Investigate the relationship between break_and_enter crimes and other variables with the Logistic Model

Baoying Xuan 1004808149

Dec 22, 2020

Abstract

This project aims to investigate the relationship between the break_and_enter crime rates and some other variables, including population, robbery crime rates, and homicide crime rates. For data analysis, the logistics regression model is used for seeking the relationship among variables. By conducting the logistic regression, the outcomes indicate that the homicide crime rates have little influence on the break_and_enter crime rates, while the other two variables, population and robbery crime rates, have a significant relationship with the break_and_enter crime rates.

Keywords

Neighborhoods Crime Rates, 2014-2019 Crime Data,2016 Census Population, Logistic model

Introduction

Crime rates in different areas across Toronto vary widely. This is particularly the case for the break and enter crime rates which is also a property crime. For example, break and enter crime rates of neighbourhoods in downtown Toronto are more than those in midtown Toronto, although these two areas are with a relatively similar population (Toronto Neighborhoods, 2020). What is the relationship between break and enter crime rates and other variables? How do those variables influence the break and enter crime rates?

For the purposes of this paper, 140 neighbourhoods were chosen for this paper. This data was collected by the Census Population in 2016 by Toronto Neighborhoods. The dataset is about the crime rate changes from 2014 to 2019 in different neighbourhoods. Assault, Auto Theft, Break-and-Enter, Robbery, Theft Over and Homicide are the available types of crime listed in this dataset, including the averages of these five-year data and crime rates per 100,000 people in different neighbourhoods.

This topic is important and interesting because crime rates are associated with everyone's health and safety who is living in Toronto or going to settle down in Toronto. Crime statistics can be a helpful tool for the police to predict the future risk of crimes and can also be used to prevent the predicted crimes in a specific area. If people want to choose a neighbourhood to live in, crime statistics will also be a good tool for them to make decisions.

Specifically, variations in the break and enter crime rates significantly affect how a community views property crime and the significance of this crime. One community may focus more attention on break and enter crimes while another may be more rigid in punishing prostitution. In attempting to explain variation in the break and enter crime rates, I am endeavoring to understand what the relationship between variables such as robbery crime rates and population will be, and how these variations influence the break and enter rates so that the property safety in this community can be determined. This determination of variations contributes significantly to the break and enter crime rate in any given neighbourhood, which in turn affects the stableness of that community. Neighbourhoods, which have a higher level of break and enter crime rates, are undeniably less attractive for people who are looking for a safe area to settle down in City Toronto. In a generation where people prior to living in a safe community, it is important to begin to delve deeper beneath the surface and attempt to determine the potential relationship between break and enter crime rate and other variables.

Data

There are five variables from the dataset were selected:

- “BreakandEnter_AVG” (Average break and enters from 2014-2019)
- “BreakandEnter_Rate_2019” (Rate of break and enters for 2019 per 100,000 population)
- “Population” (2016 Census Population)
- “Robbery_AVG” (Average robberies from 2014-2019)
- “Homicide_AVG” (Average homicides from 2014-2019)
- “Neighbourhood” (Name of City of Toronto neighbourhood)

Moreover, these variables are chosen because this report is mainly about what factors could be related to break_and_enter crime rates. The chosen factors are what normal people are interested in, and they are also the factors that come to mind for the first time. In this way, I can be more able to resonate with respondents when they do a survey and make people think more deeply.

In order to show a better view of listed variables’ raw data, making summaries and scatter plots for them will be clear and helpful.

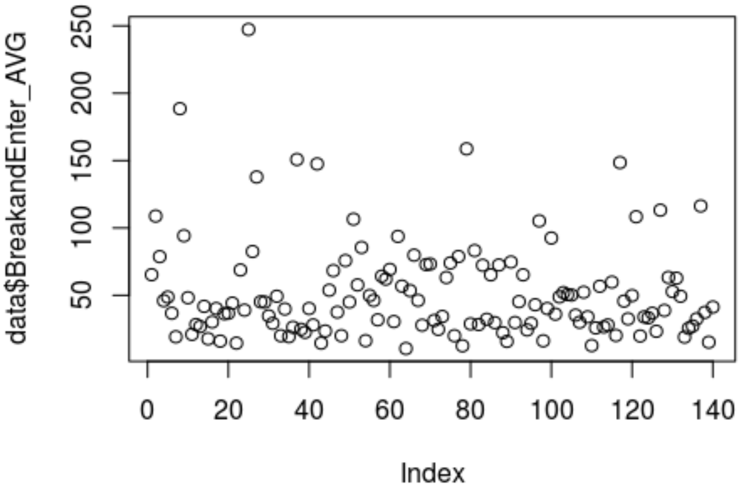
Make a summary of all variable's raw data to make a choice of which topic should be chosen as the topic of this report.

Types of crime rates	Min	1st Qu.	Median	Mean	3rd.Qu	Max
BreakandEnter_AVG	10.5	28	40.75	51.55	64.45	247.3
Robbery_AVG	3.3	11.68	20.1	25.65	30.4	135.7
Homicide_AVG	0	0.2	0.3	0.5136	0.725	2.5
AutoTheft_AVG	2.7	13.28	18.8	27.84	30.98	366.7
Theftover_AVG	1.2	3.5	5.2	8.083	8.35	56.2

It is clear to see that the average of BreankandEnter_AVG is the highest, which indicates that the break and enter crime issues are getting more and more serious in the Toronto community. Therefore, this topic would be getting more and more attention from the society as well. Thus, investigating the break_and_enter crime rates and see how this crime rate affects society will be the best idea based on this dataset.

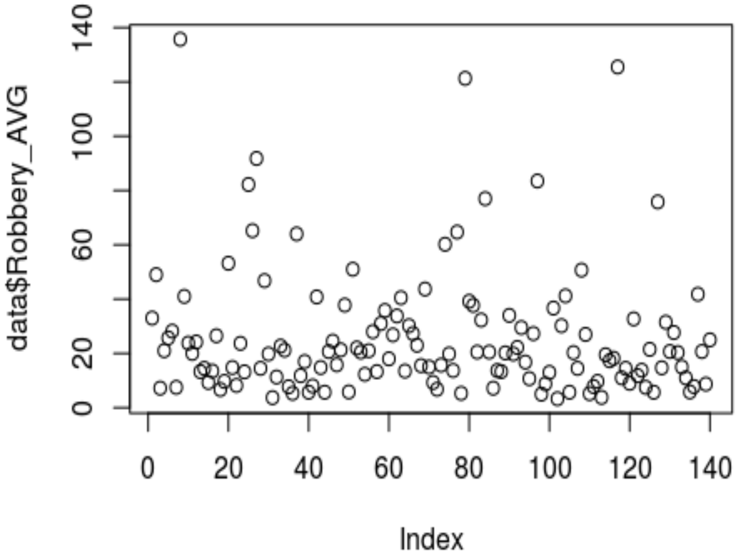
On top of that, I still scatter plot the raw data to show a better view of all chosen listed variables:

1) plot raw data of BreakandEnter_AVG:



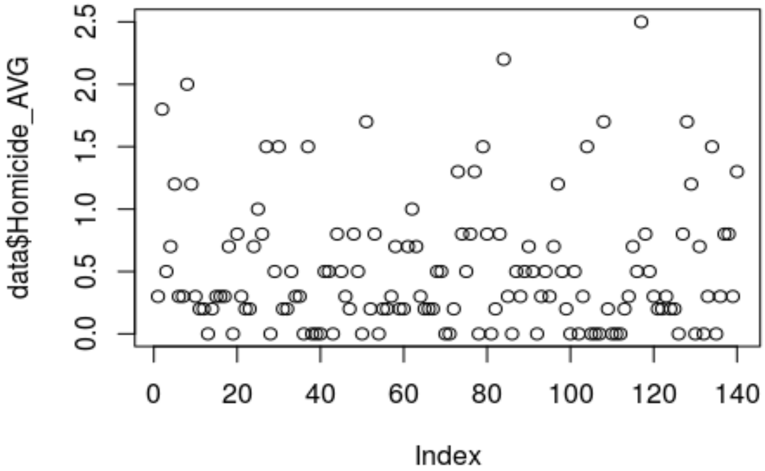
It is not hard to notice that most of the average break_and_enter crime from 2014 to 2019 are below the level of 100.

2) Plot raw data of Robbery_AVG:



It is not hard to notice that most of the average Robbery crime rates located below the level of 40.

3) Plot raw data of Homicide_AVG:



In terms of the Homicide average, most of the points located from level 0 to level 1. It means the average of homicides from 2014 to 2019 is from 0 to 1.

Overall, the data of Neighborhoods Crime Rates were collected from the open data from Toronto Police Services, and the number of investigated objects was 140 which is less than the expected number of target sample size.

The data consists of core data, which is related to the crime rates of break_and_enter rate, and classification data, such as population, neighbourhoods, geographic area. The target population of the Neighbourhood Crime Rates investigation is all the Canadian residents who were living in Toronto. The Toronto Police Service did simple random sampling and selected 140 neighbourhoods which can best represent all the neighbourhoods in Toronto. In 2016, Toronto Public Services collect data based on the Census population survey in 2016. If some respondents refused and ignored to take the Census survey and the assistants from Toronto police would explain the importance and encourage them to take the survey. Thus, the population data was mostly precious enough to be used as an independent variable. While there still exists some non-response problem, because there are some vagrants or refugees who are Canadian, but they might not take the Census population survey. This issue would affect the results of this topic.

For the survey, the crime rates data will be uploaded via the police system based on the previous cases. If some neighbourhoods wanted to cover the previous criminal cases and refused to announce a true amount of criminal cases due to their reputation maintenance, the data will be double-checked and recorrect by the Toronto Police Services. Hence, the cost of creating this data would be wages payment. For this dataset, it did not cover enough crime rate variables. It is not sufficient to analyze the relationship with variables from a different angle. For example, only a few variables such as populations, neighbourhoods and robbery crime rates can be used to analyze the relationship with break_and_enter rate, which would decrease the correlation between different variables in the model, and at least one more variable should be added. Moreover, some binary variables should be added to this dataset.

In conclusion, the target population of the data covered the most eligible person in Toronto and the sample size is mostly sufficient, such that the data is useful to analyze the relationship between crime rates and other variables in Toronto. The data is cleared but lacks more details about the crime rate variables.

Model

There are four variables selected to fit the logistic model:

- “BreakandEnter_AVG” (Average break and enters from 2014-2019)
- “Population” (2016 Census Population)
- “Robbery_AVG” (Average robberies from 2014-2019)
- “Homicide_AVG” (Average homicides from 2014-2019)

Since most of the variables in the dataset are categorical and want to explore the relationship between the break_and_enter crime rates and other variables. Thus, a logistic regression model would be chosen since the break_and_enter rate level is a binary variable and other variables are either categorical variables or numeric variables. Furthermore, logistic regression fits the dataset more than the other regression.

If the linear regression is used to fit the binary variable, the resulting model might not restrict the break_and_enter crime average level within 0 (Low break_and_enter crime rates level) and 1(High break_and_enter crime rates level). Besides, other assumptions of linear regression such as normality of error may get violated. Hence, if I use the logistic regression model, it is easier for us to understand the result than using other models.

The reason why the break_and_enter average level was chosen to be used as a dependent variable is that this variable is mutated as a new cell which divides the BreakandEnter_AVG into two levels, level 1 means the break_and_enter average rate is high, and level 0 means the break_and_enter average rate is low.

The reason why the Homicide_AVG and Robbery_AVG were chosen to be independent variables rather than other crime rates' average is that Robbery and break_and_enter crimes belong to property crimes. Thus, these two crimes may have a positive correlation. Also, potentially, when criminals commit any break_and_enter crime offence, they will commit homicide or kill the eyewitness. That is why I wonder if there is any relationship between break_and_enter crimes, homicide crimes and robbery crimes.

According to the Urban Scaling Theory, the number of crimes committed may follow a linear relationship as a function of the population size of the city. For example, if the population size increases by 100%, the incidence of crime may increase by 120%. Therefore, I assume the population variable has the potentials to influence the break_and_enter crimes rates (Population Size vs. Number of Crime - Is the Relationship Super-Linear, 2020)

The notation of the logistic regression model is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{Population} + \beta_2 x_{Robbery-AVG} + \beta_3 x_{Homicide-AVG} + \epsilon$$

p is the probability of the neighbouhoods expeirence high level of break_and_enter crimes. (1-p) is the probability of the neighborhoods experience low level of break_and_enter crimes. $\frac{p}{1-p}$ is the ratio of neighborhoods experience high level of break_and_enter crimes compared to neighborhoods experience low level of break_and_enter crimes. Beta0 is the intercept of the model which represents all the variables equal to zero, then the value of the log odds. beta1 coefficient represents change in log odds for every one unit of population increase in the total population. beta2 coefficient represents the average change in log odds for every one-unit increase in the number of robbery crimes. beta3 coefficient represents the average change in log odds for every one-unit increase in the number of homicide crimes.

Here is the summary of the logistic model:

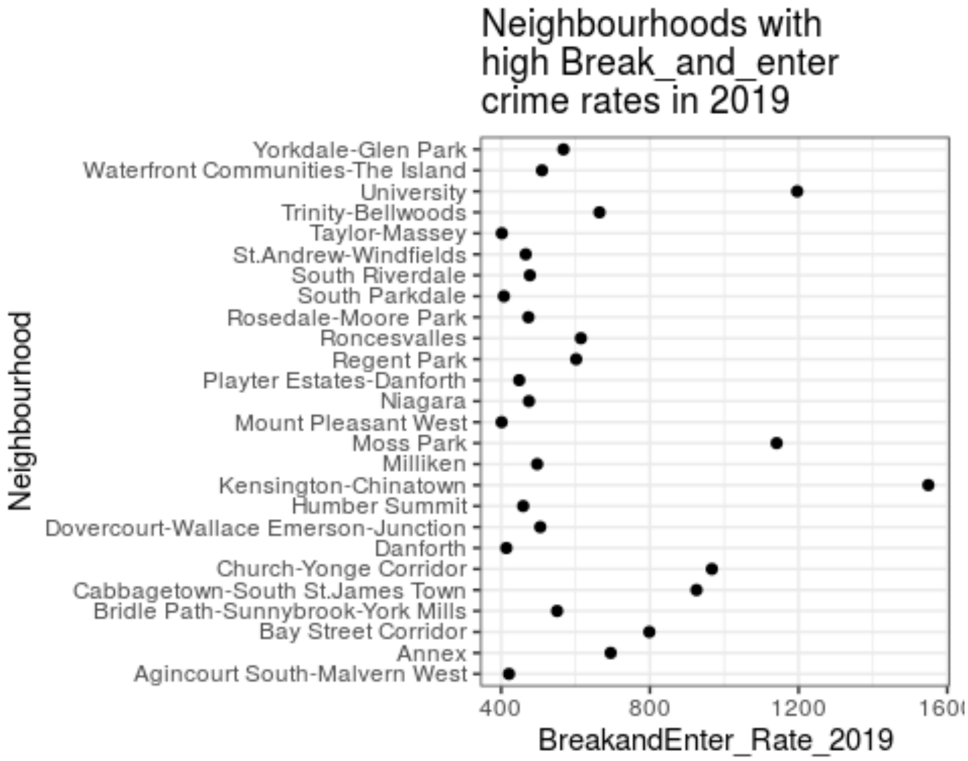
Coefficient	Estimate	Std. Error	Z-value	P-value
(Intercept)	-5.73E+00	9.25E-01	-6.194	5.86e-10 ***
Population	2.12E-04	4.66E-05	4.545 5	5.48e-06 ***
Robbery_AVG	5.37E-02	2.39E-02	2.245	0.0247 *
Homicide_AVG	-9.52E-01	7.44E-01	-1.279	0.2008

Clearly, this table shows the coefficients of each variable in the logistic model. More details will be discussed in the Discussion part.

Results

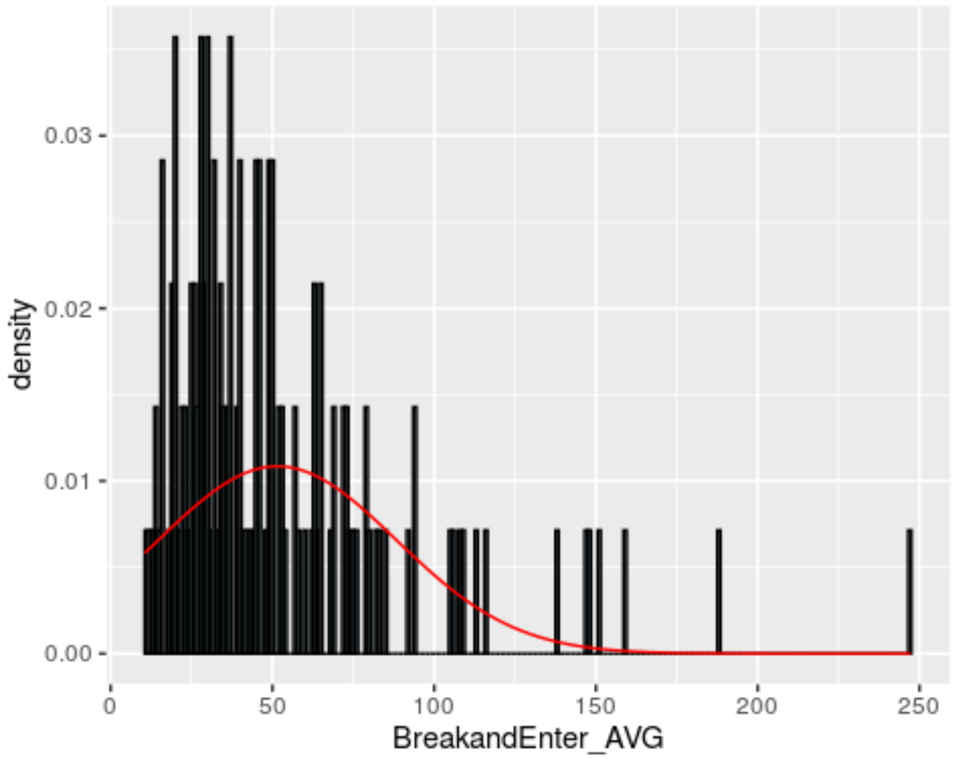
Here are some findings and results.

1) Investigate the relationship between neighbourhood and high Break_and_enter crime rates in 2019:



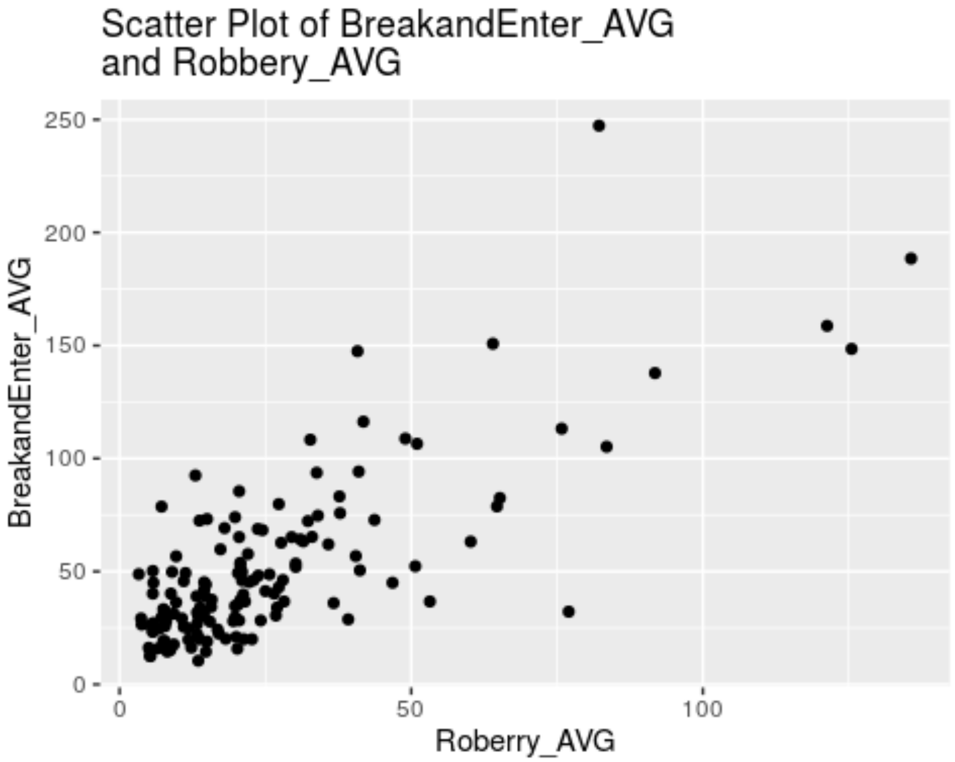
In this scatterplot, there is only one quantitative variable which is the break-and-enter rate in 2019, so there is no correlation relationship that can be discussed. Therefore, I can discuss some significant points. In the neighborhoods with high break-and-enter crime rates in 2019: The Kensington Chinatown neighborhood has the highest break-and-enter rate in 2019, which indicates that Chinatown is a dangerous place which occurred so many Break_and_enter crime in 2019. The Mount Pleasant West neighborhood has the lowest break-and-enter rate in 2019, which indicates that the place will be the safest among all the neighborhoods in the chart. And most of the points have fluctuated around the level of 400 to 800 except a few of them have extremely high crime rates.

2) Investigate the density of BreakandEnter_AVG:



As shown in this figure, the histogram of the average of break_and_enter crimes is skewed to the right (i.e. positively skewed), with the right tail being much longer and the mass of the distribution concentrating on the left of the figure, with more than 80% of the total neighbourhoods experienced break_and_enter crimes lower than 150 times from 2014 to 2019. And the rest of them have experienced an extremely a high level of break_and_enter crimes from 2014 to 2019.

3) Investigate the relationship between BreakandEnter_AVG and Robbery_AVG



In order to visualize the possible correlation between break_and_enter average crimes and robbery average crimes, a scatter plot was created. In this scatter plot, there is a point for every neighbouhood in Toronto from 2014 to 2019. As shown in this figure, there is some correlation between the two types of crimes. A neighbourhood with a higher level of robbery crimes tends to also have a higher level of break_and_enter crimes. Nevertheless, the correlation is not very strong with some points far from the majority of the points, indicating neighbourhood where one rate fails to predict the other rate.

Before I performed a basic cleaning and summary analysis of the data, I assumed that the larger the population size is, the higher the crime rate is in the neighborhood with a large population size.

However, I filter the population size and only keep the neighborhood whose population size is bigger than 40000. And also, I subset the data to include the neighborhood whose population size is bigger than 40000 and retain the corresponded columns OBJECTID, and BreakandEnter_Rate_2019.As we can see, the population of Willowdale East (50434) is bigger than L'Amoreaux (43993) but the break-and-enter crime rate of Willowdale East (174.5) is smaller than that of L'Amoreaux (195.5).

In conclusion, it turns out that the crime rates will not always increase when the population size increases, it all depends on the neighborhood environment.

4) Summarize the break_and_enter rate from 2014 to 2019:

Year	Mean	Standard deviation
2019	305.5507	216.7786
2018	53.97857	45.59652
2017	49.15714	37.71748
2016	45.7	31.42118
2015	49.28571	33.12485
2014	51.27857	32.18093

Before I performed a basic cleaning and summary analysis of the data, I assumed that the break-and-enter crime rate will decrease with the growth of the city.

However, I summarize the average and standard deviation of the break-and-enter crime rate from 2014 to 2019 and compare every year's data. As we can see, the average break-and-enter crime rate in 2019 is the highest over these five years, jumping from 51.27857 in 2014 to 305.5507 in 2019. Standard deviation follows the same trend.

In conclusion, it turns out that the break-and-enter crime rate will not always decrease with the growth of the city, it all depends on the environment of the city.

Discussion

[Interpretations -Logistic Regression Model]

In this part, I am going to interpret the logistic model data by examining the eviance Residuals and coefficients table. Recall the summary table of the coefficients.

Coefficient	Estimate	Std. Error	Z-value	P-value
(Intercept)	-5.728E+00	9.25E-01	-6.194	5.86e-10 ***
Population	2.12E-04	4.66E-05	4.545 5	5.48e-06 ***
Robbery_AVG	5.37E-02	2.39E-02	2.245	0.0247 *
Homicide_AVG	-9.52E-01	7.44E-01	-1.279	0.2008

First of all, as given by the summary of the logistic model, I notice the Deviance Residuals look good since they are close to being centred on '0' and roughly symmetrical. Secondly, I interpret the coefficients table by examining the coefficients table, which shows the estimated regression beta and the associated t-statistic p-values.

(a). Betas

[beta_0_hat is -5.728]: it shows that the estimated intercept is -5.728.

[beta1_hat is 2.120*10^ (-4) (i.e. the coefficient of “Population”)]: meaning that holding all other predictor variables – Robbery_AVG and Homicide_AVG unchanged, the increase of one unit in population, on average, will lead to the level of Break_and_enter crimes decrease by 2.12010^(-4).

[beta2_hat is 5.371*10^ (-2) (i.e. the coefficient of “Robbery_AVG”)]: meaning that holding all other predictor variables - Population and Homicide_AVG unchanged, the increase of robbery crimes by one unit, on average, will lead to the Break_and_enter crimes increase by 5.371*10^(-2).

[beta3_hat is -9.522*10^ (-1) (i.e. the coefficient of “Homicide_AVG”)]: meaning that holding all other predictor variables - Population and Robbery_AVG unchanged, the increase of Homicide_AVG by one unit, on average, will lead to Break_and_enter crimes increase by -9.52210^ (-1).

(b). Standard Errors

Each estimate of the regression contains relatively small uncertainty and is considered to be good estimates since all of the corresponding standard errors of the betas are relatively small. (c). t-statistic & p-values Assuming the significance level equal to 0.05, I set the hypothesis here for each coefficient is shown below: [Null Hypothesis H0]: the corresponding predictor has no correlation to the dependent variable (i.e. “Break_and_enter crimes”). [Alternative Hypothesis Ha]: the corresponding predictor does have a correlation to “Break_and_enter crimes”. As given by the summary table, the intercept and population p-values are well-below 0.05 and thus, these two variables are significantly related to the dependent variable, “Break_and_enter crimes”. Also, there are *** signs right next to all p-values, which also proves that they are all statistically significant. In addition, the “Robbery_AVG” is significantly correlated to the dependent variable.

(d). AIC

AIC is just the Residual Deviance adjusted for the number of parameters in the model.

e) Findings

What the world will know after reading this report:

Despite what I summarized above, there are also some fun facts found from the research:

1. The crime rates will not always increase when the population size increases, it all depends on the neighborhood environment.

2. The break-and-enter crime rate will not always decrease as the growth of the city; it all depends on the environment of the city. 3. Although certain researchers believe that robbery crimes rates will be higher when the neighbourhood is experiencing a high level of break_and_enter crimes. However, they have a positive but weak relationship.

Weakness

In general, there are a variety of weaknesses and limitations existing in this investigation. More specific details and thoughts are expressed below: Data Limitations

The data was based on the Census population in 2016, so the data set is not precious enough to predict the crime rates in the future. As known, people will move out or move in from 2016. A dataset based on the current year's population should be used next time. Moreover, this dataset only recorded 140 objects which are too small to represent the whole of Toronto city. The precision of the logistics summary results lowered due to the limited categorical data.

Model Weaknesses

In terms of the model weaknesses, the model is mainly based on past results. Generally, situations are changing every year, even every moment.

One example that the model result may not apply for now is the coronavirus pandemic, one of the worlds' greatest challenges happening in 2020. There are a variety of articles and reports mentioned that the pandemic would significantly increase the crime rates as it makes society more unstable as a whole. Therefore, the correlation between the break_and_enter crime rates and variables like population may not be what it is supposed to be in 2020.

In addition, due to the limited objects and variables, it is hard to apply other methods like post-stratification, which narrowed the way to find out the relationship between break_and_enter crime rates and other variables. Hence, the result may be not precious enough.

Communication Failings

From the communication perspective, my research and investigation are merely in the theoretical stage, where the real-world situations are more complex and might be different from what I found. Therefore, it is required to practically do more field studies and surveys to avoid communication failings and get a better understanding of my investigation.

Next step

As known, my research and investigation mainly focus on what variables would impact the break_and_enter crime rates, and I choose only several variables like "population", "robbery crime rates" and so on. However, these variables are mainly potentially internal factors, and it is believed that adding more external factors would be beneficial.

For example, city growth would be the one that could impact the crime rates of break and enter. More specifically, when the economy is booming, the break and enter crime rates would be relatively lower since people's income is higher, making them economically independent. Thus, people don’t need to rely on stealing and commit crimes much to maintain good living standards, and the crime rates would generally decrease.

Moreover, it is necessary to upgrade the ways of collecting the data to ensure that the researcher could get the data as precise as possible but also save time and money. Since the amount of data after filtering out the missing values is only about 140 neighbourhoods, which is too small to reflect the real situation of Toronto. Thus, it is also important to collect more data that could cover all areas to represent the entire population that I am going to investigate.

All in all, social problems of family investigation such as the break and enter crime rates is inextricably related to many other external factors such as the law, economic conditions and so on. Thus, there is actually much to desire in its completion method. For now, I have just created and implemented a simple direction of investigating the relationship between the break and enter crime rates and several variables like population, robbery crime rates, etc. and I am looking forward to others exploring more and improving it.

Reference

Barrett, Tyson S. R For Researchers: An Introduction. 13 May 2019, tysonbarrett.com/Rstats/chapter-2-working-with-and-cleaning-your-data.html.

Contributors, Data Carpentry. Aggregating and Analyzing Data with Dplyr.Karl Broman, 1 June 2016, kbroman.org/datacarpentry_R_2016-06-01/03-dplyr.html.

Chang, Y. S. (2013, August 2). Population Size vs. Number of Crime - Is the Relationship Super-Linear? by Yu Sang Chang, SungSup Choi, Jinsoo Lee, Won Chang Jin :: SSRN. Population Size vs. Number of Crime - Is the Relationship Super-Linear? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2305136

Toronto Neighbourhoods. (2020). DigiMarCon Canada 2022 - Digital Marketing Conference & Exhibition. https://digimarconcanada.ca/toronto-neighbourhoods/

Wickham, Hadley. Dbplyr. 17 June 2019, www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/summarise.