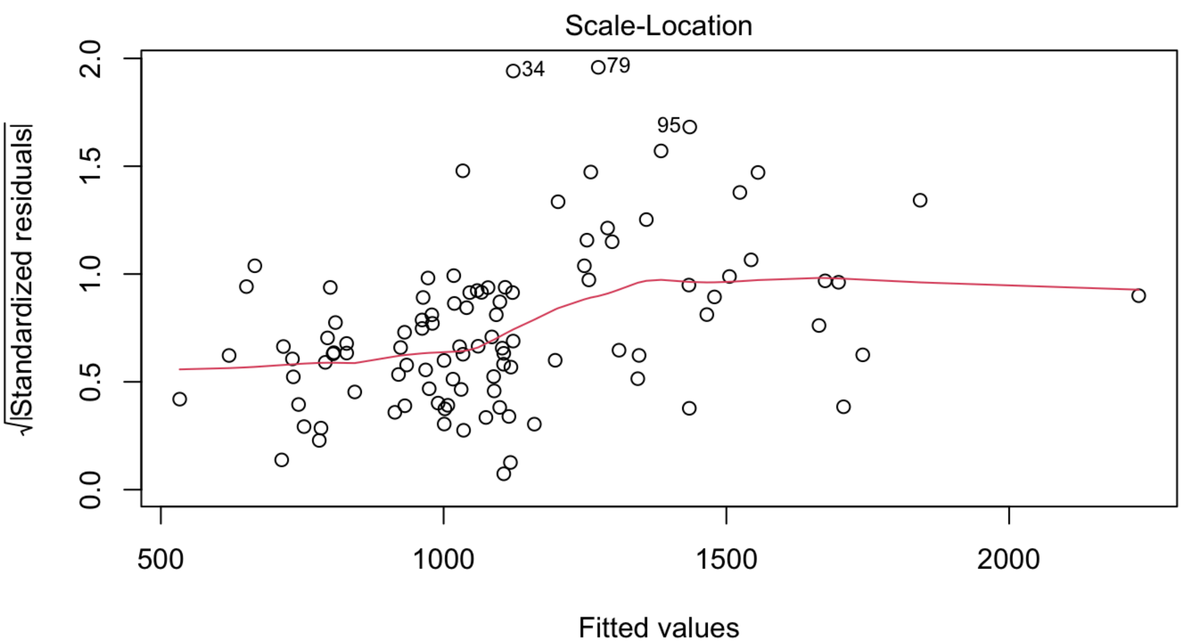


Fig. 6 Plot of scale-location

3.2.4 Normality assumption

Hypothesis statements for normality assumption

H0: the sample data are significantly normally distributed
Ha: the sample data are not significantly normally distributed

The multiple linear regression analysis requires that the errors between observed and predicted values should be normally distributed. This normality assumption of our model can be checked by looking at a histogram, a Q-Q-Plot and Shapiro-Wilk normality test. Fig. 7 shows that the residual data do not have normal distribution (from histogram and Q-Q plot). From the results of the Shapiro-Wilk normality test (W = 0.91616, p-value = 6.69e-06 < 0.05), we have sufficient evidence to reject the null hypothesis, suggesting that our model fails to have normality.

To fix this problem, we first removed outliers with the rule of thumb to identify an observation yi as influential if its leverage value hi > 2p/n. However, the residual data still do not have normal distribution from histogram and Q-Q plot in Fig. 8. Again, from the results of the Shapiro-Wilk normality test using new data (W = 0.91716, p-value = 1.455e-05), we have sufficient evidence to reject the null hypothesis, suggesting that our model using new data fails to have normality.

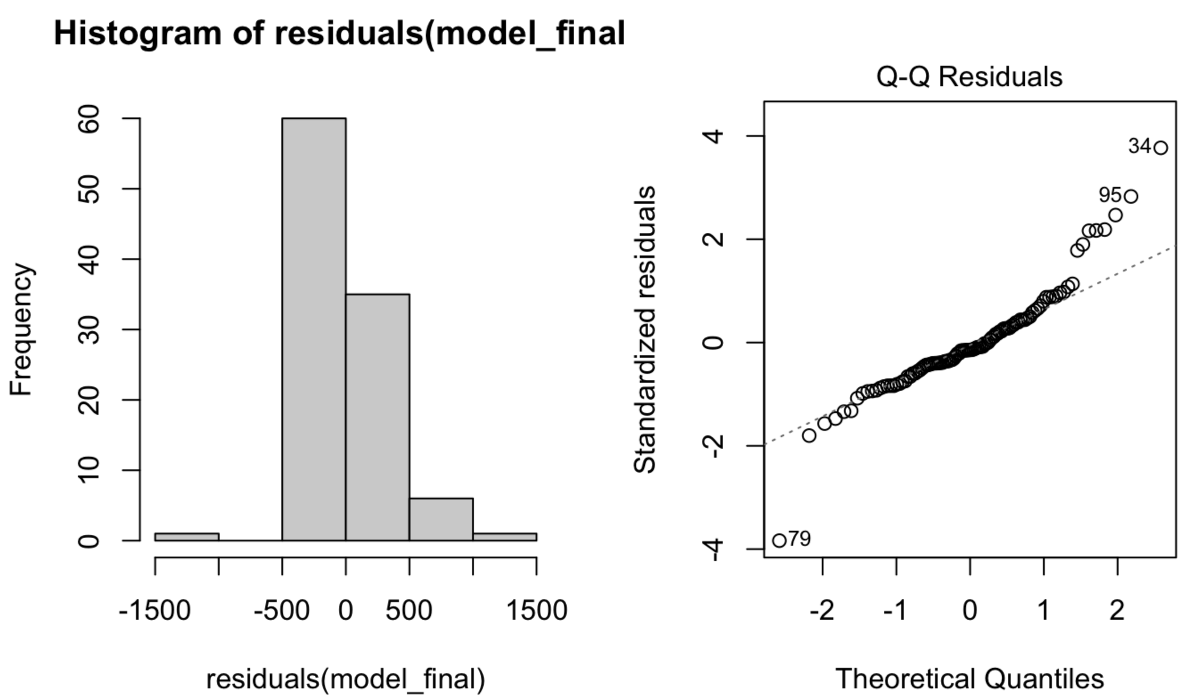


Fig. 7 Plots of histogram of residuals and Q-Q residuals

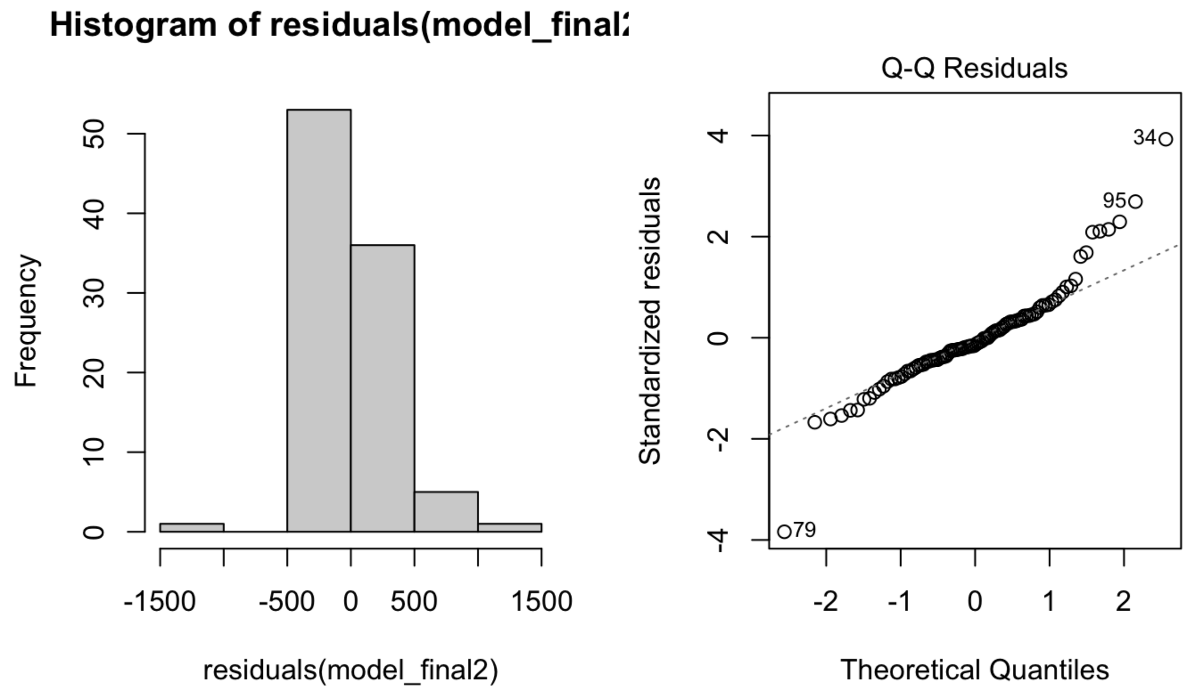


Fig. 8 Plots of histogram of residuals and Q-Q residuals after removing outliers

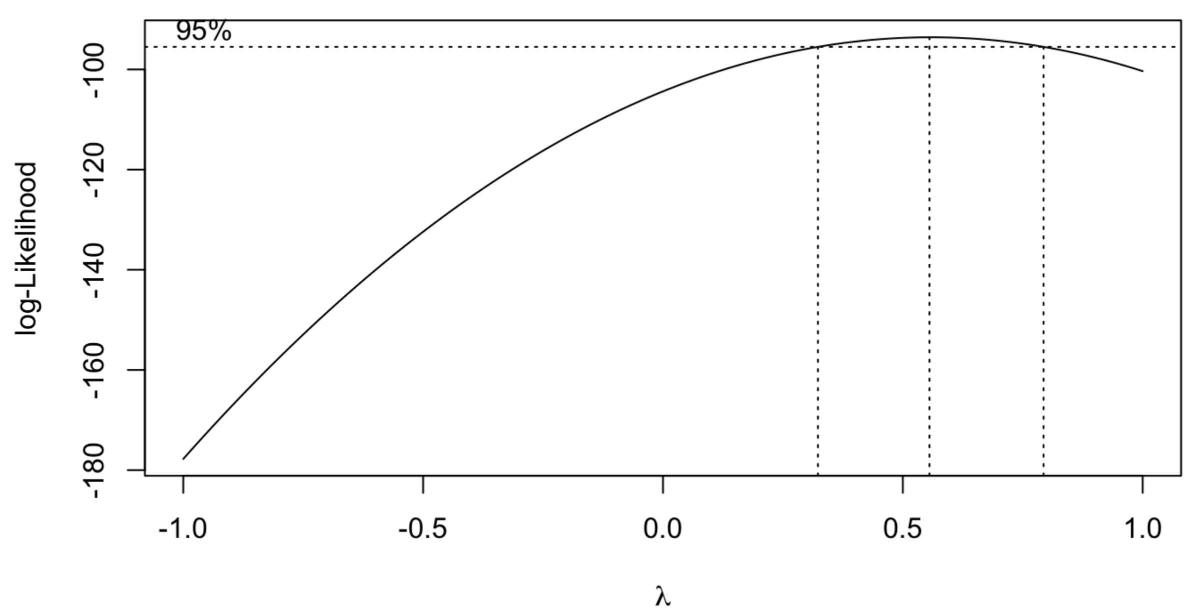


Fig. 9 Plot for Box-Cox transformations

Moreover, we conducted Box-Cox transformations for non-normality. To remedy these departures from a normal distribution, we did a transformation on Y, since the shapes and spreads of the distributions of Y need to be changed. From the output we found that “bestlambda” is approximately between 0.3 to 0.7 (Fig. 9). We chose λ=0.5555 and refitted the model. However, from the results of the Shapiro-Wilk normality test using new data (W = 0.89227, p-value = 9.587e-07), we have sufficient evidence to reject the null hypothesis, suggesting that our transformed model using new data fails to have normality.

3.2.5 Multicollinearity assumption

To test for multicollinearity of our model, we used Variance Inflation Factors (VIF) to examine any potential multicollinearity between predictors. The results show that VIF Method Failed to detect multicollinearity with VIF for five first-order predictors being 5.6752 for crime, 5.1838 for disorder, 2.7269 for schools_per_person, 3.0671 for commu_center, and 1.1282 for attraction. In addition, we also ran a ggpairs function to ensure that there were no extremely high correlations (r > 0.80) in our model (Fig. 10). The crime and disorder may have potential multicollinearity problem, but it’s still acceptable since we add an interaction term between crime and disorder in our model to include their interactive effects.

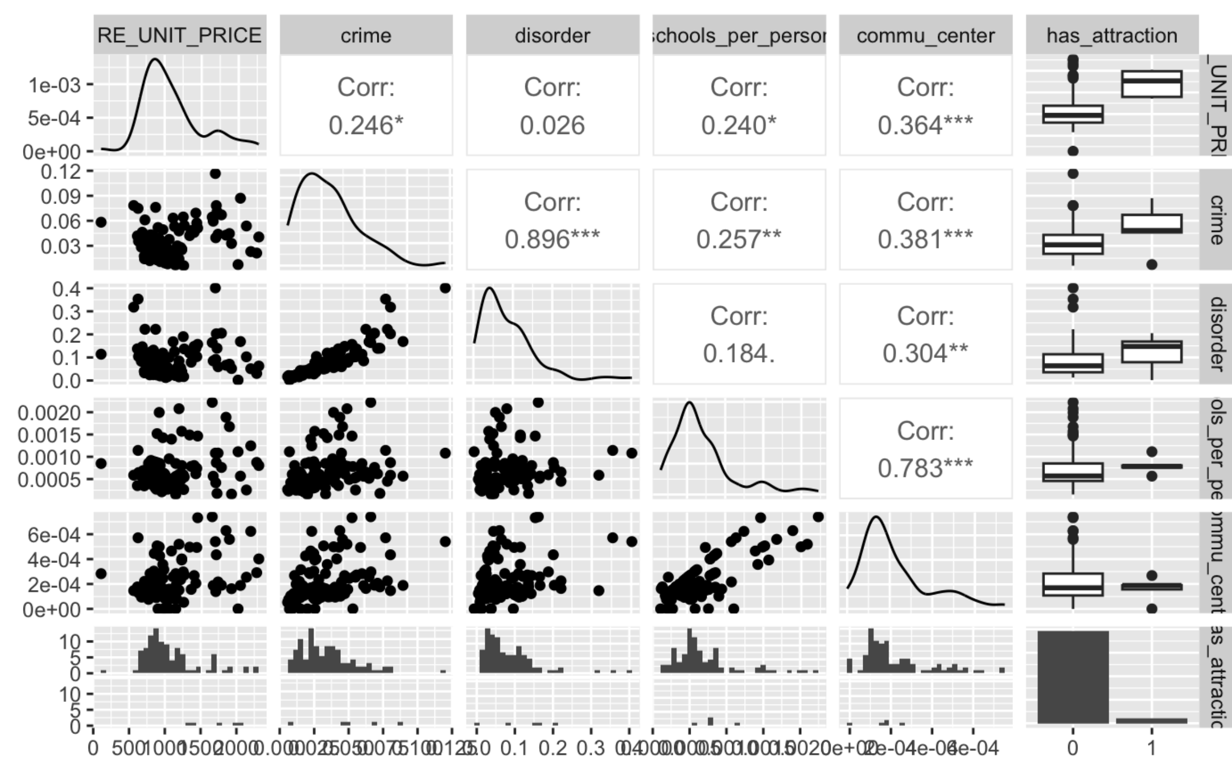


Fig. 10 Plots of ggpairs

3.2.6 Influential points and outliers

Influential points could have a great impact on our model If the parameter estimates change dramatically when the influential point is removed. To check for this we plot the residuals vs leverage against Cook's distance lines shown as dashed lines in Fig. 11. The plot is the typical look when there is no influential case because we can barely see Cook's distance lines (dashed lines) because all cases are well inside of the Cook's distance lines, suggesting that there are no influential points that have a disproportionate impact on our regression results.

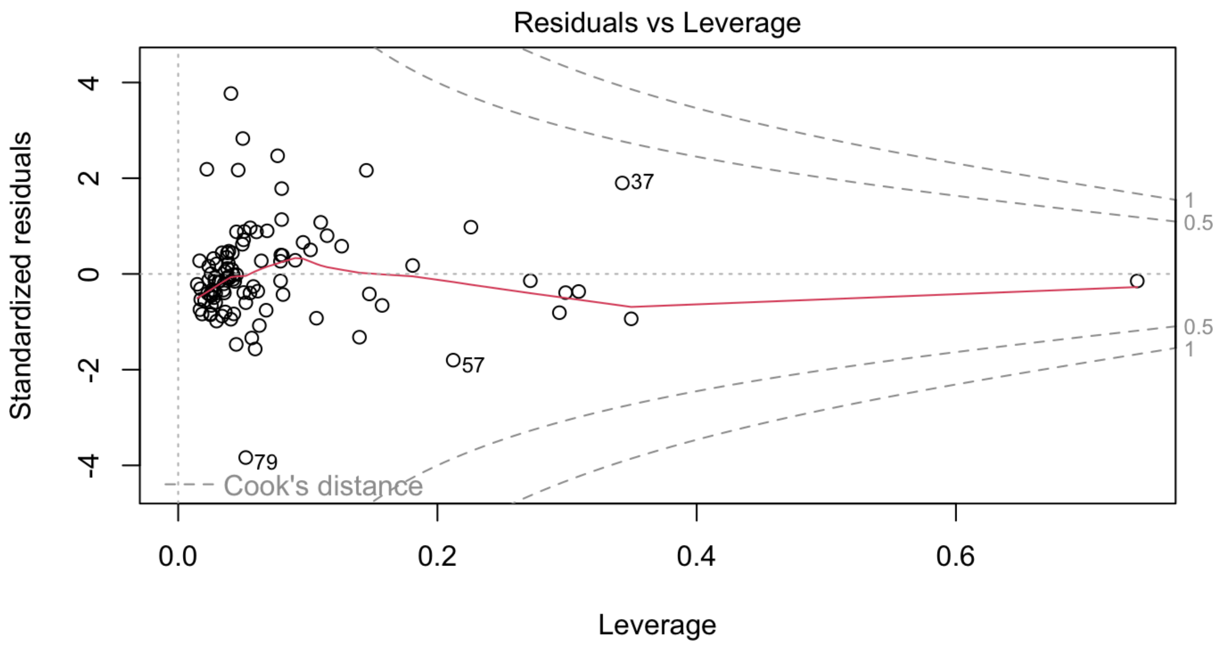


Fig. 11 Plot of residuals versus leverage

The Cook's distance plotted for each observation shown in Fig. 12 confirms our finding. This plot helps us indicate the overall influence the outlier points have on our regression by clearly identifying the observation number and the extent of its effect. The most prominent points of interest include observation number 37, 57, and 79 as they show the highest Cook's distance. However, their Cook's Distance value is all less than 0.25, so they are not influential.

Next we used the leverage plot (Fig. 13) to remove outliers beyond 2p/n and 3p/n thresholds. Our model was then refitted for both of these thresholds. For the refitted model removing outliers beyond 3p/n, there were no substantial changes to our adjusted R-squared (0.3956, smaller than our best model with adjusted R-squared 0.4604). For the refitted model removing outliers beyond 2p/n, the model fails to be fitted due to data missing in the factor variable Attraction.

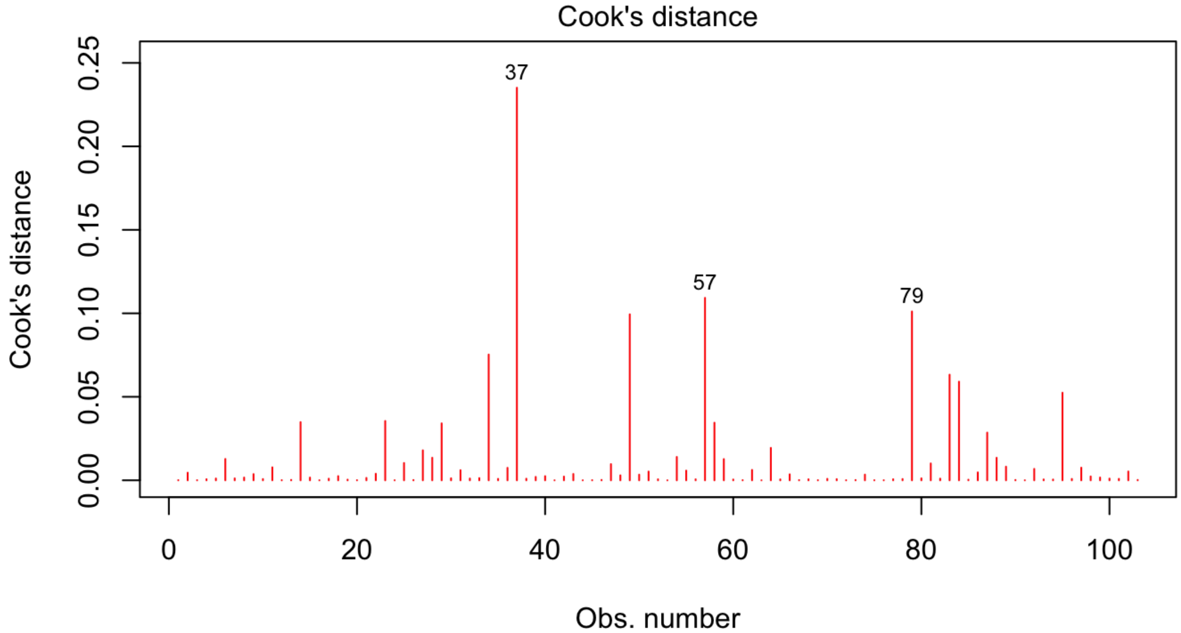


Fig. 12 Plot of Cook's distance

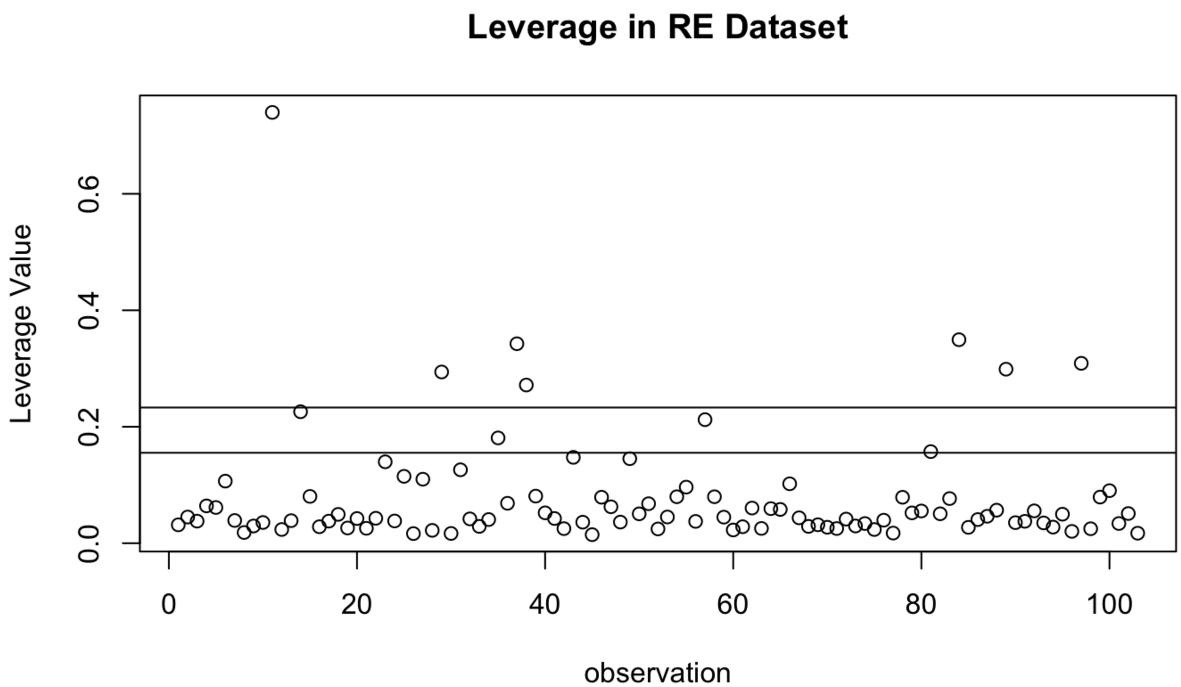


Fig. 13 Plot of leverage in RE dataset

3.3 Interpreting Coefficients

We can interpret our final model in 3 different ways.

1) Final model with all terms

$$\widehat{y_{RE}} = 1077 + 16990 * X_{\text{crime}} - 8227 * X_{\text{disorder}} - 278100 * X_{\text{schools}} - 132800 * X_{\text{comm\_ctr}} + 645.4 * X_{\text{attraction}} + 35170 * X_{\text{crime}} \times X_{\text{disorder}} + 2308000000 * X^2_{\text{comm\_ctr}}$$

2) Final model with crime collected

$$\widehat{y_{RE}} = 1077 + (16990 + 35170 \times X_{\text{disorder}}) \times X_{\text{crime}} - 8227 \times X_{\text{disorder}} - 278100 \times X_{\text{schools}} - 132800 \times X_{\text{comm\_ctr}} + 645.4 \times X_{\text{attraction}} + 2308000000 \times X^2_{\text{comm\_ctr}}$$

3) Final model with disorder collected

$$\widehat{y_{RE}} = 1077 + 16990 * X_{\text{crime}} + (35170 * X_{\text{crime}} - 8227) * X_{\text{disorder}} - 278100 * X_{\text{schools}} - 132800 * X_{\text{comm\_ctr}} + 645.4 * X_{\text{attraction}} + 2308000000 * X^2_{\text{comm\_ctr}}$$

Explanation of the coefficients:

$\widehat{y_{RE}}$: the response variable

$Intercept$: it means the response variable value when all predictor variables are zero

β_{crime}: it means for each unit increase in crime rate, the predicted value will increase 16990 units when other predictors are held constant.

$\beta_{disorder}$: it means for each unit increase in disorder rate, the predicted value will decrease 8227 units when other predictors are held constant.

$\beta_{schools}$: it means for each unit increase in schools/per person, the predicted value will decrease 278100 units when other predictors held constant.

$\beta_{attraction}$: it means the average price differences between communities with and without attractions.

$\beta_{crim*disorder}$: it represents the combined effect of the predictor crime and disorder on the response variable.

$\beta_{comm-ctr}$: since there is a higher order term of predictor attraction, therefore, the predictor ‘community center/ per persion’ here does not have a specific meaning.

$\beta^2_{comm-ctr}$: it represents there is quadratic effect of the predictor.

The Adjusted R-squared and RMSE of the best fitted model:

The R_{adj}^2 is 0.4604, meaning that 46.04% of the variance of the response variable can be explained by this model.

The $RMSE$ is 303.1, it indicates the standard deviation of the unexplained variation in estimation of response variable is 303.1.

4 Conclusion

To summarize our findings from the analysis, we observed significant main effects of crime, disorder, schools, community centers, and attractions on property prices. However, other variables showed less significance based on pairwise tests. Interactions between crime and disorder were found to be significant, while higher-order terms revealed the significance of community centers.

Upon combining main effects, interaction terms, and higher-order factors, our model met all assumptions except for the normality test. Despite the sensitivity of the Shapiro-Wilk test and limitations in the dataset, further analysis revealed a bell-shaped curve in the histogram plot, indicating a tendency towards normality. Therefore, we deemed the dataset acceptable for future analyses.

In conclusion, our analysis of property prices in Calgary has revealed insightful patterns and relationships that contribute to the City of Calgary house price. Through data collection, rigorous methodology, and thorough interpretation, we have uncovered key factors influencing property prices at the community level.

Our findings highlight the influencers for property valuation, with safety factors, public services, and demographic characteristics playing pivotal roles. The significance of crime rates, disorder incidents, and the factors such as schools, community centers, and attractions underscores the importance of both security and convenience in property value's valuation. Additionally, demographic features such as population density and language distribution provide valuable insights with shaping property prices across different communities.

By adhering to fundamental assumptions and employing robust regression modeling techniques, we have constructed a comprehensive framework for understanding the determinants of property prices in Calgary. Our analysis provides a valuable resource for comprehending the complex interaction of diverse factors influencing property prices in Calgary.

5 Discussion

In the discussion of our report, we found that the final model does not satisfy the normality assumption based on the Q-Q plot and Shapiro-Wilk test (p-value = 0.00000669, which is less than 0.05). Despite attempting a Box-Cox Transformation with $\lambda=0.5757576$ and removing outliers, the transformed data still failed the normality test with a p-value of 0.0000009188, suggesting that some adjustments to the factors might be necessary. However, since our model meets the other assumption checks, it can still be utilized to explain the data.

The adjusted R-squared value for our final model is 0.4604, meaning that 46.10% of the variance in residential unit price can be explained by our model. Although this value is not particularly high and could be improved by incorporating additional significant factors, data limitations prevented us from including potential predictors. Despite this, our final model's adjusted R-squared value is an improvement over the First Order Model's value of 0.4005, indicating some statistical significance.

Upon examining outliers, we found seven elements higher than 3p/n. However, when refitting our model with these outliers removed, the adjusted R-squared value decreased to 0.3956, and the normality assumption still failed (p-value = 0.00001455).

An intriguing finding in our model is the positive correlation between crime rate and residential property price and the negative correlation between disorder rate and residential property price. Despite the strong positive correlation between these two factors, as indicated by the interaction term, high crime rates might lead to increased security costs, such as alarm systems, security guards, or insurance premiums. Disorderly conduct, like vandalism, graffiti, or poorly maintained public spaces, can create an atmosphere of neglect and decrease a neighborhood's overall appeal. Disorder can also signal a lack of community investment and social cohesion, further reducing a neighborhood's attractiveness.

For future work, we acknowledge that our model's adjusted R-squared value is relatively low and could be improved by considering additional factors. Due to data limitations, we suggest the following factors be explored in future studies:

1. Differentiating between houses and apartments in residential property prices and accounting for non-residential properties with residential units, which could impact the accuracy of average residential property prices.

2. Addressing missing information in the public service dataset and incorporating additional factors, such as distance to public facilities, nearby main streets/roads, and school size/teaching quality.

3. Including household income data from demographic datasets, as it significantly affects purchasing power.

4. Investigating the impact of safety levels on property prices by considering factors like the number of homeless individuals and traffic conditions.

5. Incorporating data from recent years to examine trends in property prices and other factors, enabling predictions of future property prices in each community.

Reference

[1] https://www150.statcan.gc.ca/n1/daily-quotidien/221026/dq221026a-eng.htm
[2] Hardin, H. (2023). Breaking the Immigration Taboo: Canada needs to reduce immigration dramatically until the housing crisis is resolved, especially in Vancouver. Inroads: A Journal of Opinion, (53), 159-178.
[3] https://www.nesto.ca/mortgage-basics/calgary-housing-market-outlook/
[4] Leishman, C. (2024). Understanding the Role of New Housing Supply Through Macro, Micro and Behavioural Perspectives. In *The Routledge Handbook of Housing Economics* (pp. 121-132). Routledge.
[5] Li, N., Li, R. Y. M., & Nuttapong, J. (2022). Factors affect the housing prices in China: a systematic review of papers indexed in Chinese Science Citation Database. *Property Management*, *40*(5), 780-796.
[6] https://data.calgary.ca/Government/Historical-Property-Assessments-Parcel-/4ur7-wsgc/about_data
[7] https://data.calgary.ca/Demographics/Census-by-Community-2019/rkfr-buzb/about_data
[8] https://data.calgary.ca/Services-and-Amenities/Schools-in-Communities/xmep-aasr
[9] https://data.calgary.ca/Services-and-Amenities/Calgary-Health-Clinics-and-Hospitals/tsqf-wjr5
[10] https://data.calgary.ca/Health-and-Safety/Community-Disorder-Statistics/h3h6-kgme/data_preview
[11] https://data.calgary.ca/Health-and-Safety/Community-Crime-Statistics/78gh-n26t/data_preview

Appendix

R markdown codes for modeling and assumption check

DATA603_Group_Project

2024-03-31

read merged datasets

```
property_price = read.csv("new_data.csv")
head(property_price)
```

```
##   COMM_CODE       CLASS RE_UNIT_PRICE has_attraction commu_center
## 1       BED Residential      886.2819              0            1
## 2       BRE Residential     1060.1257              0            1
## 3       CHW Residential     1149.7259              0            1
## 4       ACA Residential      814.4904              0            1
## 5       CAM Residential     1091.5291              0            1
## 6       CAP Residential     1433.2848              0            2
##   commu_center_per_person has_hospital has_library has_phs_clinic
## 1             9.06618e-05            0           0              0
## 2             1.59109e-04            0           1              0
## 3             2.83286e-04            0           0              0
## 4             1.01937e-04            0           0              1
## 5             4.65116e-04            0           0              0
## 6             4.93827e-04            0           0              0
##   schools_per_person has_social_dev_ctr      MALE    FEMALE English
## 1        0.000271985                  0  5,310.00  5,171.00    8635
## 2        0.000954654                  0  2,756.00  2,825.00    5325
## 3        0.000849858                  0  3,561.00  3,864.00    3260
## 4        0.000917431                  0       204       189    8555
## 5        0.001395349                  0  1,551.00  1,548.00    2055
## 6        0.000740741                  0 13,040.00 11,302.00    3635
##   Eng_not_spk_oft_home Eng_ratio Population                 Top_lan
guage
## 1                 2395 0.7828649      11030                    Cant
onese
## 2                  960 0.8472554       6285                     Man
darin
## 3                  270 0.9235127       3530                     Man
darin
## 4                 1255 0.8720693       9810 Tagalog (Pilipino, Fili
pino)
## 5                   95 0.9558140       2150                      Sp
anish
## 6                  415 0.8975309       4050                       K
orean
##   Top_language_num Top_language_per Top_2_language Top_2_language_n
um
## 1              910             0.08       Mandarin                3
10
## 2              200             0.03      Cantonese                1
25
## 3               45             0.01         German                 
45
## 4              255             0.02        Spanish                1
60
## 5               45             0.02          Greek
```

```
10
## 6                80              0.02      Cantonese
65
##   Top_2_language_per  Top_3_language Top_3_language_num Top_3_langu
age_per
## 1               0.03         Spanish                185
0.02
## 2               0.02 Persian (Farsi)                 70
0.01
## 3               0.01          Arabic                 25
0.01
## 4               0.02         Russian                 75
0.01
## 5               0.00               0                  0
0.00
## 6               0.02         Spanish                 60
0.01
##   crime_per_person disorder_per_person
## 1       0.01504986          0.04315503
## 2       0.03770883          0.09992045
## 3       0.02181303          0.03257790
## 4       0.04322120          0.12405708
## 5       0.02232558          0.04139535
## 6       0.05728395          0.10345679
```

data type transformation

```
property_price$RE_UNIT_PRICE <- as.numeric(property_price$RE_UNIT_PRIC
E)
property_price$commu_center <- as.numeric(property_price$commu_center_
per_person)
property_price$has_hospital <- as.character(property_price$has_hospita
l)
property_price$has_library <- as.character(property_price$has_library)
property_price$has_attraction <- as.character(property_price$has_attra
ction)
property_price$has_phs_clinic <- as.character(property_price$has_phs_c
linic)
property_price$has_social_ctr <- as.character(property_price$has_socia
l_dev_ctr)
property_price$schools_per_person <- as.numeric(property_price$schools
_per_person)
property_price$Population <- as.numeric(property_price$Population)
property_price$Eng_ratio <- as.numeric(property_price$Eng_ratio)
property_price$crime <- as.numeric(property_price$crime_per_person)
property_price$disorder <- as.numeric(property_price$disorder_per_pers
on)

#remove null value in response variable
re_unit_price = property_price[!is.na(property_price[,'RE_UNIT_PRIC
E']),]
head(re_unit_price)
```

```
##   COMM_CODE       CLASS RE_UNIT_PRICE has_attraction commu_center
## 1       BED Residential      886.2819              0  9.06618e-05
## 2       BRE Residential     1060.1257              0  1.59109e-04
## 3       CHW Residential     1149.7259              0  2.83286e-04
## 4       ACA Residential      814.4904              0  1.01937e-04
## 5       CAM Residential     1091.5291              0  4.65116e-04
## 6       CAP Residential     1433.2848              0  4.93827e-04
##   commu_center_per_person has_hospital has_library has_phs_clinic
## 1             9.06618e-05            0           0              0
## 2             1.59109e-04            0           1              0
## 3             2.83286e-04            0           0              0
## 4             1.01937e-04            0           0              1
## 5             4.65116e-04            0           0              0
## 6             4.93827e-04            0           0              0
##   schools_per_person has_social_dev_ctr      MALE    FEMALE English
## 1        0.000271985                  0  5,310.00  5,171.00    8635
## 2        0.000954654                  0  2,756.00  2,825.00    5325
## 3        0.000849858                  0  3,561.00  3,864.00    3260
## 4        0.000917431                  0       204       189    8555
## 5        0.001395349                  0  1,551.00  1,548.00    2055
## 6        0.000740741                  0 13,040.00 11,302.00    3635
##   Eng_not_spk_oft_home Eng_ratio Population                 Top_lan
guage
## 1                 2395 0.7828649      11030                    Cant
onese
## 2                  960 0.8472554       6285                     Man
darin
## 3                  270 0.9235127       3530                     Man
darin
## 4                 1255 0.8720693       9810 Tagalog (Pilipino, Fili
pino)
## 5                   95 0.9558140       2150                      Sp
anish
## 6                  415 0.8975309       4050                       K
orean
##   Top_language_num Top_language_per Top_2_language Top_2_language_n
um
## 1              910             0.08       Mandarin                3
10
## 2              200             0.03      Cantonese                1
25
## 3               45             0.01         German
45
## 4              255             0.02        Spanish                1
60
## 5               45             0.02          Greek
```

```
10
## 6                 80              0.02      Cantonese
65
##   Top_2_language_per  Top_3_language Top_3_language_num Top_3_langu
age_per
## 1                0.03         Spanish                185
0.02
## 2                0.02 Persian (Farsi)                 70
0.01
## 3                0.01          Arabic                 25
0.01
## 4                0.02         Russian                 75
0.01
## 5                0.00               0                  0
0.00
## 6                0.02         Spanish                 60
0.01
##   crime_per_person disorder_per_person has_social_ctr      crime
disorder
## 1       0.01504986          0.04315503              0 0.01504986 0.
04315503
## 2       0.03770883          0.09992045              0 0.03770883 0.
09992045
## 3       0.02181303          0.03257790              0 0.02181303 0.
03257790
## 4       0.04322120          0.12405708              0 0.04322120 0.
12405708
## 5       0.02232558          0.04139535              0 0.02232558 0.
04139535
## 6       0.05728395          0.10345679              0 0.05728395 0.
10345679
```

get full model and the best fitted first order model

```
# get full model
full = lm(RE_UNIT_PRICE~ Eng_ratio + crime + disorder + schools_per_pe
rson + commu_center + has_social_ctr + has_phs_clinic + has_attraction
+ has_library + has_hospital, data=re_unit_price)
summary(full)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ Eng_ratio + crime + disorder + schools
_per_person +
##     commu_center + has_social_ctr + has_phs_clinic + has_attraction
+
##     has_library + has_hospital, data = re_unit_price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1200.4  -171.1    -8.0   136.4  1027.5
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               333.63     323.31   1.032 0.304810
## Eng_ratio                 579.03     385.50   1.502 0.136517
## crime                   15101.70    3787.37   3.987 0.000134 ***
## disorder                -4794.20    1067.48  -4.491 2.05e-05 ***
## schools_per_person -294871.04  127525.84  -2.312 0.022995 *
## commu_center       1361363.22  366609.17   3.713 0.000350 ***
## has_social_ctr1            64.36     138.17   0.466 0.642452
## has_phs_clinic1           -28.91     127.97  -0.226 0.821778
## has_attraction1           644.75     163.51   3.943 0.000157 ***
## has_library1              127.79      99.49   1.284 0.202221
## has_hospital1             242.76     347.53   0.699 0.486605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 320.4 on 92 degrees of freedom
## Multiple R-squared:  0.4563, Adjusted R-squared:  0.3973
## F-statistic: 7.722 on 10 and 92 DF,  p-value: 7.404e-09
```

```
# get the best fitted first order model
first_model = lm(RE_UNIT_PRICE~crime + disorder + schools_per_person +
commu_center  + has_attraction, data=re_unit_price)
summary(first_model)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction, data = re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1209.69  -172.11   -32.12   143.21  1067.61
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            804.6       77.6  10.369  < 2e-16 ***
## crime                16356.5     3668.7   4.458 2.22e-05 ***
## disorder             -5084.5     1030.5  -4.934 3.34e-06 ***
## schools_per_person -250205.1   123187.2  -2.031    0.045 *
## commu_center       1361984.3   334859.1   4.067 9.68e-05 ***
## has_attraction1        661.1      155.6   4.248 4.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 319.5 on 97 degrees of freedom
## Multiple R-squared:  0.4299, Adjusted R-squared:  0.4005
## F-statistic: 14.63 on 5 and 97 DF,  p-value: 1.132e-10
```

stepwise selection for the best first order model

```
stepmod=ols_step_both_p(full,p_enter = 0.05, p_remove = 0.05, details=
TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1. Eng_ratio
## 2. crime
## 3. disorder
## 4. schools_per_person
## 5. commu_center
## 6. has_social_ctr
## 7. has_phs_clinic
## 8. has_attraction
## 9. has_library
## 10. has_hospital
##
##
## Step   => 0
## Model  => RE_UNIT_PRICE ~ 1
## R2     => 0
##
## Initiating stepwise selection...
##
## Step      => 1
## Selected  => commu_center
## Model     => RE_UNIT_PRICE ~ commu_center
## R2        => 0.132
##
## Step      => 2
## Selected  => has_attraction
## Model     => RE_UNIT_PRICE ~ commu_center + has_attraction
## R2        => 0.265
##
##
## No more variables to be added or removed.
```

```
summary(stepmod$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -987.78 -225.18  -53.63  139.35 1105.12
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        850.7       60.3  14.107  < 2e-16 ***
## commu_center    984341.1   214569.6   4.588 1.30e-05 ***
## has_attraction1    698.3      164.4   4.248 4.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 357.3 on 100 degrees of freedom
## Multiple R-squared:  0.265,  Adjusted R-squared:  0.2503
## F-statistic: 18.03 on 2 and 100 DF,  p-value: 2.056e-07
```

forward selection for the best first order model

```
ExecSalFor=ols_step_forward_p(full, p_val = 0.1, details=TRUE)
```

```
## Forward Selection Method
## ------------------------
##
## Candidate Terms:
##
## 1. Eng_ratio
## 2. crime
## 3. disorder
## 4. schools_per_person
## 5. commu_center
## 6. has_social_ctr
## 7. has_phs_clinic
## 8. has_attraction
## 9. has_library
## 10. has_hospital
##
##
## Step   => 0
## Model  => RE_UNIT_PRICE ~ 1
## R2     => 0
##
## Initiating stepwise selection...
##
##                          Selection Metrics Table
## -------------------------------------------------------------------
------
## Predictor              Pr(>|t|)    R-Squared    Adj. R-Squared
AIC
## -------------------------------------------------------------------
------
## commu_center            0.00016        0.132             0.124    15
23.334
## Eng_ratio               0.00024        0.125             0.117    15
24.177
## has_attraction          0.00061        0.110             0.102    15
25.923
## crime                   0.01225        0.061             0.051    15
31.538
## schools_per_person      0.01476        0.057             0.048    15
31.875
## has_hospital            0.19087        0.017             0.007    15
36.215
## has_phs_clinic          0.27531        0.012             0.002    15
36.749
## has_social_ctr          0.46456        0.005            -0.005    15
37.420
```

```
## has_library            0.56863        0.003             -0.007    15
37.635
## disorder               0.79328        0.001             -0.009    15
37.898
## --------------------------------------------------------------------
------
##
## Step       => 1
## Selected   => commu_center
## Model      => RE_UNIT_PRICE ~ commu_center
## R2         => 0.132
##
##                           Selection Metrics Table
## --------------------------------------------------------------------
------
## Predictor               Pr(>|t|)    R-Squared    Adj. R-Squared
AIC
## --------------------------------------------------------------------
------
## has_attraction            5e-05        0.265             0.250    15
08.250
## Eng_ratio              0.01095        0.187             0.171    15
18.633
## crime                  0.21216        0.146             0.129    15
23.723
## has_social_ctr         0.30966        0.141             0.124    15
24.266
## disorder               0.34011        0.140             0.123    15
24.392
## schools_per_person     0.43615        0.138             0.120    15
24.707
## has_phs_clinic         0.47889        0.137             0.120    15
24.815
## has_library            0.76790        0.133             0.116    15
25.244
## has_hospital           0.85331        0.133             0.115    15
25.299
## --------------------------------------------------------------------
------
##
## Step       => 2
## Selected   => has_attraction
## Model      => RE_UNIT_PRICE ~ commu_center + has_attraction
## R2         => 0.265
##
##                           Selection Metrics Table
## --------------------------------------------------------------------
```

```
------
## Predictor               Pr(>|t|)    R-Squared    Adj. R-Squared
AIC
## ----------------------------------------------------------------------
------
## Eng_ratio                0.06049        0.291              0.269    15
06.565
## schools_per_person       0.08771        0.286              0.265    15
07.201
## disorder                 0.09387        0.286              0.264    15
07.315
## has_library              0.41491        0.270              0.248    15
09.555
## has_social_ctr           0.53594        0.268              0.246    15
09.849
## has_phs_clinic           0.65672        0.267              0.244    15
10.044
## crime                    0.69035        0.266              0.244    15
10.084
## has_hospital             0.85514        0.265              0.243    15
10.215
## ----------------------------------------------------------------------
------
##
## Step      => 3
## Selected  => Eng_ratio
## Model     => RE_UNIT_PRICE ~ commu_center + has_attraction + Eng_ra
tio
## R2        => 0.291
##
##                           Selection Metrics Table
## ----------------------------------------------------------------------
------
## Predictor               Pr(>|t|)    R-Squared    Adj. R-Squared
AIC
## ----------------------------------------------------------------------
------
## schools_per_person       0.05123        0.318              0.290    15
04.550
## disorder                 0.24330        0.301              0.272    15
07.127
## has_library              0.32531        0.298              0.269    15
07.543
## crime                    0.37465        0.297              0.268    15
07.733
## has_social_ctr           0.53070        0.294              0.265    15
08.150
```

```
## has_phs_clinic         0.60344        0.293              0.264    15
08.280
## has_hospital           0.73598        0.292              0.263    15
08.445
## ----------------------------------------------------------------------
------
##
## Step      => 4
## Selected  => schools_per_person
## Model     => RE_UNIT_PRICE ~ commu_center + has_attraction + Eng_ra
tio + schools_per_person
## R2        => 0.318
##
##                        Selection Metrics Table
## ----------------------------------------------------------------------
--
## Predictor         Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
## ----------------------------------------------------------------------
--
## disorder           0.16791        0.331              0.297    1504.5
21
## has_library        0.26611        0.327              0.292    1505.2
30
## has_hospital       0.44886        0.322              0.287    1505.9
38
## crime              0.47850        0.322              0.287    1506.0
14
## has_social_ctr     0.53756        0.321              0.286    1506.1
45
## has_phs_clinic     0.72085        0.319              0.284    1506.4
14
## ----------------------------------------------------------------------
--
##
##
## No more variables to be added.
##
## Variables Selected:
##
## => commu_center
## => has_attraction
## => Eng_ratio
## => schools_per_person
```

```
summary(ExecSalFor$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1028.82  -192.74   -49.74   149.25  1026.77
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              238.3      321.1   0.742 0.459800
## commu_center         1326770.9   347097.0   3.822 0.000232 ***
## has_attraction1          687.7      166.1   4.141 7.33e-05 ***
## Eng_ratio                837.4      393.6   2.128 0.035887 *
## schools_per_person -263976.1   133748.0  -1.974 0.051234 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 347.7 on 98 degrees of freedom
## Multiple R-squared:  0.318,  Adjusted R-squared:  0.2901
## F-statistic: 11.42 on 4 and 98 DF,  p-value: 1.19e-07
```

backward elimination selection for the best first order model

```
ExecSalBack=ols_step_backward_p(full, p_val = 0.1, details=TRUE)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1. Eng_ratio
## 2. crime
## 3. disorder
## 4. schools_per_person
## 5. commu_center
## 6. has_social_ctr
## 7. has_phs_clinic
## 8. has_attraction
## 9. has_library
## 10. has_hospital
##
##
## Step   => 0
## Model  => RE_UNIT_PRICE ~ Eng_ratio + crime + disorder + schools_pe
r_person + commu_center + has_social_ctr + has_phs_clinic + has_attrac
tion + has_library + has_hospital
## R2     => 0.456
##
## Initiating stepwise selection...
##
## Step     => 1
## Removed  => has_phs_clinic
## Model    => RE_UNIT_PRICE ~ Eng_ratio + crime + disorder + schools_
per_person + commu_center + has_social_ctr + has_attraction + has_libr
ary + has_hospital
## R2       => 0.45604
##
## Step     => 2
## Removed  => has_social_ctr
## Model    => RE_UNIT_PRICE ~ Eng_ratio + crime + disorder + schools_
per_person + commu_center + has_attraction + has_library + has_hospita
l
## R2       => 0.45493
##
## Step     => 3
## Removed  => has_hospital
## Model    => RE_UNIT_PRICE ~ Eng_ratio + crime + disorder + schools_
per_person + commu_center + has_attraction + has_library
## R2       => 0.4521
##
## Step     => 4
```

```
## Removed  => has_library
## Model    => RE_UNIT_PRICE ~ Eng_ratio + crime + disorder + schools_
per_person + commu_center + has_attraction
## R2       => 0.44241
##
## Step     => 5
## Removed  => Eng_ratio
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
commu_center + has_attraction
## R2       => 0.42986
##
##
## No more variables to be removed.
##
## Variables Removed:
##
## => has_phs_clinic
## => has_social_ctr
## => has_hospital
## => has_library
## => Eng_ratio
```

```
summary(ExecSalBack$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(c(include, cterms), collaps
e = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1209.69  -172.11   -32.12   143.21  1067.61
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              804.6       77.6  10.369  < 2e-16 ***
## crime                  16356.5     3668.7   4.458 2.22e-05 ***
## disorder               -5084.5     1030.5  -4.934 3.34e-06 ***
## schools_per_person -250205.1   123187.2  -2.031    0.045 *
## commu_center       1361984.3   334859.1   4.067 9.68e-05 ***
## has_attraction1        661.1      155.6   4.248 4.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 319.5 on 97 degrees of freedom
## Multiple R-squared:  0.4299, Adjusted R-squared:  0.4005
## F-statistic: 14.63 on 5 and 97 DF,  p-value: 1.132e-10
```

add interactive terms

```
model_inter =lm(RE_UNIT_PRICE~(crime + disorder + schools_per_person +
commu_center  + has_attraction)^2, data=re_unit_price)
summary(model_inter)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ (crime + disorder + schools_per_person
+
##     commu_center + has_attraction)^2, data = re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1107.50  -128.42   -17.11    96.65  1093.58
##
## Coefficients:
##                                      Estimate Std. Error t value Pr
(>|t|)
## (Intercept)                         9.681e+02  1.384e+02   6.996 5.
17e-10 ***
## crime                               2.895e+04  9.102e+03   3.180 0.
002039 **
## disorder                           -1.147e+04  3.354e+03  -3.419 0.
000958 ***
## schools_per_person                 -5.014e+04  3.772e+05  -0.133 0.
894562
## commu_center                       -5.067e+05  9.837e+05  -0.515 0.
607797
## has_attraction1                     9.174e+04  1.960e+05   0.468 0.
640863
## crime:disorder                      3.914e+04  2.102e+04   1.862 0.
065994 .
## crime:schools_per_person           -3.279e+07  1.579e+07  -2.077 0.
040764 *
## crime:commu_center                  4.714e+07  3.457e+07   1.364 0.
176204
## crime:has_attraction1              -6.131e+05  1.287e+06  -0.477 0.
634884
## disorder:schools_per_person         8.071e+06  5.009e+06   1.611 0.
110710
## disorder:commu_center              -1.185e+07  9.479e+06  -1.250 0.
214507
## disorder:has_attraction1            9.696e+04  1.911e+05   0.507 0.
613240
## schools_per_person:commu_center     1.136e+09  5.579e+08   2.036 0.
044815 *
## schools_per_person:has_attraction1 -7.746e+07  1.677e+08  -0.462 0.
645204
## commu_center:has_attraction1       -5.609e+07  1.129e+08  -0.497 0.
620519
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 300.6 on 87 degrees of freedom
## Multiple R-squared:  0.5475, Adjusted R-squared:  0.4695
## F-statistic: 7.018 on 15 and 87 DF,  p-value: 9.588e-10
```

```
model_inter1 =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person +
commu_center  + has_attraction +
                  crime*disorder + crime*schools_per_person + schools_
per_person*commu_center, data=re_unit_price)
summary(model_inter1)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction + crime * disorder + crime *
##     schools_per_person + schools_per_person * commu_center, data =
re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1085.30  -148.32   -32.01   121.71  1064.41
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|
t|)
## (Intercept)                       9.943e+02  1.341e+02   7.413 5.35e
-11 ***
## crime                             1.962e+04  4.682e+03   4.190 6.31e
-05 ***
## disorder                         -8.321e+03  1.549e+03  -5.373 5.60e
-07 ***
## schools_per_person               -2.311e+05  2.265e+05  -1.020  0.31
024
## commu_center                      3.150e+05  5.819e+05   0.541  0.58
951
## has_attraction1                   6.608e+02  1.533e+02   4.310 4.01e
-05 ***
## crime:disorder                    4.844e+04  1.668e+04   2.904  0.00
459 **
## crime:schools_per_person         -8.494e+06  6.388e+06  -1.330  0.18
687
## schools_per_person:commu_center  1.013e+09  5.451e+08   1.857  0.06
637 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307.9 on 94 degrees of freedom
## Multiple R-squared:  0.4869, Adjusted R-squared:  0.4433
## F-statistic: 11.15 on 8 and 94 DF,  p-value: 5.77e-11
```

```
model_inter2 =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person +
commu_center  + has_attraction +
                  crime*disorder + schools_per_person*commu_center, da
ta=re_unit_price)
summary(model_inter2)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction + crime * disorder + schools_per_
person *
##     commu_center, data = re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1153.92  -147.43   -33.48   119.34  1075.26
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|
t|)
## (Intercept)                      1.051e+03  1.278e+02   8.223 1.02e
-12 ***
## crime                            1.557e+04  3.573e+03   4.358 3.32e
-05 ***
## disorder                        -8.083e+03  1.544e+03  -5.233 9.94e
-07 ***
## schools_per_person              -3.888e+05  1.938e+05  -2.006   0.0
477 *
## commu_center                     7.366e+05  4.899e+05   1.504   0.1
360
## has_attraction1                  6.671e+02  1.538e+02   4.336 3.61e
-05 ***
## crime:disorder                   3.837e+04  1.492e+04   2.571   0.0
117 *
## schools_per_person:commu_center  5.223e+08  4.031e+08   1.296   0.1
982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 309.2 on 95 degrees of freedom
## Multiple R-squared:  0.4773, Adjusted R-squared:  0.4388
## F-statistic: 12.39 on 7 and 95 DF,  p-value: 3.532e-11
```

```
# get the best fitted interactive model as model_inter3
model_inter3 =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person +
commu_center  + has_attraction +
                  crime*disorder, data=re_unit_price)
summary(model_inter3)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction + crime * disorder, data = re_uni
t_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1192.09  -160.42   -35.09   107.59  1025.23
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            932.36      89.69  10.395  < 2e-16 ***
## crime                15370.46    3581.85   4.291 4.24e-05 ***
## disorder             -8185.64    1547.91  -5.288 7.76e-07 ***
## schools_per_person -192849.75  121585.45  -1.586 0.115999
## commu_center       1206504.76  330475.77   3.651 0.000426 ***
## has_attraction1        629.22     151.57   4.151 7.16e-05 ***
## crime:disorder       39274.44   14958.17   2.626 0.010066 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 310.2 on 96 degrees of freedom
## Multiple R-squared:  0.4681, Adjusted R-squared:  0.4348
## F-statistic: 14.08 on 6 and 96 DF,  p-value: 1.951e-11
```

```
# anova table to show the significance of interactive terms
print(anova(first_model,model_inter3))
```

```
## Analysis of Variance Table
##
## Model 1: RE_UNIT_PRICE ~ crime + disorder + schools_per_person + co
mmu_center +
##     has_attraction
## Model 2: RE_UNIT_PRICE ~ crime + disorder + schools_per_person + co
mmu_center +
##     has_attraction + crime * disorder
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     97 9904029
## 2     96 9240461  1    663568 6.8939 0.01007 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

backward elimination method to get the best interactive model

```
ExecSalBack=ols_step_backward_p(model_inter, p_val = 0.05, details=TRUE)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1. crime
## 2. disorder
## 3. schools_per_person
## 4. commu_center
## 5. has_attraction
## 6. crime:disorder
## 7. crime:schools_per_person
## 8. crime:commu_center
## 9. crime:has_attraction
## 10. disorder:schools_per_person
## 11. disorder:commu_center
## 12. disorder:has_attraction
## 13. schools_per_person:commu_center
## 14. schools_per_person:has_attraction
## 15. commu_center:has_attraction
##
##
## Step   => 0
## Model  => RE_UNIT_PRICE ~ crime + disorder + schools_per_person + c
ommu_center + has_attraction + crime:disorder + crime:schools_per_pers
on + crime:commu_center + crime:has_attraction + disorder:schools_per_
person + disorder:commu_center + disorder:has_attraction + schools_per
_person:commu_center + schools_per_person:has_attraction + commu_cente
r:has_attraction
## R2     => 0.548
##
## Initiating stepwise selection...
##
## Step     => 1
## Removed  => schools_per_person:has_attraction
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
commu_center + has_attraction + crime:disorder + crime:schools_per_per
son + crime:commu_center + crime:has_attraction + disorder:schools_per
_person + disorder:commu_center + disorder:has_attraction + schools_pe
r_person:commu_center + commu_center:has_attraction
## R2       => 0.54642
##
## Step     => 2
## Removed  => disorder:commu_center
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
commu_center + has_attraction + crime:disorder + crime:schools_per_per
```

```
son + crime:commu_center + crime:has_attraction + disorder:schools_per
_person + disorder:has_attraction + schools_per_person:commu_center +
commu_center:has_attraction
## R2       => 0.53773
##
## Step     => 3
## Removed  => crime:commu_center
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
commu_center + has_attraction + crime:disorder + crime:schools_per_per
son + crime:has_attraction + disorder:schools_per_person + disorder:ha
s_attraction + schools_per_person:commu_center + commu_center:has_attr
action
## R2       => 0.53475
##
## Step     => 4
## Removed  => disorder:schools_per_person
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
commu_center + has_attraction + crime:disorder + crime:schools_per_per
son + crime:has_attraction + disorder:has_attraction + schools_per_per
son:commu_center + commu_center:has_attraction
## R2       => 0.52522
##
## Step     => 5
## Removed  => crime:schools_per_person
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
commu_center + has_attraction + crime:disorder + crime:has_attraction
+ disorder:has_attraction + schools_per_person:commu_center + commu_ce
nter:has_attraction
## R2       => 0.51987
##
## Step     => 6
## Removed  => schools_per_person:commu_center
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
commu_center + has_attraction + crime:disorder + crime:has_attraction
+ disorder:has_attraction + commu_center:has_attraction
## R2       => 0.50406
##
## Step     => 7
## Removed  => crime:has_attraction
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
commu_center + has_attraction + crime:disorder + disorder:has_attracti
on + commu_center:has_attraction
## R2       => 0.48799
##
## Step     => 8
## Removed  => disorder:has_attraction
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
```

```
commu_center + has_attraction + crime:disorder + commu_center:has_attr
action
## R2       => 0.48378
##
## Step     => 9
## Removed  => commu_center:has_attraction
## Model    => RE_UNIT_PRICE ~ crime + disorder + schools_per_person +
commu_center + has_attraction + crime:disorder
## R2       => 0.46806
##
## Step     => 10
## Removed  => schools_per_person
## Model    => RE_UNIT_PRICE ~ crime + disorder + commu_center + has_a
ttraction + crime:disorder
## R2       => 0.45412
##
##
## No more variables to be removed.
##
## Variables Removed:
##
## => schools_per_person:has_attraction
## => disorder:commu_center
## => crime:commu_center
## => disorder:schools_per_person
## => crime:schools_per_person
## => schools_per_person:commu_center
## => crime:has_attraction
## => disorder:has_attraction
## => commu_center:has_attraction
## => schools_per_person
```

```
summary(ExecSalBack$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(c(include, cterms), collaps
e = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1200.97  -155.58   -27.65   108.58  1018.54
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          895.11      87.24  10.260  < 2e-16 ***
## crime              15419.15    3609.59   4.272 4.53e-05 ***
## disorder           -8470.63    1549.41  -5.467 3.55e-07 ***
## commu_center      795730.34  206887.06   3.846 0.000215 ***
## has_attraction1      575.68     148.91   3.866 0.000200 ***
## crime:disorder     43537.06   14829.30   2.936 0.004153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 312.7 on 97 degrees of freedom
## Multiple R-squared:  0.4541, Adjusted R-squared:  0.426
## F-statistic: 16.14 on 5 and 97 DF,  p-value: 1.487e-11
```

setpwise method to get the best interactive model

```
stepmod=ols_step_both_p(model_inter,p_enter = 0.05, p_remove = 0.1, de
tails=TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1. crime
## 2. disorder
## 3. schools_per_person
## 4. commu_center
## 5. has_attraction
## 6. crime:disorder
## 7. crime:schools_per_person
## 8. crime:commu_center
## 9. crime:has_attraction
## 10. disorder:schools_per_person
## 11. disorder:commu_center
## 12. disorder:has_attraction
## 13. schools_per_person:commu_center
## 14. schools_per_person:has_attraction
## 15. commu_center:has_attraction
##
##
## Step   => 0
## Model  => RE_UNIT_PRICE ~ 1
## R2     => 0
##
## Initiating stepwise selection...
##
## Step      => 1
## Selected  => crime:commu_center
## Model     => RE_UNIT_PRICE ~ crime:commu_center
## R2        => 0.136
##
## Step      => 2
## Selected  => disorder
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder
## R2        => 0.234
##
## Step      => 3
## Selected  => has_attraction
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder + has_at
traction
## R2        => 0.374
##
## Step      => 4
## Selected  => schools_per_person
```

```
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder + has_at
traction + schools_per_person
## R2        => 0.404
##
## Step      => 5
## Selected  => crime
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder + has_at
traction + schools_per_person + crime
## R2        => 0.464
##
## Step      => 6
## Selected  => commu_center
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder + has_at
traction + schools_per_person + crime + commu_center
## R2        => 0.468
##
##
## No more variables to be added or removed.
```

```
summary(stepmod$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1220.45  -166.21   -42.04   136.29  1107.02
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              997.7      105.3   9.476 1.98e-15 ***
## disorder               -5941.7     1052.4  -5.646 1.67e-07 ***
## has_attraction1          670.5      151.1   4.436 2.44e-05 ***
## schools_per_person  -246908.0   119613.0  -2.064  0.04169 *
## crime                  12887.5     3799.2   3.392  0.00101 **
## commu_center          396808.5   490730.6   0.809  0.42074
## crime:commu_center 23523997.7  8958768.7   2.626  0.01006 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 310.2 on 96 degrees of freedom
## Multiple R-squared:  0.4681, Adjusted R-squared:  0.4348
## F-statistic: 14.08 on 6 and 96 DF,  p-value: 1.95e-11
```

forward selection method to get the best interactive model

```
ExecSalFor=ols_step_forward_p(model_inter, p_val = 0.1, details=TRUE)
```

```
## Forward Selection Method
## ------------------------
##
## Candidate Terms:
##
## 1. crime
## 2. disorder
## 3. schools_per_person
## 4. commu_center
## 5. has_attraction
## 6. crime:disorder
## 7. crime:schools_per_person
## 8. crime:commu_center
## 9. crime:has_attraction
## 10. disorder:schools_per_person
## 11. disorder:commu_center
## 12. disorder:has_attraction
## 13. schools_per_person:commu_center
## 14. schools_per_person:has_attraction
## 15. commu_center:has_attraction
##
##
## Step   => 0
## Model  => RE_UNIT_PRICE ~ 1
## R2     => 0
##
## Initiating stepwise selection...
##
##                                 Selection Metrics Table
## -------------------------------------------------------------------
---------------------
## Predictor                            Pr(>|t|)    R-Squared    Adj.
R-Squared      AIC
## -------------------------------------------------------------------
---------------------
## commu_center:has_attraction              2e-05        0.198
0.182    1517.295
## crime:commu_center                     0.00013        0.136
0.127    1522.959
## commu_center                           0.00016        0.132
0.124    1523.334
## schools_per_person:has_attraction      0.00024        0.153
0.136    1522.829
## has_attraction                         0.00061        0.110
0.102    1525.923
## schools_per_person:commu_center        0.00125        0.098
```

```
0.089    1527.299
## crime:has_attraction                        0.00201        0.117
0.099    1527.174
## crime:schools_per_person                    0.00470        0.076
0.067    1529.777
## crime                                       0.01225        0.061
0.051    1531.538
## schools_per_person                          0.01476        0.057
0.048    1531.875
## disorder:has_attraction                     0.01709        0.078
0.060    1531.586
## disorder:commu_center                       0.03480        0.043
0.034    1533.402
## disorder:schools_per_person                 0.19819        0.016
0.007    1536.271
## crime:disorder                              0.19979        0.016
0.006    1536.283
## disorder                                    0.79328        0.001
-0.009    1537.898
## -----------------------------------------------------------------------
---------------------
##
## Step      => 1
## Selected  => crime:commu_center
## Model     => RE_UNIT_PRICE ~ crime:commu_center
## R2        => 0.136
##
##                                   Selection Metrics Table
## -----------------------------------------------------------------------
---------------------
## Predictor                                  Pr(>|t|)    R-Squared    Adj.
R-Squared      AIC
## -----------------------------------------------------------------------
---------------------
## disorder                                    0.00000        0.234
0.218    1512.545
## disorder:has_attraction                     0.00000        0.345
0.325    1498.366
## disorder:schools_per_person                 0.00000        0.306
0.292    1502.295
## disorder:commu_center                         3e-05        0.274
0.260    1506.954
## has_attraction                                6e-05        0.242
0.227    1511.393
## schools_per_person:has_attraction           0.00126        0.245
0.222    1513.068
## crime:has_attraction                        0.00257        0.211
```

```
0.195    1515.551
## schools_per_person                         0.00327        0.136
0.119    1524.919
## crime                                      0.00360        0.137
0.120    1524.767
## commu_center:has_attraction                0.01342        0.187
0.171    1518.631
## crime:disorder                             0.02034        0.181
0.165    1519.387
## commu_center                               0.18648        0.148
0.130    1523.528
## crime:schools_per_person                   0.18668        0.151
0.134    1523.154
## schools_per_person:commu_center            0.51845        0.139
0.122    1524.527
## ----------------------------------------------------------------------
---------------------
##
## Step      => 2
## Selected  => disorder
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder
## R2        => 0.234
##
##                                  Selection Metrics Table
## ----------------------------------------------------------------------
---------------------
## Predictor                                 Pr(>|t|)    R-Squared    Adj.
R-Squared      AIC
## ----------------------------------------------------------------------
---------------------
## has_attraction                             0.00000        0.374
0.355    1493.698
## schools_per_person                         0.00000        0.249
0.226    1512.524
## schools_per_person:has_attraction            1e-05        0.400
0.376    1491.304
## crime:has_attraction                         4e-05        0.353
0.334    1497.081
## disorder:has_attraction                      8e-05        0.345
0.325    1498.366
## crime                                        2e-04        0.350
0.330    1497.622
## commu_center                               0.00034        0.245
0.222    1512.986
## commu_center:has_attraction                0.00052        0.322
0.301    1501.970
## disorder:schools_per_person                0.00133        0.310
```

```
0.289    1503.771
## disorder:commu_center                    0.00775       0.287
0.265    1507.131
## schools_per_person:commu_center          0.05099       0.263
0.240    1510.562
## crime:disorder                           0.09445       0.255
0.233    1511.621
## crime:schools_per_person                 0.27425       0.243
0.220    1513.295
## ----------------------------------------------------------------------
---------------------
##
## Step      => 3
## Selected  => has_attraction
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder + has_at
traction
## R2        => 0.374
##
##                                  Selection Metrics Table
## ----------------------------------------------------------------------
---------------------
## Predictor                               Pr(>|t|)    R-Squared    Adj.
R-Squared      AIC
## ----------------------------------------------------------------------
---------------------
## schools_per_person                       0.00000       0.404
0.379    1490.729
## crime                                      1e-05       0.444
0.421    1483.526
## commu_center                             0.00022       0.379
0.354    1494.914
## disorder:schools_per_person              0.00026       0.454
0.432    1481.640
## disorder:commu_center                    0.01552       0.411
0.387    1489.511
## crime:schools_per_person                 0.04972       0.398
0.374    1491.630
## schools_per_person:commu_center          0.05231       0.398
0.373    1491.720
## schools_per_person:has_attraction        0.09220       0.404
0.373    1492.635
## crime:disorder                           0.15870       0.387
0.362    1493.599
## commu_center:has_attraction              0.45742       0.378
0.352    1495.115
## crime:has_attraction                     0.69466       0.375
0.350    1495.535
```

```
## disorder:has_attraction                    0.73171       0.375
0.349    1495.574
## ----------------------------------------------------------------------
---------------------
##
## Step      => 4
## Selected  => schools_per_person
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder + has_at
traction + schools_per_person
## R2        => 0.404
##
##                                 Selection Metrics Table
## ----------------------------------------------------------------------
---------------------
## Predictor                                Pr(>|t|)    R-Squared    Adj.
R-Squared      AIC
## ----------------------------------------------------------------------
---------------------
## crime                                     0.00000       0.464
0.437    1481.650
## commu_center                              0.00021       0.404
0.374    1492.611
## disorder:commu_center                     0.00338       0.454
0.426    1483.563
## disorder:schools_per_person               0.00349       0.454
0.426    1483.621
## commu_center:has_attraction               0.31896       0.410
0.379    1491.669
## crime:disorder                            0.58899       0.405
0.375    1492.418
## crime:schools_per_person                  0.75390       0.404
0.374    1492.624
## schools_per_person:has_attraction         0.76681       0.404
0.373    1492.635
## schools_per_person:commu_center           0.89024       0.404
0.373    1492.709
## disorder:has_attraction                   0.89482       0.404
0.373    1492.711
## crime:has_attraction                      0.93232       0.404
0.373    1492.721
## ----------------------------------------------------------------------
---------------------
##
## Step      => 5
## Selected  => crime
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder + has_at
traction + schools_per_person + crime
```

```
## R2        => 0.464
##
##                                  Selection Metrics Table
## ----------------------------------------------------------------------
---------------------
## Predictor                                Pr(>|t|)    R-Squared    Adj.
R-Squared      AIC
## ----------------------------------------------------------------------
---------------------
## commu_center                              0.01006        0.468
0.435    1482.951
## crime:schools_per_person                  0.06310        0.483
0.451    1479.925
## schools_per_person:has_attraction         0.18143        0.474
0.442    1481.724
## commu_center:has_attraction               0.18596        0.474
0.441    1481.763
## disorder:schools_per_person               0.18661        0.474
0.441    1481.769
## crime:has_attraction                      0.37580        0.469
0.436    1482.804
## crime:disorder                            0.40047        0.468
0.435    1482.888
## schools_per_person:commu_center           0.51590        0.467
0.433    1483.195
## disorder:commu_center                     0.54948        0.466
0.433    1483.264
## disorder:has_attraction                   0.75664        0.465
0.432    1483.546
## ----------------------------------------------------------------------
---------------------
##
## Step      => 6
## Selected  => commu_center
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder + has_at
traction + schools_per_person + crime + commu_center
## R2        => 0.468
##
##                                  Selection Metrics Table
## ----------------------------------------------------------------------
---------------------
## Predictor                                Pr(>|t|)    R-Squared    Adj.
R-Squared      AIC
## ----------------------------------------------------------------------
---------------------
## crime:schools_per_person                  0.08517        0.484
0.447    1481.721
```

```
## commu_center:has_attraction           0.09194        0.484
0.446    1481.855
## schools_per_person:has_attraction     0.10243        0.483
0.445    1482.044
## crime:disorder                        0.20428        0.477
0.439    1483.194
## crime:has_attraction                  0.25096        0.475
0.437    1483.514
## disorder:schools_per_person           0.29562        0.474
0.435    1483.759
## disorder:has_attraction               0.56052        0.470
0.431    1484.582
## disorder:commu_center                 0.71005        0.469
0.430    1484.800
## schools_per_person:commu_center       0.73686        0.469
0.430    1484.828
## -------------------------------------------------------------------
---------------------
##
## Step      => 7
## Selected  => crime:schools_per_person
## Model     => RE_UNIT_PRICE ~ crime:commu_center + disorder + has_at
traction + schools_per_person + crime + commu_center + crime:schools_p
er_person
## R2        => 0.484
##
##                                  Selection Metrics Table
## -------------------------------------------------------------------
---------------------
## Predictor                            Pr(>|t|)    R-Squared    Adj.
R-Squared      AIC
## -------------------------------------------------------------------
---------------------
## schools_per_person:commu_center       0.12052        0.498
0.455    1481.066
## crime:disorder                        0.33197        0.490
0.446    1482.684
## commu_center:has_attraction           0.35329        0.489
0.446    1482.772
## schools_per_person:has_attraction     0.38223        0.489
0.445    1482.880
## disorder:commu_center                 0.52774        0.487
0.443    1483.282
## crime:has_attraction                  0.73567        0.485
0.441    1483.595
## disorder:has_attraction               0.74022        0.485
0.441    1483.600
```

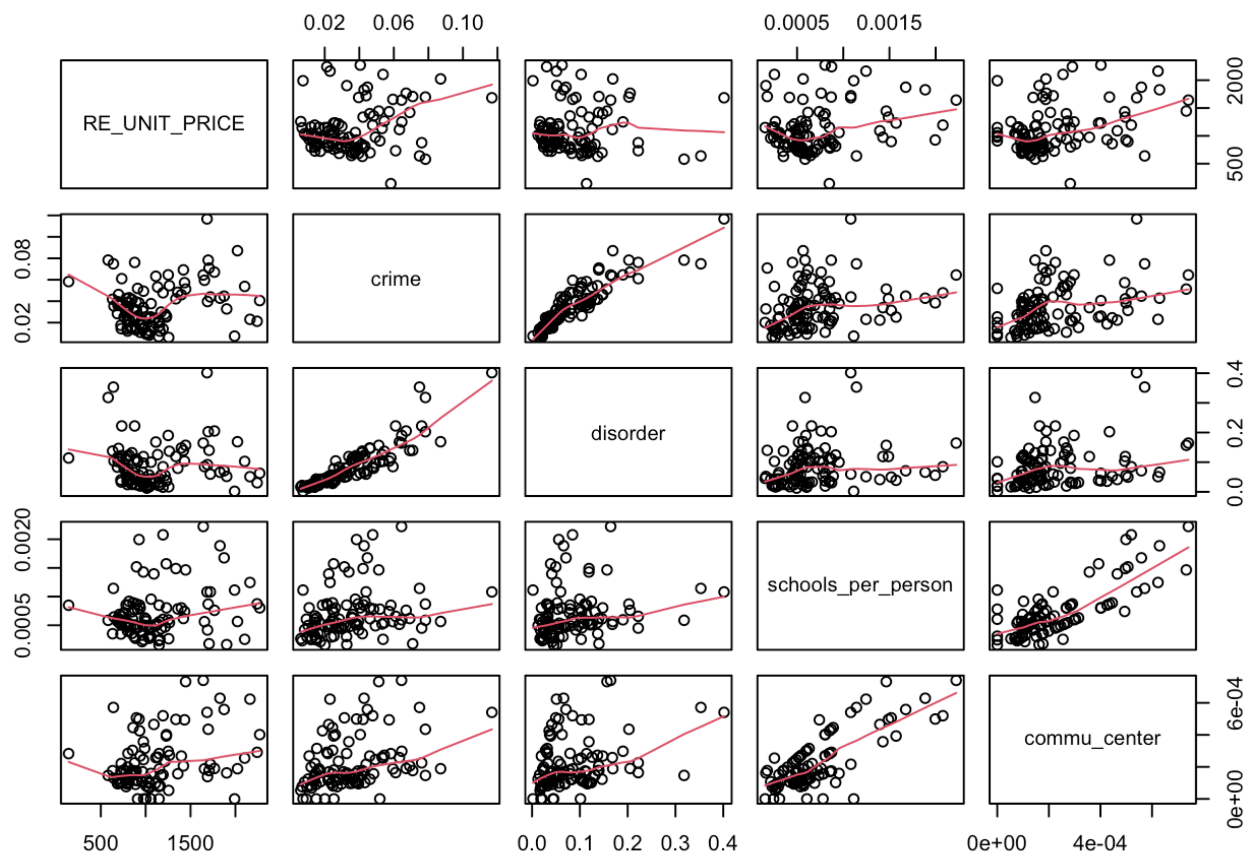
```
## disorder:schools_per_person          0.82360       0.485
0.441    1483.666
## ------------------------------------------------------------------
---------------------
##
##
## No more variables to be added.
##
## Variables Selected:
##
## => crime:commu_center
## => disorder
## => has_attraction
## => schools_per_person
## => crime
## => commu_center
## => crime:schools_per_person
```

```
summary(ExecSalFor$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1190.30  -167.64   -13.02   114.14  1074.11
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                8.676e+02  1.282e+02   6.766 1.08e-09 ***
## disorder                  -5.880e+03  1.042e+03  -5.643 1.72e-07 ***
## has_attraction1            6.321e+02  1.512e+02   4.182 6.45e-05 ***
## schools_per_person         2.055e+05  2.857e+05   0.719 0.473777
## crime                      1.590e+04  4.139e+03   3.841 0.000221 ***
## commu_center              -2.680e+05  6.180e+05  -0.434 0.665498
## crime:commu_center         4.184e+07  1.376e+07   3.040 0.003059 **
## crime:schools_per_person -1.173e+07  6.745e+06  -1.740 0.085173 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307 on 95 degrees of freedom
## Multiple R-squared:  0.4845, Adjusted R-squared:  0.4465
## F-statistic: 12.75 on 7 and 95 DF,  p-value: 1.896e-11
```

pairs plots to explore potential high order terms

```
pairs(~RE_UNIT_PRICE+crime + disorder + schools_per_person + commu_cen
ter, data = re_unit_price, panel = panel.smooth)
```



add high order terms

```
model_high =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person + c
ommu_center  + has_attraction +
                 crime*disorder + I(crime^2), data=re_unit_price)
summary(model_high)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction + crime * disorder + I(crime^2),
##     data = re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1193.00  -160.18   -36.31   109.13  1028.21
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            941.8      115.8   8.136 1.55e-12 ***
## crime                13796.6    12596.2   1.095 0.276155
## disorder             -7804.7     3310.4  -2.358 0.020444 *
## schools_per_person -189810.2   124416.3  -1.526 0.130431
## commu_center       1207437.0   332257.8   3.634 0.000453 ***
## has_attraction1        624.4      156.8   3.983 0.000133 ***
## I(crime^2)           23583.1   180871.8   0.130 0.896537
## crime:disorder       32769.9    52103.2   0.629 0.530896
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 311.9 on 95 degrees of freedom
## Multiple R-squared:  0.4682, Adjusted R-squared:  0.429
## F-statistic: 11.95 on 7 and 95 DF,  p-value: 7.684e-11
```

```
model_high1 =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person +
commu_center  + has_attraction +
                  crime*disorder + I(disorder^2), data=re_unit_price)
summary(model_high1)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction + crime * disorder + I(disorder^
2),
##     data = re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1191.84  -160.60   -35.21   107.76  1025.95
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            933.98      96.12   9.717 6.64e-16 ***
## crime                15084.44    6894.09   2.188 0.031120 *
## disorder             -8115.75    2117.75  -3.832 0.000228 ***
## schools_per_person -192283.87  122774.42  -1.566 0.120637
## commu_center       1207646.58  333034.46   3.626 0.000465 ***
## has_attraction1        628.48     153.12   4.104 8.57e-05 ***
## I(disorder^2)         -539.65   11092.55  -0.049 0.961300
## crime:disorder       41236.13   43034.97   0.958 0.340394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 311.9 on 95 degrees of freedom
## Multiple R-squared:  0.4681, Adjusted R-squared:  0.4289
## F-statistic: 11.94 on 7 and 95 DF,  p-value: 7.737e-11
```

```
model_high2 =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person +
commu_center  + has_attraction +
                  crime*disorder + I(schools_per_person^2), data=re_un
it_price)
summary(model_high2)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction + crime * disorder + I(schools_pe
r_person^2),
##     data = re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1149.10  -142.84   -28.56   115.72  1070.74
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.040e+03  1.313e+02   7.919 4.45e-12 ***
## crime                    1.488e+04  3.604e+03   4.130 7.80e-05 ***
## disorder                -7.899e+03  1.567e+03  -5.041 2.21e-06 ***
## schools_per_person      -4.912e+05  2.931e+05  -1.676 0.097050 .
## commu_center             1.209e+06  3.301e+05   3.664 0.000408 ***
## has_attraction1          6.610e+02  1.540e+02   4.292 4.27e-05 ***
## I(schools_per_person^2)  1.428e+08  1.277e+08   1.118 0.266234
## crime:disorder           3.850e+04  1.495e+04   2.574 0.011584 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 309.8 on 95 degrees of freedom
## Multiple R-squared:  0.475,  Adjusted R-squared:  0.4363
## F-statistic: 12.28 on 7 and 95 DF,  p-value: 4.31e-11
```

```
# get the best fitted high order model as model_high3
model_high3 =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person +
commu_center  + has_attraction +
                  crime*disorder + I(commu_center^2), data=re_unit_pri
ce)
summary(model_high3)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction + crime * disorder + I(commu_cent
er^2),
##     data = re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1132.18  -154.10   -38.54   122.75  1119.24
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.077e+03  1.069e+02  10.070  < 2e-16 ***
## crime                1.699e+04  3.566e+03   4.763 6.82e-06 ***
## disorder            -8.227e+03  1.513e+03  -5.439 4.15e-07 ***
## schools_per_person -2.781e+05  1.242e+05  -2.240   0.0275 *
## commu_center        -1.328e+05  6.536e+05  -0.203   0.8394
## has_attraction1      6.454e+02  1.483e+02   4.353 3.38e-05 ***
## I(commu_center^2)    2.308e+09  9.792e+08   2.357   0.0205 *
## crime:disorder       3.517e+04  1.472e+04   2.390   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 303.1 on 95 degrees of freedom
## Multiple R-squared:  0.4974, Adjusted R-squared:  0.4604
## F-statistic: 13.43 on 7 and 95 DF,  p-value: 6.028e-12
```

```
model_high4 =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person +
commu_center  + has_attraction +
                  crime*disorder + I(commu_center^2) + I(commu_center^
3), data=re_unit_price)
summary(model_high4)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction + crime * disorder + I(commu_cent
er^2) +
##     I(commu_center^3), data = re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1130.67  -158.49   -35.16   124.07  1127.04
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.111e+03  1.244e+02   8.927 3.51e-14 ***
## crime               1.711e+04  3.586e+03   4.770 6.73e-06 ***
## disorder           -7.992e+03  1.579e+03  -5.062 2.06e-06 ***
## schools_per_person -2.957e+05  1.288e+05  -2.295   0.0239 *
## commu_center       -7.661e+05  1.342e+06  -0.571   0.5694
## has_attraction1     6.492e+02  1.490e+02   4.358 3.35e-05 ***
## I(commu_center^2)   4.765e+09  4.647e+09   1.025   0.3078
## I(commu_center^3)  -2.364e+12  4.370e+12  -0.541   0.5898
## crime:disorder      3.242e+04  1.563e+04   2.075   0.0407 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 304.3 on 94 degrees of freedom
## Multiple R-squared:  0.499,  Adjusted R-squared:  0.4564
## F-statistic:  11.7 on 8 and 94 DF,  p-value: 2.024e-11
```

get the final best fitted model

```
model_final =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person +
commu_center  + has_attraction +
                  crime*disorder + I(commu_center^2), data=re_unit_pri
ce)
summary(model_final)
```

```
##
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person
+
##     commu_center + has_attraction + crime * disorder + I(commu_cent
er^2),
##     data = re_unit_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1132.18  -154.10   -38.54   122.75  1119.24
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.077e+03  1.069e+02  10.070  < 2e-16 ***
## crime                1.699e+04  3.566e+03   4.763 6.82e-06 ***
## disorder            -8.227e+03  1.513e+03  -5.439 4.15e-07 ***
## schools_per_person -2.781e+05  1.242e+05  -2.240   0.0275 *
## commu_center        -1.328e+05  6.536e+05  -0.203   0.8394
## has_attraction1      6.454e+02  1.483e+02   4.353 3.38e-05 ***
## I(commu_center^2)    2.308e+09  9.792e+08   2.357   0.0205 *
## crime:disorder       3.517e+04  1.472e+04   2.390   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 303.1 on 95 degrees of freedom
## Multiple R-squared:  0.4974, Adjusted R-squared:  0.4604
## F-statistic: 13.43 on 7 and 95 DF,  p-value: 6.028e-12
```

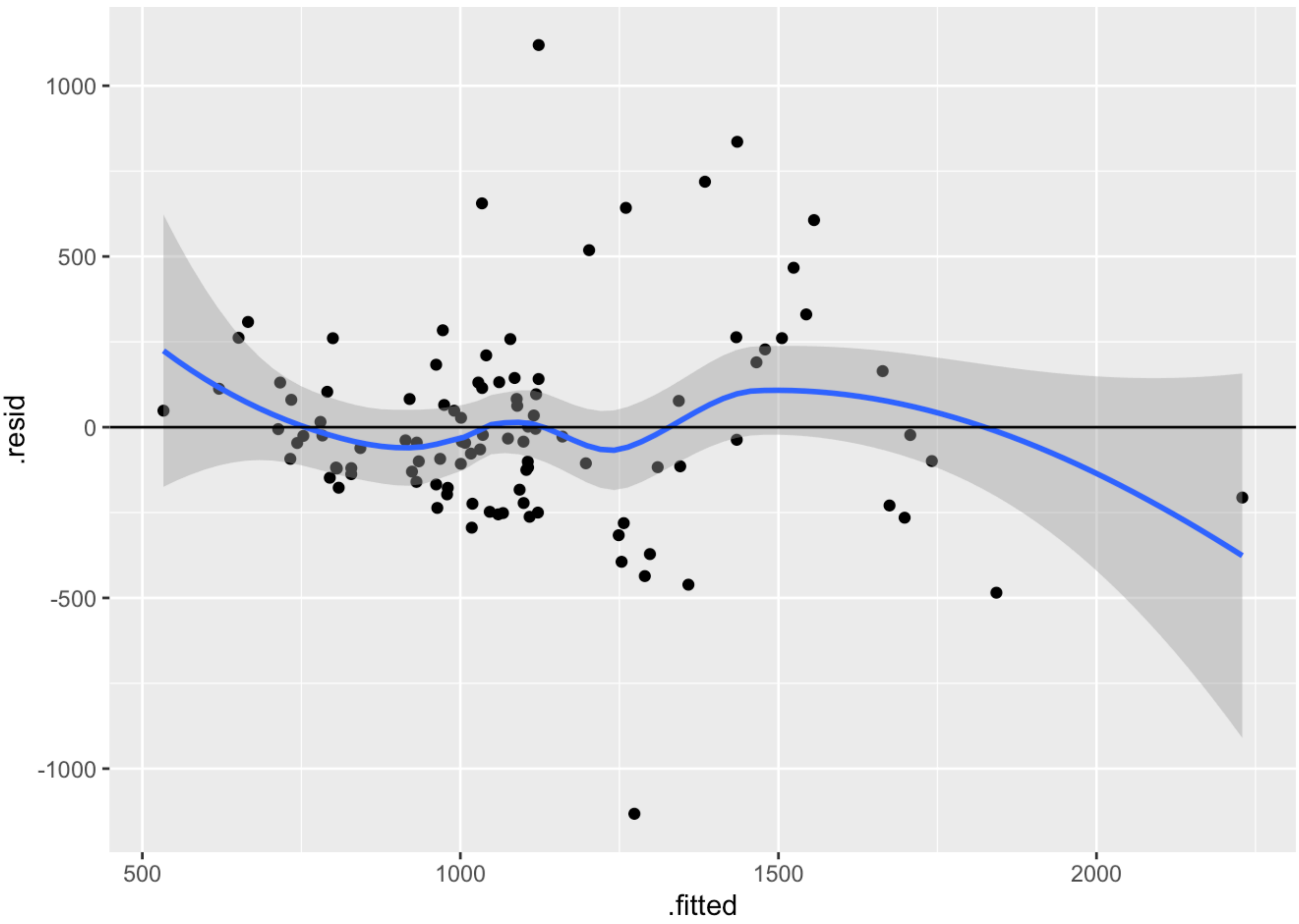
```
# anova table to show the significance of high order terms
print(anova(model_inter3,model_final))
```

```
## Analysis of Variance Table
##
## Model 1: RE_UNIT_PRICE ~ crime + disorder + schools_per_person + co
mmu_center +
##     has_attraction + crime * disorder
## Model 2: RE_UNIT_PRICE ~ crime + disorder + schools_per_person + co
mmu_center +
##     has_attraction + crime * disorder + I(commu_center^2)
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     96 9240461
## 2     95 8730042  1    510418 5.5544 0.02049 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

assumptions check for the final model

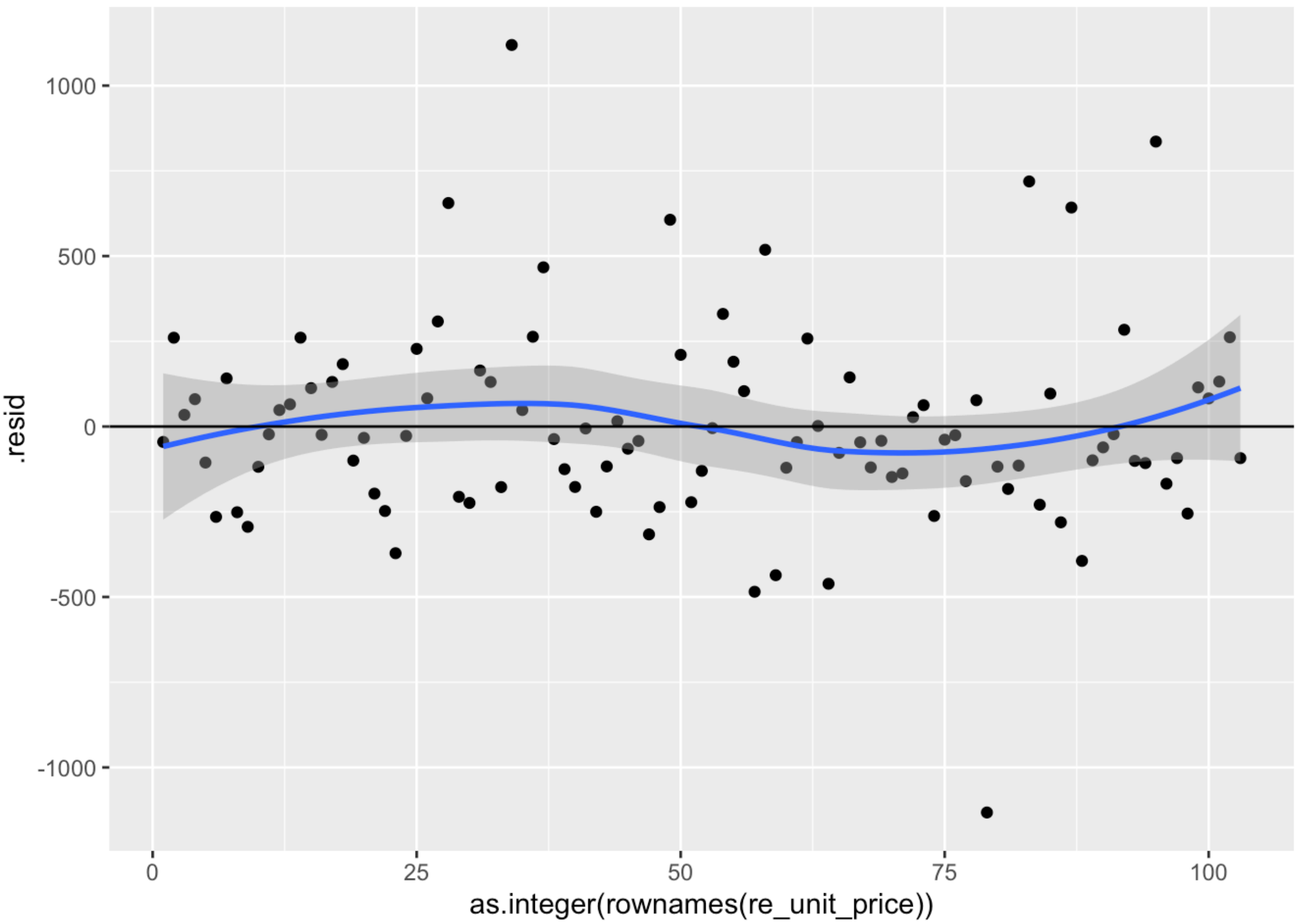
```
# linearity
ggplot(model_final, aes(x=.fitted, y=.resid)) +
  geom_point() +geom_smooth()+
  geom_hline(yintercept = 0)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

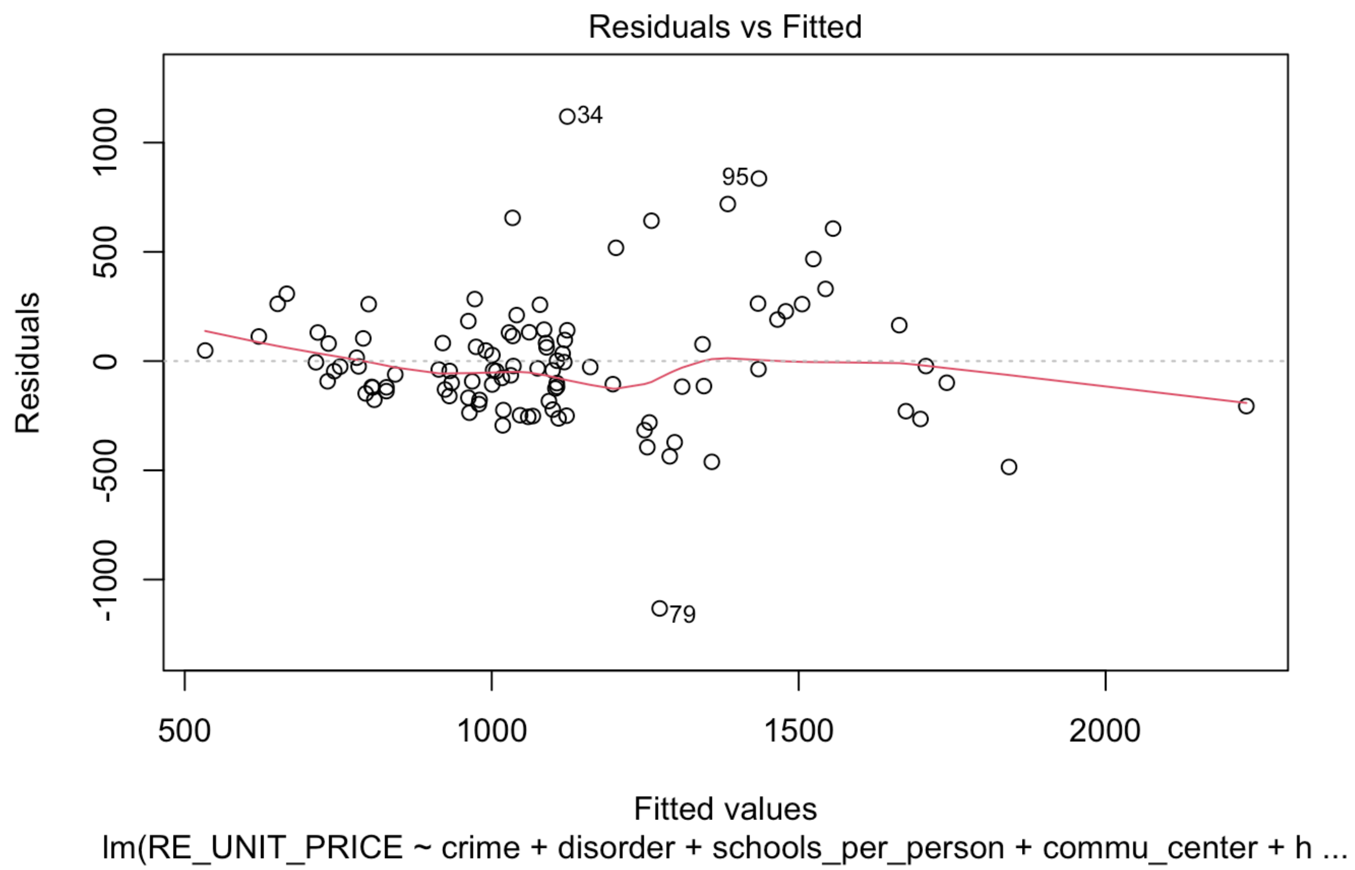



```
# independence
ggplot(model_final, aes(x=as.integer(rownames(re_unit_price)), y=.resi
d)) +
  geom_point() +geom_smooth()+
  geom_hline(yintercept = 0)
```

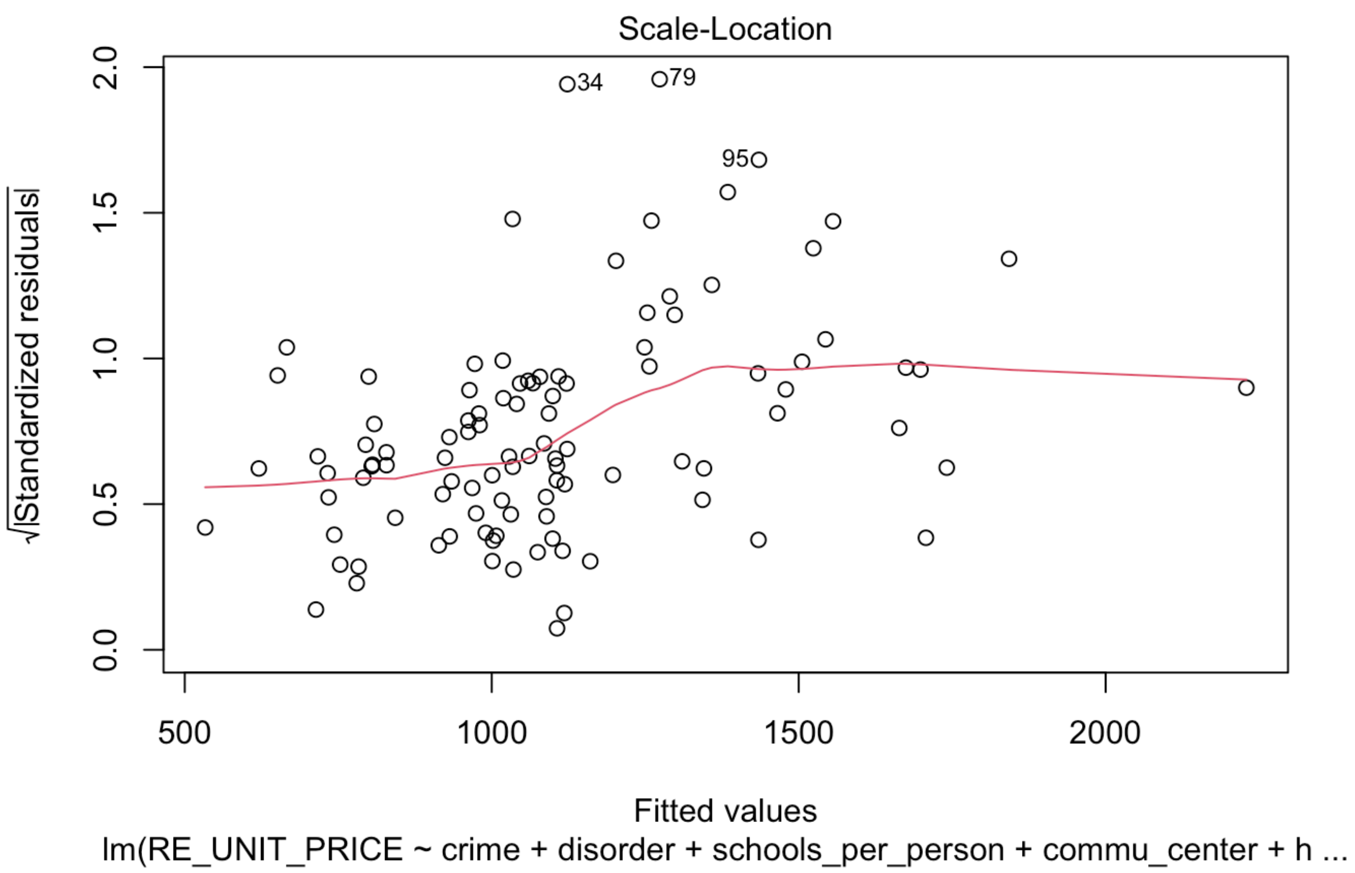
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
# Equal Variance
plot(model_final, which=1) #residuals plot
```



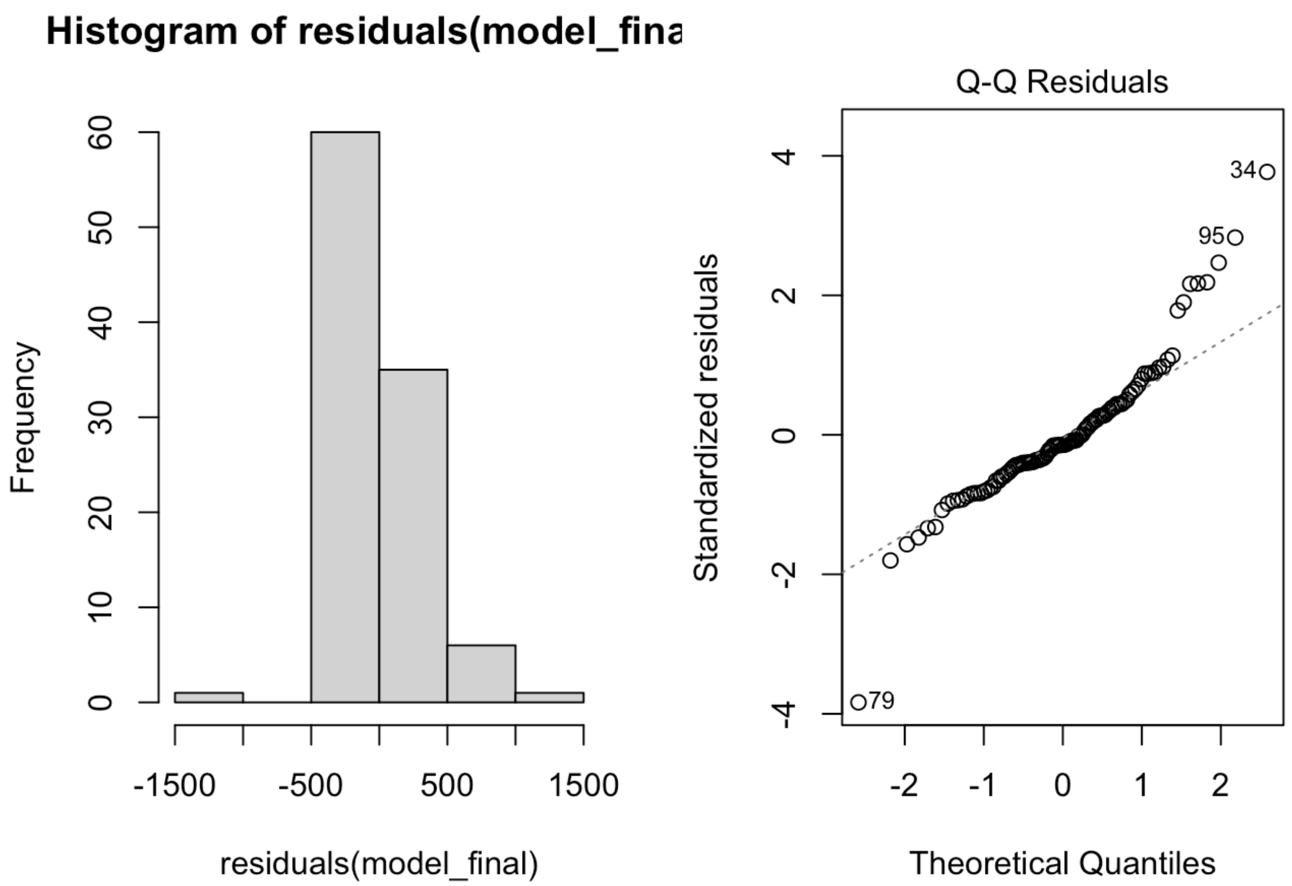

```
plot(model_final, which=3) #a scale location plot
```



```
bptest(model_final)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_final
## BP = 12.779, df = 7, p-value = 0.07769
```

```
# Normality
par(mfrow=c(1,2))
hist(residuals(model_final))
plot(model_final, which=2) #a Normal plot
```




```
#Testing for Normality
shapiro.test(residuals(model_final))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_final)
## W = 0.91616, p-value = 6.69e-06
```

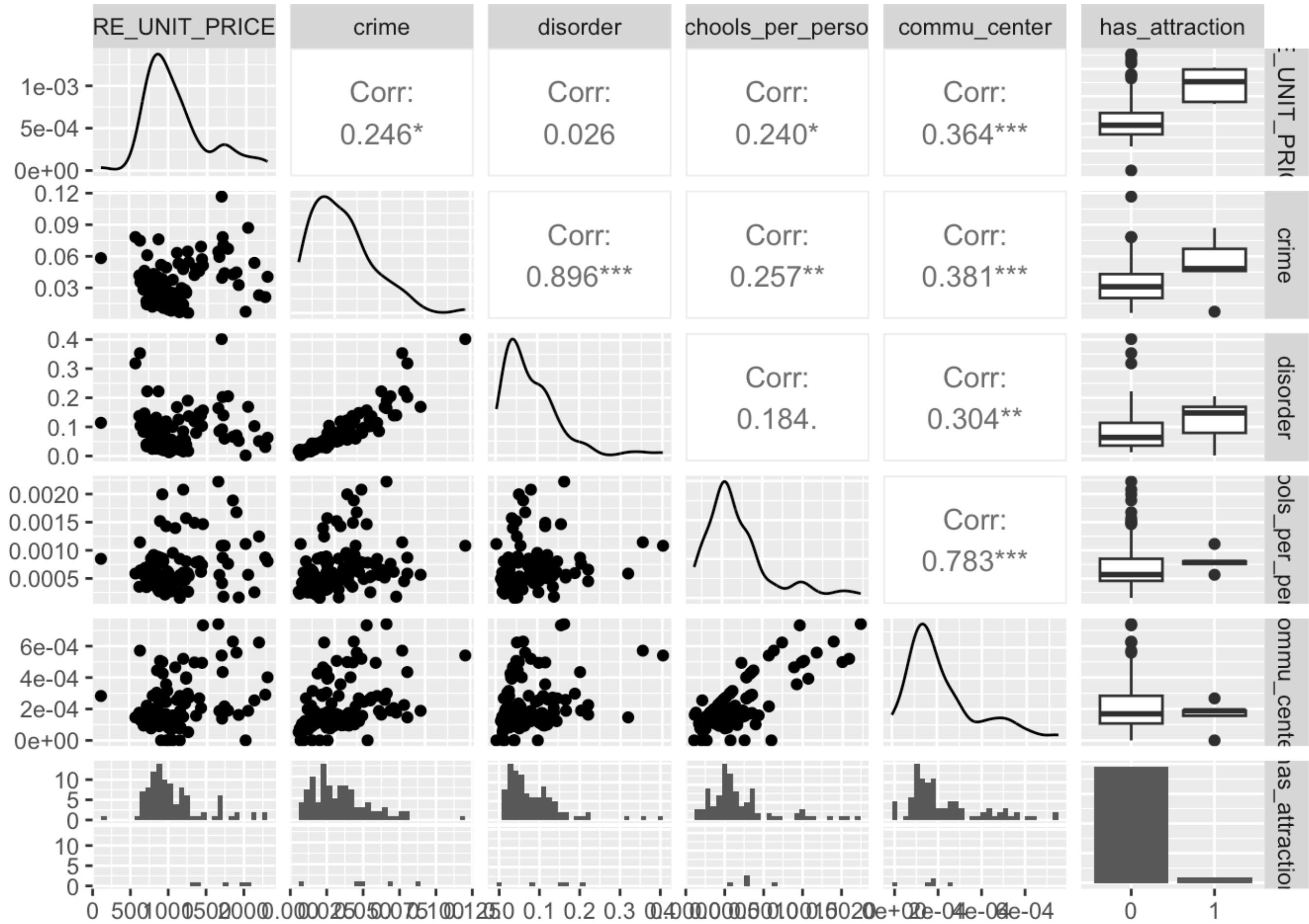
```
# Multilinearity
imcdiag(first_model, method="VIF")
```

```
##
## Call:
## imcdiag(mod = first_model, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                       VIF detection
## crime              5.6752         0
## disorder           5.1838         0
## schools_per_person 2.7269         0
## commu_center       3.0671         0
## has_attraction1    1.1282         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

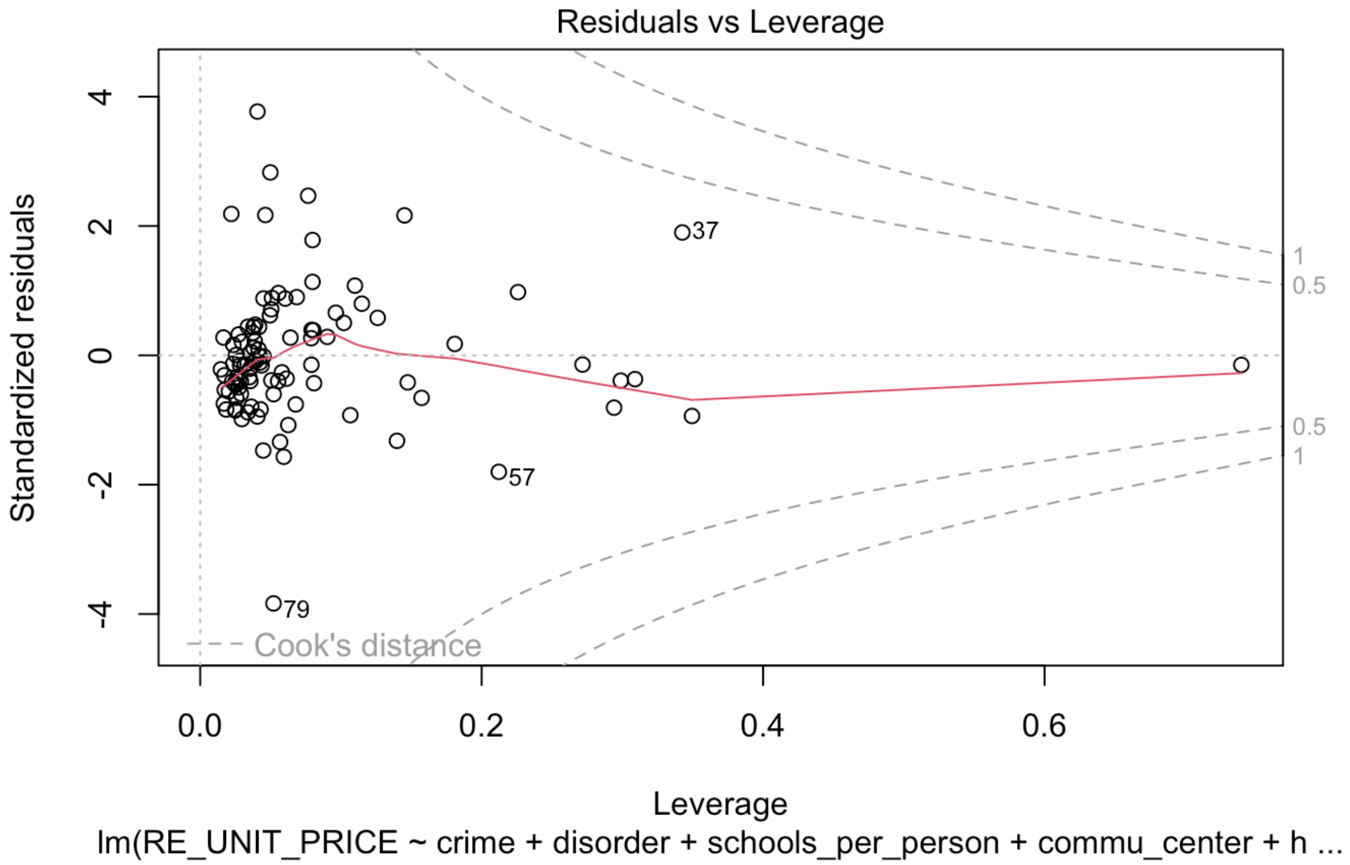
```
df = re_unit_price[,c("RE_UNIT_PRICE","crime", "disorder","schools_per
_person", "commu_center", "has_attraction")]

ggpairs(df)
```

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



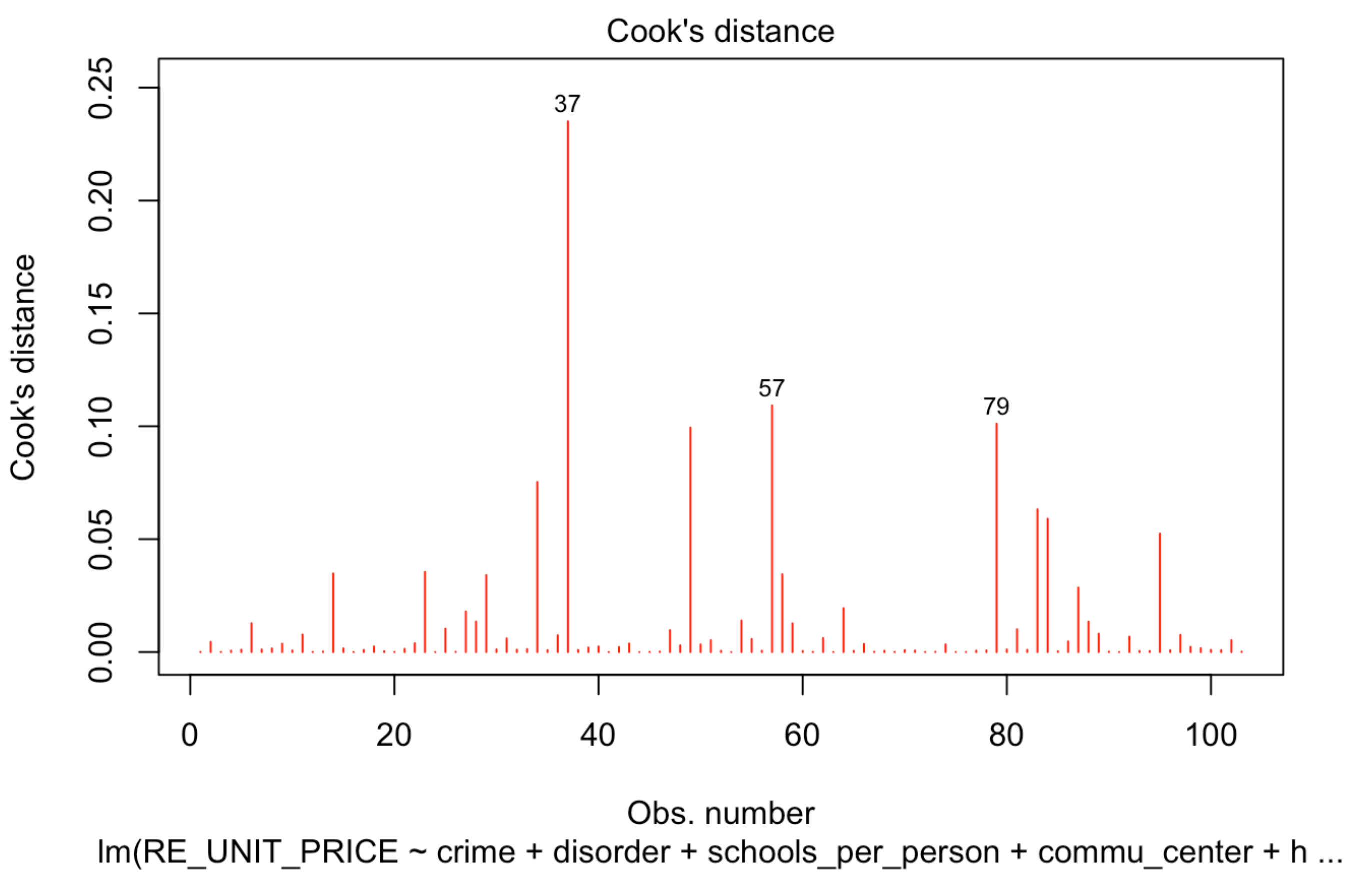

```
# Outliers
plot(model_final,which=5)
```




```
re_unit_price[cooks.distance(model_final)>0.5,]
```

```
##  [1] COMM_CODE                CLASS                    RE_UNIT_PRICE
##  [4] has_attraction           commu_center             commu_center_p
er_person
##  [7] has_hospital             has_library              has_phs_clinic
## [10] schools_per_person       has_social_dev_ctr       MALE
## [13] FEMALE                   English                  Eng_not_spk_of
t_home
## [16] Eng_ratio                Population               Top_language
## [19] Top_language_num         Top_language_per         Top_2_language
## [22] Top_2_language_num       Top_2_language_per       Top_3_language
## [25] Top_3_language_num       Top_3_language_per       crime_per_pers
on
## [28] disorder_per_person      has_social_ctr           crime
## [31] disorder
## <0 rows> (or 0-length row.names)
```

```
plot(model_final,pch=18,col="red",which=c(4))
```




```
lev=hatvalues(model_final)
p = length(coef(model_final))
n = nrow(re_unit_price)
outlier2p = lev[lev>(2*p/n)]
outlier3p = lev[lev>(3*p/n)]
print("h_I>2p/n, outliers are")
```

```
## [1] "h_I>2p/n, outliers are"
```

```
print(outlier2p)
```

```
##        11        14        29        35        37        38
57        81
## 0.7398493 0.2258141 0.2940827 0.1808550 0.3425908 0.2716626 0.21218
89 0.1573208
##        84        89        97
## 0.3494825 0.2988768 0.3089506
```

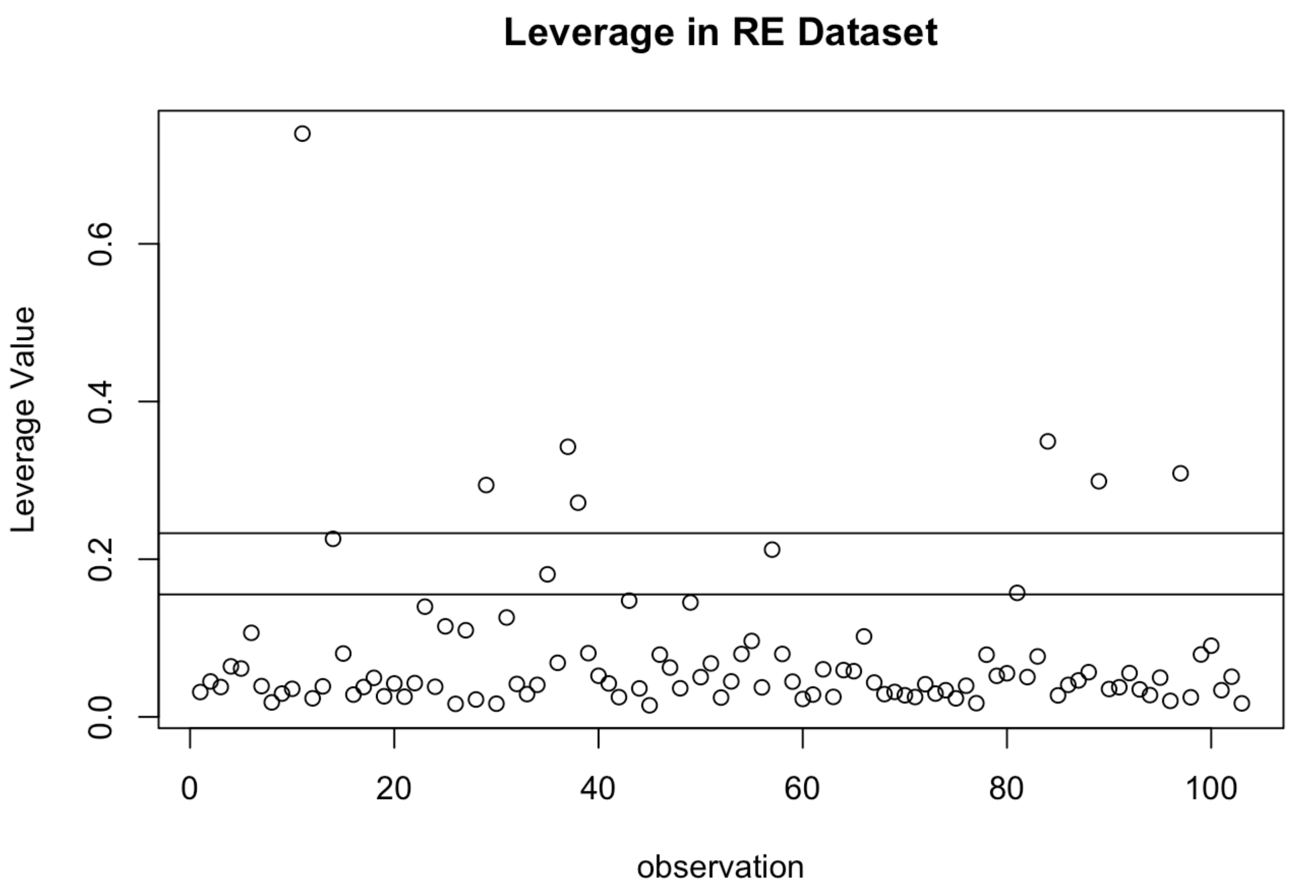
```
print("h_I>3p/n, outliers are")
```

```
## [1] "h_I>3p/n, outliers are"
```

```
print(outlier3p)
```

```
##        11        29        37        38        84        89
97
## 0.7398493 0.2940827 0.3425908 0.2716626 0.3494825 0.2988768 0.30895
06
```

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



remove outliers to refit the final model

```
newdata = re_unit_price[-c(11,29,37,38,84,89,97),]
newdata2 = re_unit_price[-c(11,14,29,35,37,38,57,81,84,89,97),]
model_final2 =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person + 
commu_center  + has_attraction + 
                  crime*disorder + I(commu_center^2), data=newdata)
summary(model_final2)
```

```
## 
## Call:
## lm(formula = RE_UNIT_PRICE ~ crime + disorder + schools_per_person 
+ 
##     commu_center + has_attraction + crime * disorder + I(commu_cent
er^2), 
##     data = newdata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1130.81  -146.83   -42.21   124.20  1165.43 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         1.095e+03  1.264e+02   8.662 2.04e-13 ***
## crime               1.691e+04  4.334e+03   3.902 0.000186 ***
## disorder           -7.414e+03  2.243e+03  -3.305 0.001375 ** 
## schools_per_person -3.966e+05  1.396e+05  -2.842 0.005575 ** 
## commu_center       -4.095e+05  8.099e+05  -0.506 0.614349    
## has_attraction1     5.375e+02  2.266e+02   2.372 0.019880 *  
## I(commu_center^2)   3.474e+09  1.406e+09   2.472 0.015378 *  
## crime:disorder      3.188e+04  2.918e+04   1.092 0.277597    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 304.7 on 88 degrees of freedom
## Multiple R-squared:  0.4401, Adjusted R-squared:  0.3956 
## F-statistic: 9.882 on 7 and 88 DF,  p-value: 4.851e-09
```

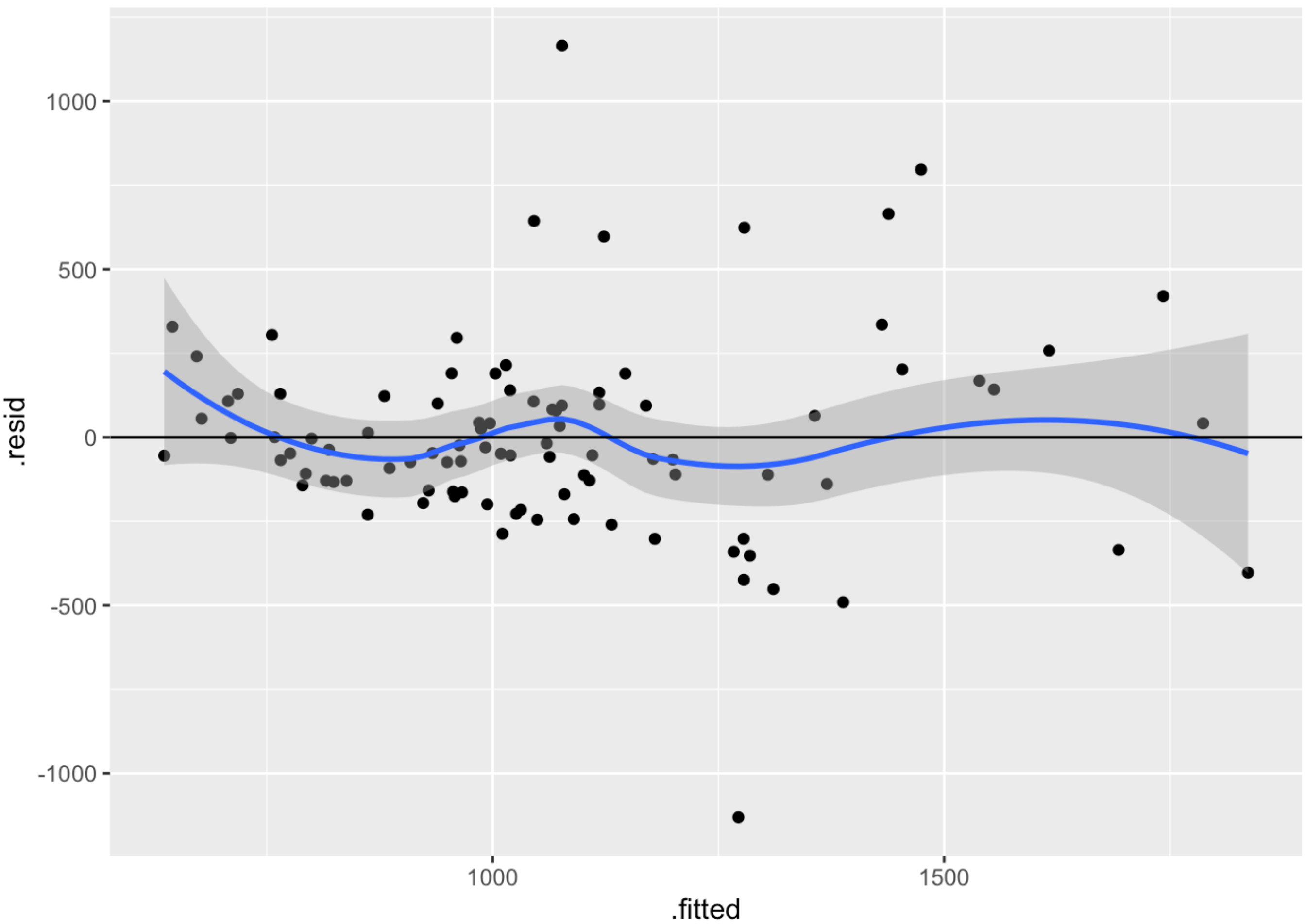
```
# codes to refit the model by removing outliers > 2p/n, but failed due 
to data avilability of attraction predictor

#model_final3 =lm(RE_UNIT_PRICE~crime + disorder + schools_per_person 
+ commu_center  + has_attraction + crime*disorder + I(commu_center^2), 
data=newdata2)
#summary(model_final3)
```

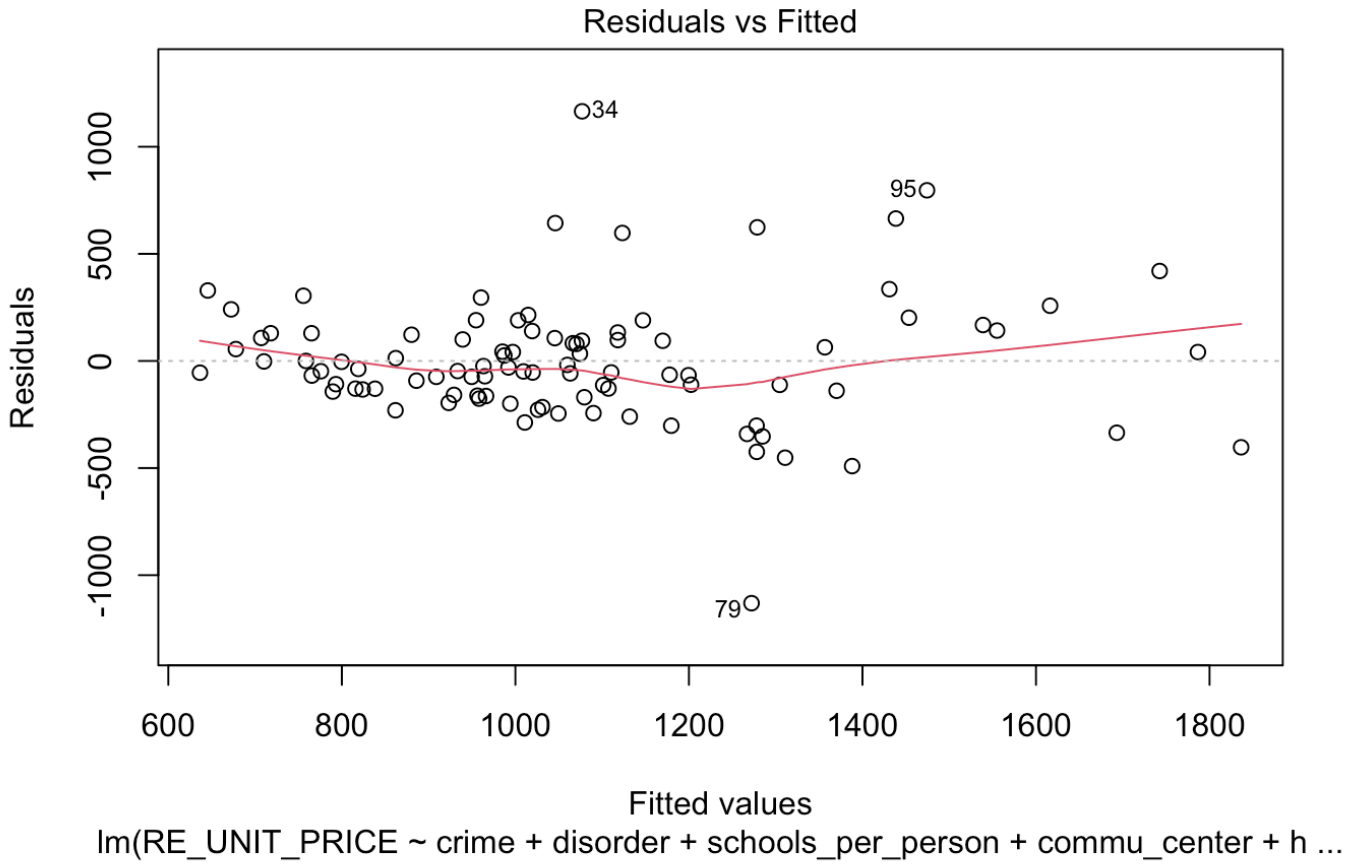
assumptions check for refitted model with outliers removed

```
# linearity
ggplot(model_final2, aes(x=.fitted, y=.resid)) +
  geom_point() +geom_smooth()+
  geom_hline(yintercept = 0)
```

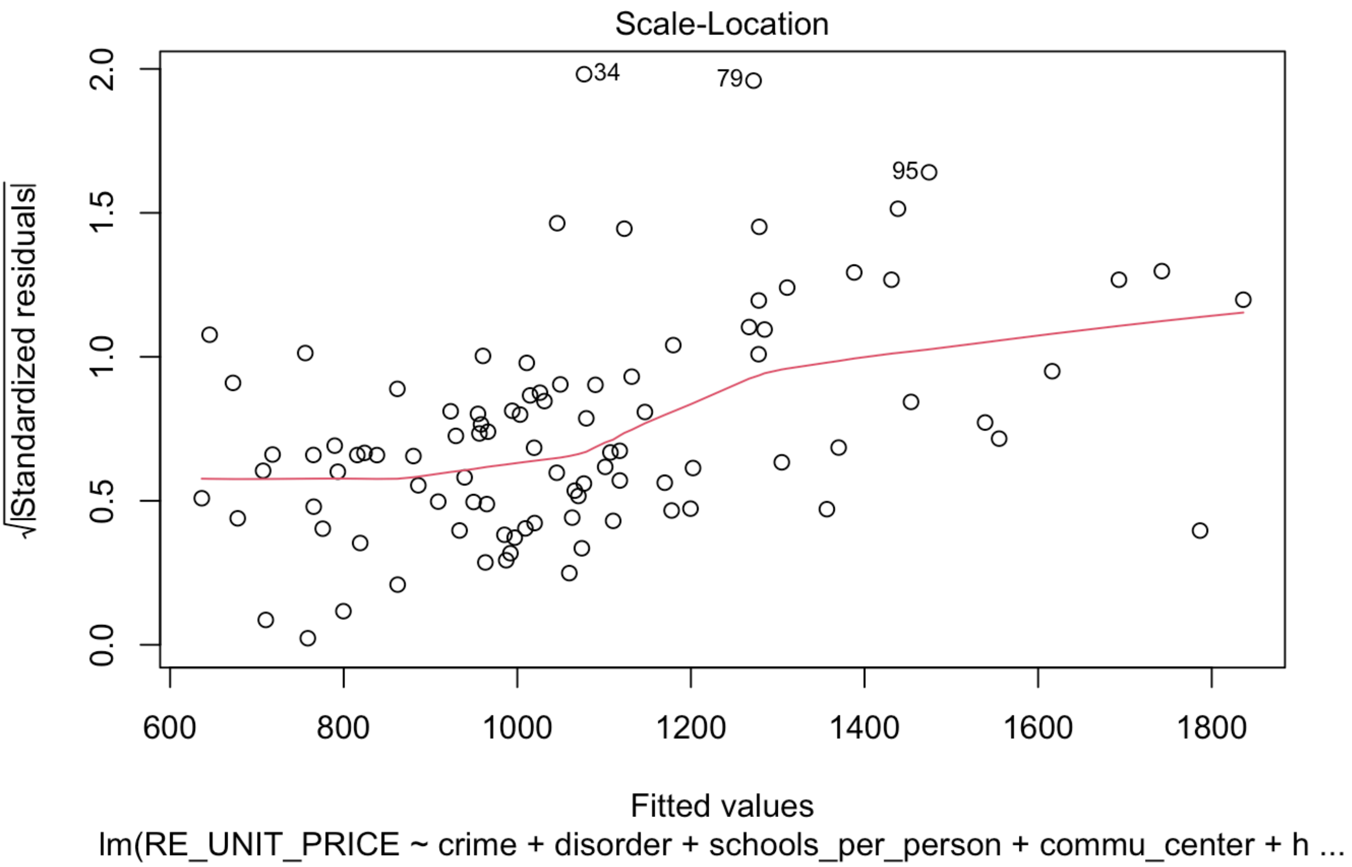
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
# Equal Variance
plot(model_final2, which=1) #residuals plot
```



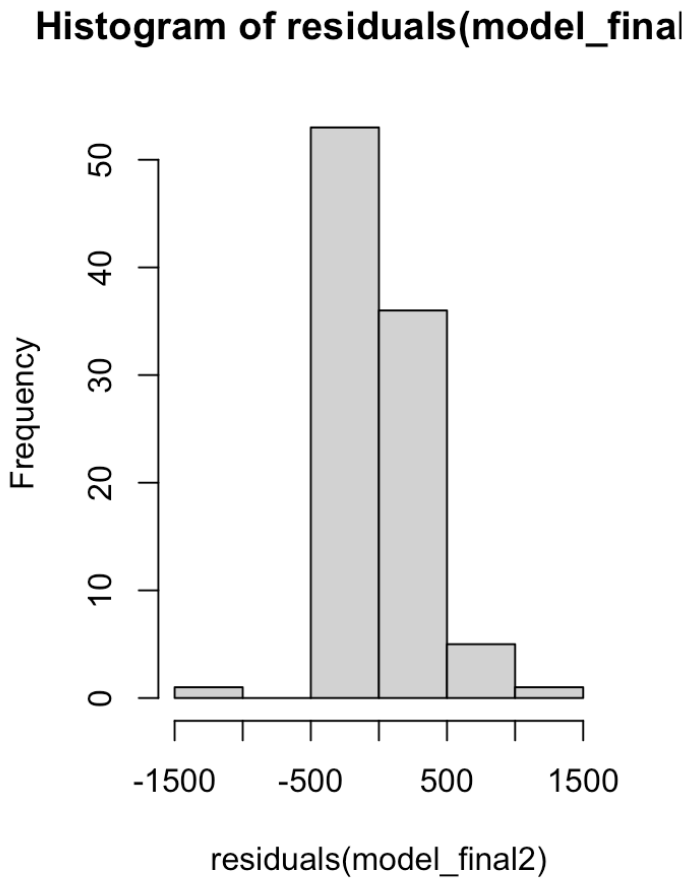

```
plot(model_final2, which=3) #a scale location plot
```

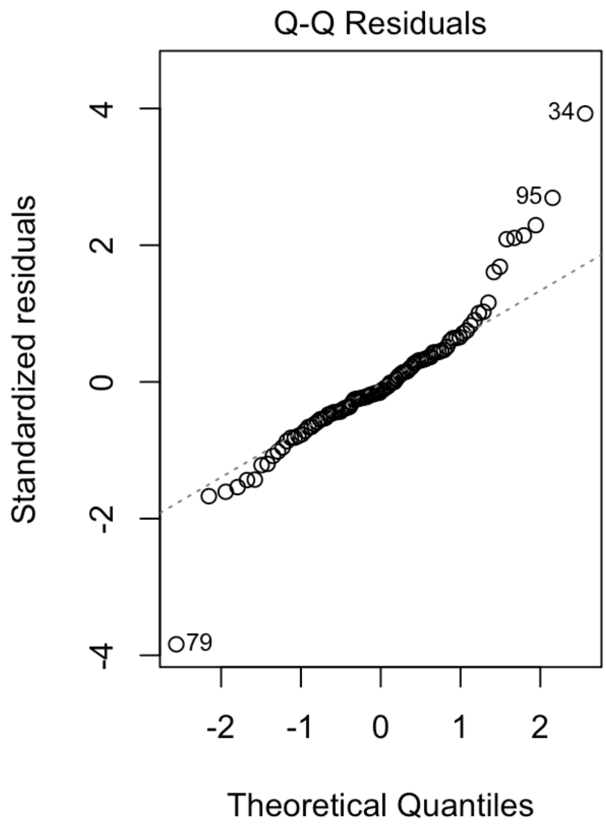


```
bptest(model_final2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_final2
## BP = 14.019, df = 7, p-value = 0.05084
```

```
# Normality
par(mfrow=c(1,2))
hist(residuals(model_final2))
plot(model_final2, which=2) #a Normal plot
```



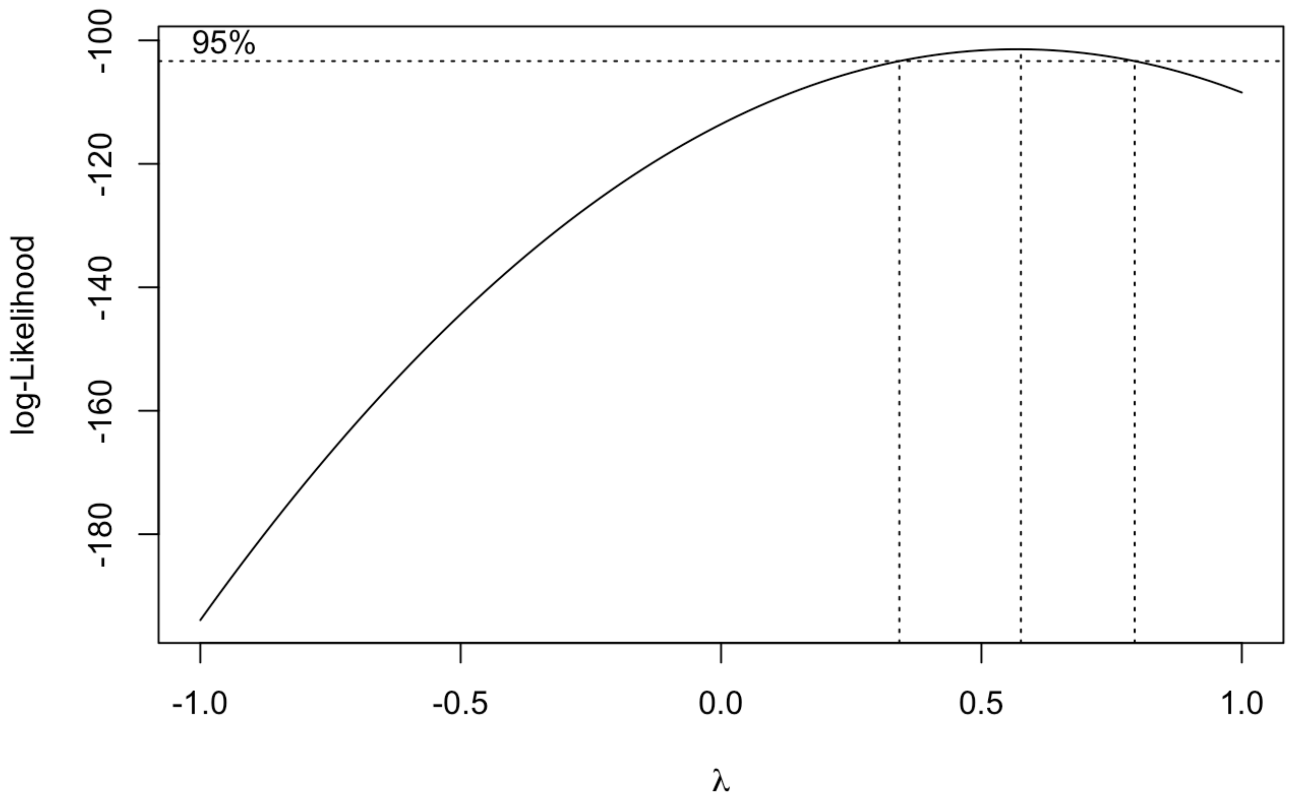



```
#Testing for Normality
shapiro.test(residuals(model_final2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_final2)
## W = 0.91716, p-value = 1.455e-05
```

box transformation to have normality

```
# transformation using original data
bc=boxcox(model_final,lambda=seq(-1,1))
```




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

```
## [1] 0.5757576
```

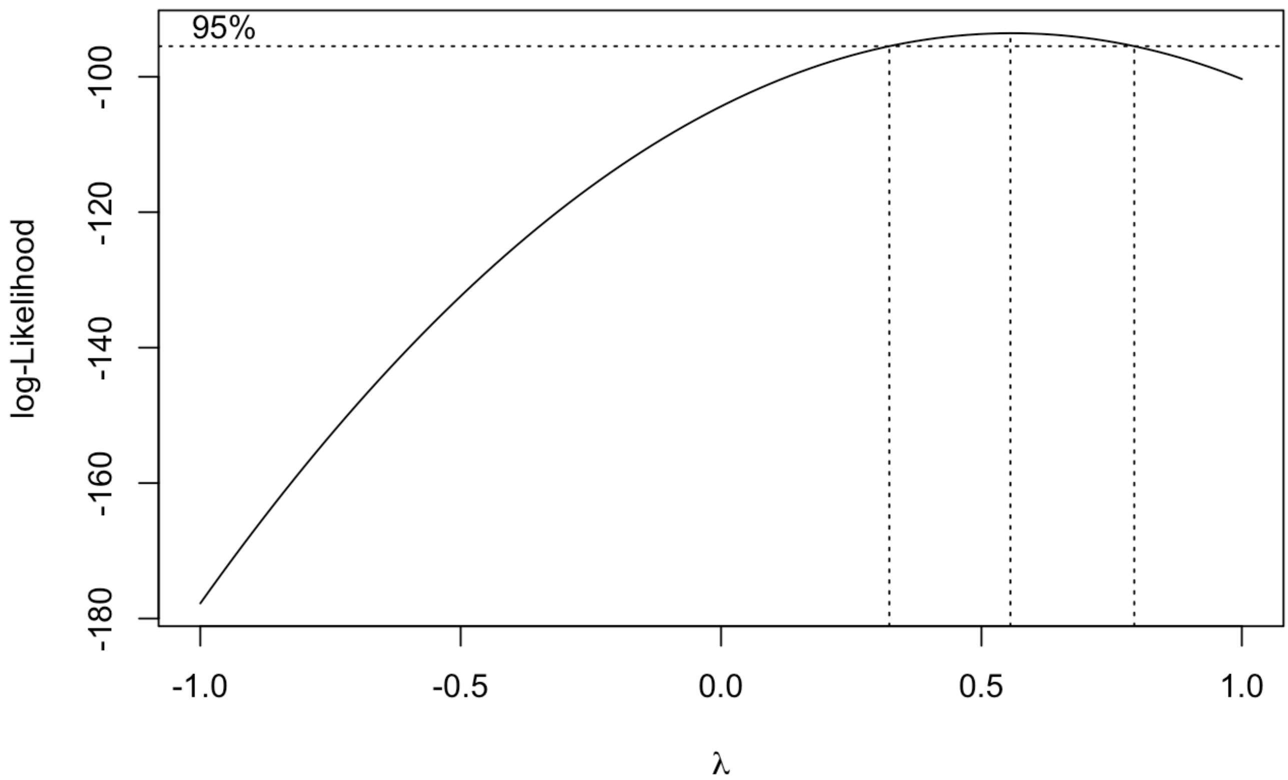
```r
bcmodel1=lm((((RE_UNIT_PRICE^0.5757)-1)/0.5757)~crime + disorder + sch
ools_per_person + commu_center  + has_attraction +
                  crime*disorder + I(commu_center^2), data=re_unit_pri
ce)
summary(bcmodel1)
```

```
##
## Call:
## lm(formula = (((RE_UNIT_PRICE^0.5757) - 1)/0.5757) ~ crime +
##     disorder + schools_per_person + commu_center + has_attraction +
##     crime * disorder + I(commu_center^2), data = re_unit_price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.926  -8.130  -1.565   6.799  49.815
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.502e+01  5.395e+00  17.612  < 2e-16 ***
## crime               8.002e+02  1.800e+02   4.446 2.37e-05 ***
## disorder           -4.083e+02  7.633e+01  -5.350 6.07e-07 ***
## schools_per_person -1.372e+04  6.267e+03  -2.189   0.0311 *
## commu_center       -9.666e+03  3.298e+04  -0.293   0.7701
## has_attraction1     3.161e+01  7.481e+00   4.225 5.49e-05 ***
## I(commu_center^2)   1.183e+08  4.941e+07   2.394   0.0186 *
## crime:disorder      1.813e+03  7.427e+02   2.440   0.0165 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.3 on 95 degrees of freedom
## Multiple R-squared:  0.479,  Adjusted R-squared:  0.4406
## F-statistic: 12.48 on 7 and 95 DF,  p-value: 3.055e-11
```

```r
shapiro.test(residuals(bcmodel1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bcmodel1)
## W = 0.89886, p-value = 9.181e-07
```

```r
bc=boxcox(model_final2,lambda=seq(-1,1))
```




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

```
## [1] 0.5555556
```

```
# transforamtion using data with outliers removed
bcmodel2=lm((((RE_UNIT_PRICE^0.5555)-1)/0.5555)~crime + disorder + sch
ools_per_person + commu_center  + has_attraction +
                  crime*disorder + I(commu_center^2), data=newdata)
summary(bcmodel2)
```

```
##
## Call:
## lm(formula = (((RE_UNIT_PRICE^0.5555) - 1)/0.5555) ~ crime +
##     disorder + schools_per_person + commu_center + has_attraction +
##     crime * disorder + I(commu_center^2), data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.949  -6.698  -1.013   6.265  44.809
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.636e+01  5.584e+00  15.465  < 2e-16 ***
## crime               6.831e+02  1.915e+02   3.567 0.000586 ***
## disorder           -3.278e+02  9.912e+01  -3.307 0.001368 **
## schools_per_person -1.636e+04  6.167e+03  -2.653 0.009466 **
## commu_center       -2.120e+04  3.578e+04  -0.592 0.555138
## has_attraction1     2.388e+01  1.001e+01   2.385 0.019235 *
## I(commu_center^2)   1.509e+08  6.211e+07   2.430 0.017131 *
## crime:disorder      1.554e+03  1.289e+03   1.205 0.231427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.46 on 88 degrees of freedom
## Multiple R-squared:  0.4152, Adjusted R-squared:  0.3687
## F-statistic: 8.925 on 7 and 88 DF,  p-value: 2.868e-08
```

```
#Testing for Normality
shapiro.test(residuals(bcmodel2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bcmodel2)
## W = 0.89227, p-value = 9.587e-07
```