Victoria University of Wellington

DATA301
Final Project

Data Analysis of New Zealand Police Data

Author:
Luke Hartfield

October 16, 2020

Contents

1 Executive Summary

For this project, I analysed the Victimisations Time and Place dataset from Police New Zealand which contains information on criminal offenses in New Zealand, and more specifically, the number of victims resulting from those offenses. I combined this dataset with the 2018 New Zealand Census dataset which provides many population measures. For my response variable, I choose to use the continuous variable, victimisation rate, which is the number of victims as a per capita percentage for that territory. I then choose to explore the relationship victimisation rate has with three independent variables, mean unemployment rate, mean age, and mean income.

The exploratory analysis showed a fairly weak positive linear relationship between victimisation rate and mean unemployment rate. It also shows a moderate negative linear relationship between victimisation rate and mean age. As for mean income, there did not appear any obvious relationship.

The initial model involving all three independent variables and their interactions did not meet the assumptions of linear regression, to solve this mean income was excluded from the model and all remaining explanatory variables were log-transformed. The log transformation slightly reduced the spread of the data points.

From this model, there is evidence that suggests is a relationship between mean age and victimisation rate (p-value = 0.05). According the the model, keeping all other explanatory variables constant, and with 95% confidence, if log(mean age) was increased by 1, victimisation rate would decrease by 19.25 and 0.04 on average. There was no evidence of a relationship between mean unemployment rate and victimisation rate (p-value = 0.18). The interaction term of mean age and mean unemployment rate was found not found to be significant (p-value = 0.15).

As mean unemployment and the interaction term were found to not be significant, another model which only contains mean age was produced. This model shows evidence that there is a significant relationship between mean age and victimisation rate (p-value <0.01). According to this model, keeping all other explanatory variables constant, and with 95% confidence, if log(mean age) was increased by 1, victimisation rate would decrease by between 12.57 and 4.78 on average.

2 Background

2.1 About The Data

The primary dataset used in this project is the Victimisation Time and Place dataset provided by New Zealand Police [1]. This publicly available dataset contains information on where and when crimes occurred in New Zealand. In addition to this, we also acquired the 2018 New Zealand Census data, supplied by Statistics New Zealand [2]. This dataset has an extensive number of population measures from all across New Zealand. Finally, the Geographic Areas dataset also provided by Statistics New Zealand will also be required to link the two other datasets together [3].

The Victimisation Time and Place dataset contains a breakdown of all criminal incidents that occurred in New Zealand between July 2014 and May 2020. Each entry includes the type of offense, where it occurred, the time of occurrence in terms of the month, year, day, and hour, and a number of other potentially useful aspects. This makes the dataset of significant interest

for analysis as it can be used to determine valuable patterns in crime. It is hoped that this information can be used to better understand crime and provide insight into when and where criminal acts are most likely to take place. This knowledge could help reduce New Zealand's crime rate by more effectively targeting areas of known higher criminal activity through better allocation of government funding and police resources.

While the Police dataset details the when and where of criminal offenses in New Zealand, further analysis into the possible 'why?' is also desired. This is why the 2018 New Zealand Census data is being combined with the Police data. The census has numerous measures regarding the country's population such as ones relating to income, employment, education, and health. These features can be compared against the crime data to see if any possible relationships can be found. An example of this could be to see if areas with higher levels of unemployment also see higher levels of crime.

The Geographic Areas dataset provides information on two different measures of area in New Zealand that are used in the data collection process. The first measurement being meshblock and the second being statistical area, the latter measurement being a grouping of the former. In order for the Police and Census data to be integrated with each other they need to have an attribute in common. In this case, a unit of area is used to link the datasets. While the census uses 2018 statistical areas, the Police data uses 2013 Census meshblocks and therefore an intermediate file that contains meshblocks and their corresponding statistical area is needed to join the datasets. The dataset also provides information on region names, constituencies, and community names relating to a particular meshblock/statistical area.

2.2 Objectives

The objective of this analysis is to answer the following question: "Is it possible use unemployment rate, mean age, and median personal income to predict victimisation rates in New Zealand?". Using this data, statistical models will be produced that are able to accurately predict the rate of victimisation in a given area when provided with new data. It is hoped that these models could be used in the future to predict victimisation rates in areas within New Zealand. These models will also be able to tell us how much unemployment, age, and personal income influence victimisation rates in New Zealand.

3 Data Description

The 2018 New Zealand Census dataset is in wide format with 483 variables and 32,521 observations, all collected during the 2018 census period. The attributes are made up of a mix of integer, character, and factor data types. Each observation provides information for that particular statistical area in New Zealand. The data provider, Statistics NZ, chooses to replace any confidential data points with 'C' as per their 'Threshold Rule'. If a variable has been deemed as sensitive, then all observations for that variable with a count less 6 are replaced with 'C' [4]. They are therefore treated as missing values. When looking to see if these values shared anything in common, many of the variables with the highest number of censored values were related to occupations, the industry people worked in, and also people's mode of transport for getting to work and education. Due to this, the dataset consists of 1,706,161 missing data points out of 14,001,482 total data points, giving the dataset a missing value percentage of

12.2%. These will therefore be excluded. The data has been preprocessed by Stats NZ so no errors should be present.

The Victimisation Time and Place dataset is also in wide format with 17 variables and 1,213,977 observations. The variables are made up of both factor and integer data types. The data uses Statistics New Zealand's 2013 Census meshblock values.There are just 2 missing values, both occurring in the hour of day variable. The dataset does have some features that should be noted. Firstly, the attribute 'area unit' which consists of the name of the suburb where the offense took place, has some unexpected values with 119 data points listed as '-29' (<0.1% of total). Area unit also has 4678 entries (<1% of total) listed as '999999' which Police NZ has used to indicate unknown areas [5]. Following on from this, the area units with those values also have unexpected meshblock codes, 4797 of the meshblock entries (<1% of total) are listed as negative values which do not correspond to a valid meshblock. Next, the location type which listed the type of location where the occurrence took place (e.g. private dwelling, warehouse/storage) has 835,400 (68.8% of total) listed as '.' which likely represents an unspecified location type. It should also be noted that 11076 instances (<1% of total) in the location type are listed as 'Unknown Location". For the day of the week attribute, 2 entries are listed as '.' and 383,362 (31.6% of total) entries are listed as 'UNKNOWN". The last unexpected value noticed is present in the 'hour of day' attribute that lists the hour that the crime occurred. There are 639,900 instances (52.7% of total) where the hour of day is listed as '99', which, according to Police NZ, denotes an hour that cannot be determined accurately [5]. Most of these variables are not being used in this analysis so they do not need to be treated. The only one of these variables being used is the meshblock code and due to there being such a small number of abnormal meshblock codes (<1% of total), they will be excluded.

4 Ethics, Privacy, and Security

4.1 Ethical Considerations

O'Neil, in Weapons of Math Destruction (2016), talks about how three characteristics of mathematical models can make them dangerous [6]. The same reasoning can be applied to any use of data, or any research. These characteristics are opacity, scale, and damage.

Opacity is about inscrutability of process - a lack of clarity about things like underlying assumptions, for instance. Research should be subject to review and critique to check that it is robust. Opacity limits opportunities for errors, oversights, unintended consequences, and bias to be detected and addressed and increases potential for misuse (deliberately or otherwise) of results. For this project, we are not constructing complex mathematical models. The exploratory data analysis is primarily composed of simple calculations, tables and plots. Data cleaning processes and decisions have been discussed among the group, and records of decisions made. All analysis has been documented and annotated.

Scale is about the level of influence your work may have. In this case, the potential for influence is very low. The project is not intended to inform policy or other decision makers or stakeholders. Neither is it likely to be published in any form at any time. The audience for this work will probably not extend beyond the group members and the person marking assignments.

Damage is about working against the interests of those affected. This might be through dis-

crimination. For instance, predictive policing has been found to amplify existing bias and exacerbate discrimination against already margin- alised groups [7]. Even though we are using police crime records and attempting to apply some kind of predictive approach, it is unlikely that this project would do damage for several key reasons. The first reason is the lack of influence already mentioned. The project is an opportunity to experience a range of tools and processes associated with data-science-in-practice and collaborative work, rather than inform the body of knowledge about crime in New Zealand. In any case, we did still consider and discuss ethical issues associated with the data and its use. The national operating model for New Zealand Police suggests that the police are already aware of the existence and location of areas of high (recorded) crime, but that they want to better understand the time and place relationships [8]. We discussed how considering the influence of systemic factors (rather than characteristics of individuals or communities) that might be associated with high crime rates could be a positive and constructive approach that might limit the potential for stigma or targeting of individuals and groups. If the scale of this work was different (with more potential for influence), it would be important to put much greater consideration into the potential impacts on already heavily-policed groups, and also the right of Māori, under the Treaty of Waitangi, to data sovereignty. Another point to note is that for the Police data, neither the offenders nor the victims were required to give consent to the use of this data as it has been heavily anonymized, though for this reason, the data requires more care and no attempts to de-anonymise it should ever take place.

4.2 Privacy

The data that we are using is publicly available official administrative data that has been de-identified (anonymised). In the case of census data, small counts of data deemed to be sensitive are suppressed in the available data set. Data is also not being combined in such a way that might allow for re-identification.

4.3 Security

Again, because all the data we are using is publicly available, there are no obvious security concerns. Depending on the analyses being carried out, and the potential for harm from results, the security of those results could be important. However, as already discussed, the work we are doing doesn't present a high risk of damage. While the integrity and availability of the data are unlikely to be at risk from adversaries (our work is not a likely target), it is still possible for integrity and availability to be compromised due to mistakes or technical problems. In order to minimise those risks, we are using cloud-based collaborative tools, including git for version control. While the git repository is set to public, this shouldn't be a privacy or security issue as the data is all publicly available, it is worth noting though.

5 Exploratory Data Analysis

5.1 Victimisation, Income and Unemployment

This section compares mean income, unemployment rate, and the rate of victimisation per head of population per territorial authority. While the number of victims per offense per territory has been summed and then adjusted to be a percentage of the territory's population, it is

possible that one person could be the victim of multiple different offenses and therefore doesn't may not show the percentage of the individuals in the area that have fallen victim, but rather the rate of victimisations per capita. In New Zealand there are 67 territorial authorities. These consist of 12 city councils, 53 districts, Auckland Council, and Chatham Islands Council [9].

Firstly, Figure 1 shows the rate of victimisations per capita for each territory. We can see that the percentage varies quite significantly between the different districts. The 5 territories that see the highest rate of victimisation per capita are: Napier, Hamilton, Christchurch, Rotorua, and Thames-Coromandel.

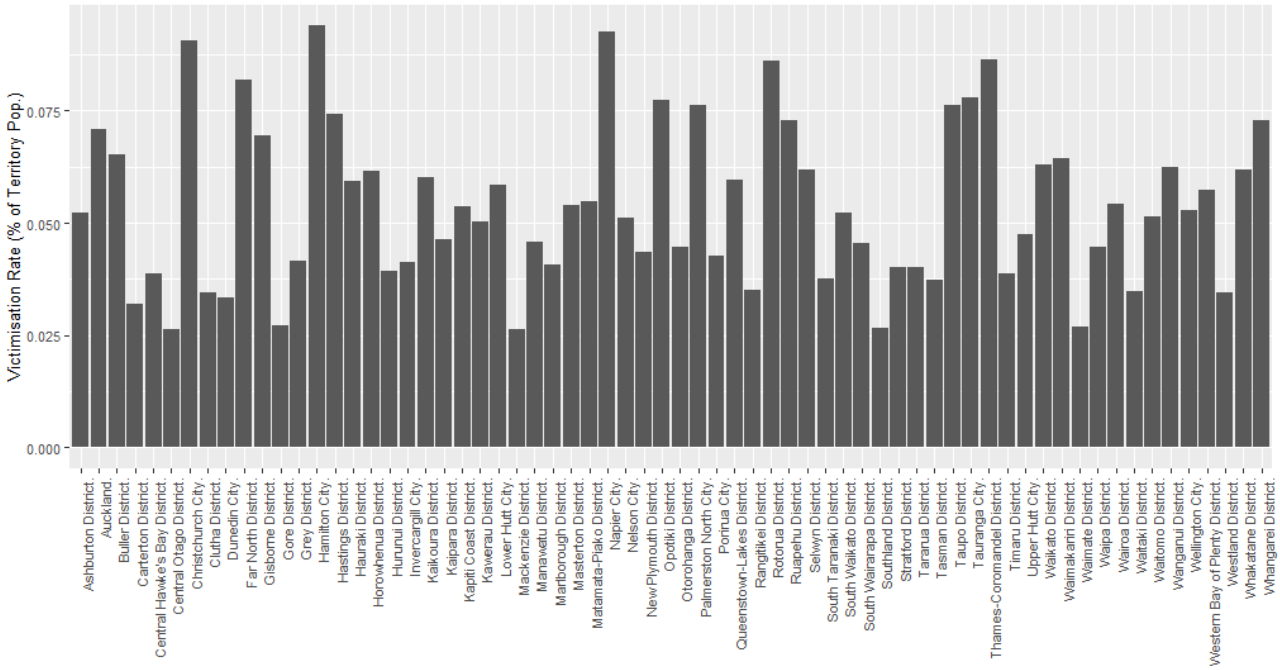

Figure 1: Victimisation percentage by territorial authority

Next, Figure 2 shows the mean personal income (in $NZD) of each territorial authority. Mean personal income was calculated by totaling all sources of personal income and then dividing that by the corresponding territory's population. Interestingly, We can see that mean income is reasonably even between the territories, more so than expected prior to analysis as it was expected that territories with large infrastructure or high tourism levels to have considerably higher mean income compared to others that do not. We can also see that Queenstown-Lakes, Selwyn, and Wellington have slightly higher incomes than the other territories. Comparing these territories' mean personal income to their victimisation rate per capita, they do appear to be reasonably low. When plotting mean personal income against victimisation rate per capita, as seen in Figure 3, there doesn't appear to be any strong linear relationship between the two.

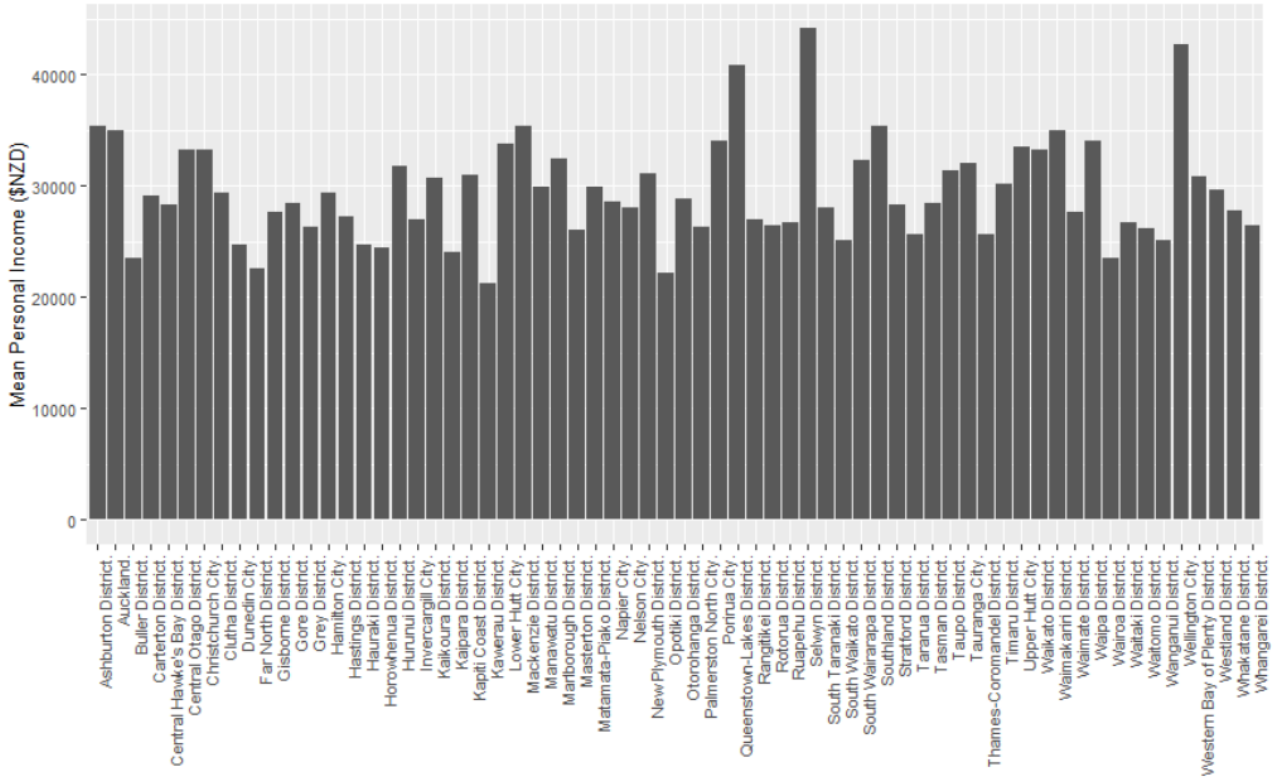

Figure 2: Mean personal income per territorial authority

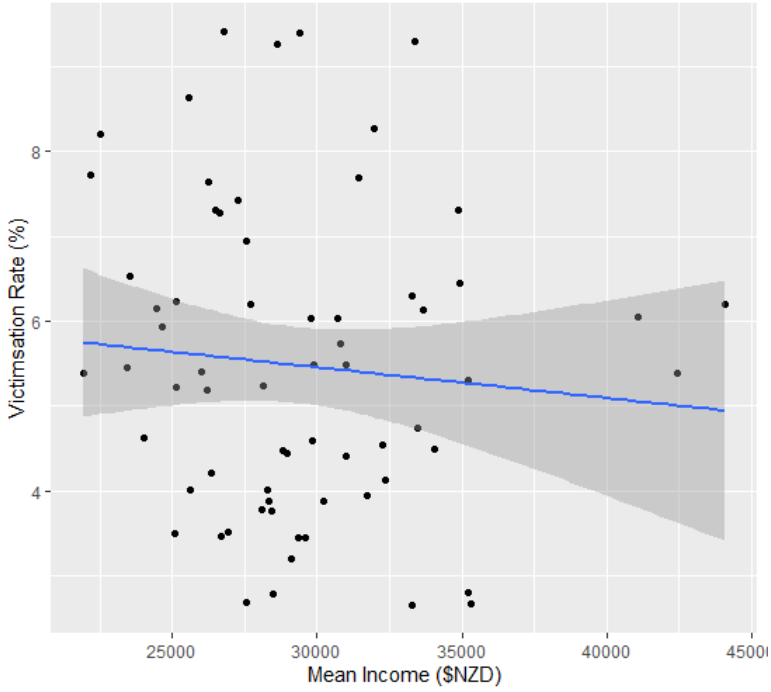

Figure 3: Victimisation rate vs. mean personal income

6 Detailed Analysis Results

6.1 Outline

For this detailed analysis, I have chosen to create a multiple linear regression model. This will be done at the territorial authority level. The model will use the continuous numerical variable, victimisation rate, as the response variable. As for the explanatory variables I'm using the continuous numerical variables: unemployment rate, mean age, and median personal income. Victimisation rate has been calculated by taking the number of victims in a territorial authority and then dividing it by the population of that territory to give us a percentage of that areas population that has been victimised. Unemployment rate is the number of people unemployed expressed as a percentage of all individuals of the working-age population (age 15 and over) who are either employed or unemployed. Stats NZ defines unemployment as being without paid work, where a person was available for and actively seeking work [10].

6.2 Data Preprocessing

Before the models can be created, some preprocessing needs to be done to the data. First, the police dataset is filtered to only include offenses that take place in 2018, this is so that it over the same period at the 2018 census data. Next, the police dataset and the census dataset are merged together using the geographical area dataset as an intermediary dataset. From there, the dataset is filtered to only include SA2018 code, territorial authority, number of victimisations, unemployment count, mean age, median personal income, and population count. The missing values in the dataset are then imputed using the mean value for that variable. The population per territory is calculated by summing all the population counts per SA code for that territory. Next, the total victimisations, mean personal income, mean age, and mean unemployment rate is calculated per territory. Victimisation rate per territory is then added by dividing the number of victims by that territory's population. This results in a dataset with 66 observations.

6.3 Results

The relationships between the three explanatory variables and the response can be illustrated in figure 4. Figure 4a shows a strong positive linear relationship between mean unemployment rate and victimisation rate. Figure 4b shows no relationship between mean personal income and victimisation rate. Figure 4c shows a strong negative linear relationship between mean age and victimisation rate. To try and improve normality, the data is then log transformed. Figure 5 shows the new relationships. We can see that this has indeed slightly improved the relationships so the log transformed variables will be used in the model. As mean personal income has shown to not have any relationship, it will be excluded from the model.

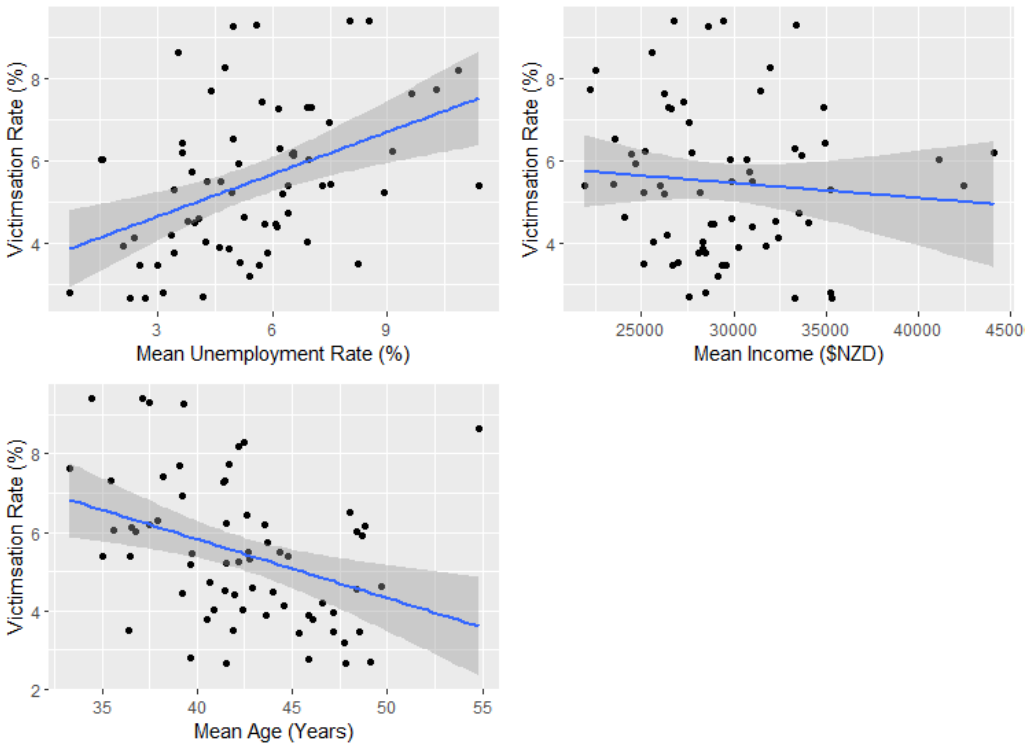

Figure 4: (a) Mean unemployment rate vs. Victimisation rate, (b) Mean personal income vs. Victimisation rate, (c) Mean age vs. Victimisation rate

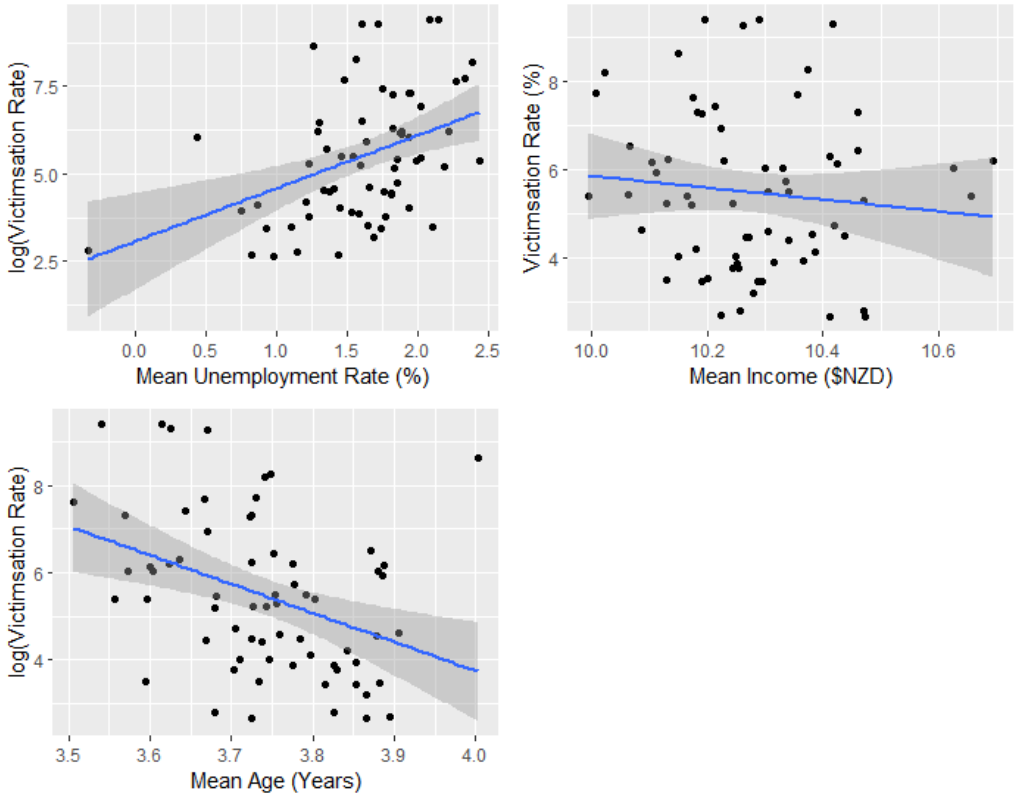

Figure 5: Explanatory variables against response variable after log transformation

A multiple linear regression model is produced with the log transformed mean age and mean unemployment rate, as well as the interaction between them, as the predictors. Figure 6 shows

the diagnostic plots for this model and indicates that the linear regression assumptions have been met.

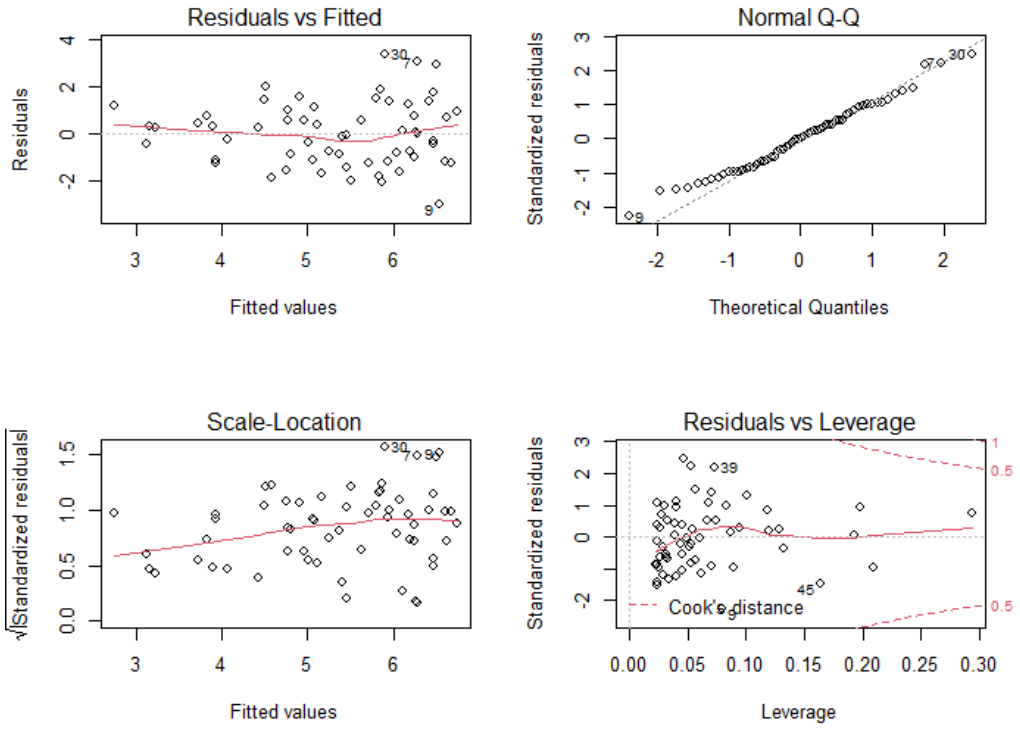

Figure 6: Diagnostic plots for multiple linear regression model

When looking at the results of the model, as shown in table 1, we can see that there was evidence of a relationship between mean age and victimisation rate, with coefficient -19.25, CI (-38.47, -0.04), p-value 0.05. We can also see there is no evidence of a significant relationship between mean unemployment rate and victimisation rate, with p-value = 0.18. We can also see that the interaction term of mean age and mean unemployment rate was non-significant, with p-value = 0.15. This model has the following equation:

$$y = 75.305 - 19.25\log(age) - 27.46\log(unemployment) - 7.70\log(age * unemployment)$$

The adjusted R^2 for this model is 33%, this indicates that the model is not very good for making predictions.

Variable	Estimate	P-value	Lower CI	Upper CI
log(mean age)	-19.25	0.05	-38.47	-0.04
log(mean unemployment rate)	-27.46	0.18	-67.57	12.65
Interaction Term	7.70	0.15	-2.97	18.37

Table 1: Results of linear regression model involving mean age and mean unemployment rate

As mean unemployment rate and the interaction term were found to be non-significant at the 5% level, I decided to make a linear regression model that just included mean age. Using

victimisation rate along with the log-transformed mean age, the resulting diagnostic plots were produced as seen in Figure 7.

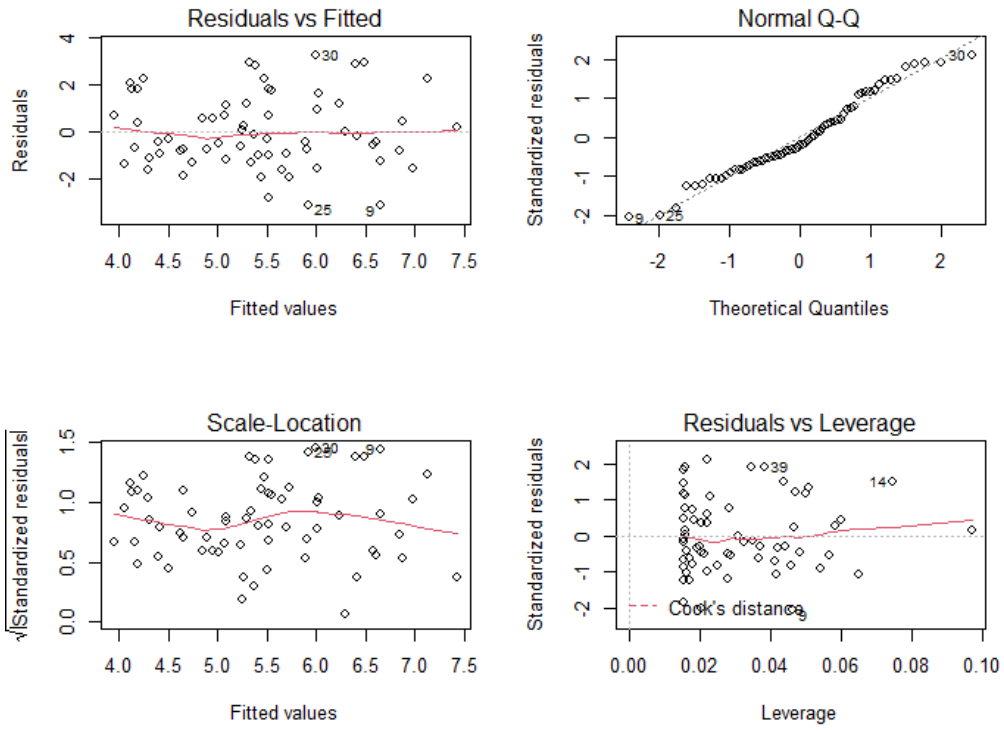

Figure 7: Diagnostic plots for linear regression model

By looking at the diagnostic plots, we can see that this model fits the linear regression assumptions even better than the first model. Looking at the results, shown by table 2, we can see that there is a very significant relationship between age and victimisation rate with coefficient -8.67, CI (-12.57, -4.78), p-value <0.01.

Variable	Estimate	P-value	Lower CI	Upper CI
log(mean age)	-8.67	0	-12.57	-4.78

Table 2: Results of linear regression model involving mean age

This model has the following equation:

$$y = 37.84 - 8.67 \log(age)$$

The adjusted R^2 for this model is 22.7%, which indicates that the model is not very good for making predictions but it does indicate that age accounts for 22.7% of the variance in victimisation rate which, as this is real world data, is quite a lot.

7 Conclusion

The models produced from this analysis have been found to not be very powerful when it comes to trying to predict victimisation rate per capita per territory. This is likely due to the limitations of the datasets being used. As the New Zealand Census is only conducted once every 5 years, with the most recent being 2018, only the Victimisation Time and Place data from the year 2018 could be used reliably as the entire dataset contains data from 2014-2020. If population values for all the years were available, the models would likely be greatly improved.

References

[1] Police New Zealand. *Victimisation Time and Place Dataset*. 2020. https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz/victimisation-time-and-place.

[2] Statistics New Zealand. *2018 New Zealand Census*. 2018. https://www.stats.govt.nz/information-releases/statistical-area-1-dataset-for-2018-census-updated-march-2020.

[3] Statistics New Zealand. *Geographic Area Dataset*. https://datafinder.stats.govt.nz/?_ga=2.201599507.38014314.1598730863-1186394837.1596602881&_gac=1.203546916.1598730894.EAIaIQobChMIo-WgwJjB6wIVjqmWCh28GAhiEAAYASABEgJK5PD_BwE.

[4] Stats NZ. *2013 Census confidentiality rules and how they are applied*. 2013. http://archive.stats.govt.nz/Census/2013-census/methodology/confidentiality-how-applied.aspx#gsc.tab=0.

[5] Police New Zealand. *Recorded Crime - Victims User Manual*. 2016. https://www.police.govt.nz/sites/default/files/publications/nz_recorded_crime_victims_manual_v1.2.pdf.

[6] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy (First edition)*. 2016.

[7] Andrew Tarantola. *Predictive policing' could amplify today's law enforcement issues*. 2020. https://www.engadget.com/predictive-policing-privacy-civil-rights-dangers-133040971.html.

[8] New Zealand Police. *Prevention First — Āraia i te tuatahi: National operating model*. 2017. https://www.police.govt.nz/sites/default/files/publications/prevention-first-2017.pdf.

[9] Stats NZ. *Territorial authority*. 2015. http://archive.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/territorial-authority/definition.aspx#gsc.tab=0.

[10] Stats NZ. *Employment and unemployment*. n.d. https://www.stats.govt.nz/topics/employment-and-unemployment.