The Study on Factors Corresponding to Potential Risks of Victimization

Ke Deng, Yongpeng Hua, Qihui Huang, Qing Wen

10/19/2020

Abstract

The victimization rate in Canada is found to be a lot lower than in 2014. Though the victimization rate is decreasing, the feeling of safety does not seem to rise with it. In this paper, the main objective is to determine factors that are potentially influential in determining people's risk of being victimized. To accomplish this, we decide to fit a logistic model with a bivariate response of being victimized or not. Some preliminary results include that older people tend to have a lower risk of being victimized. People who self-rated their mental health as poor or fair have a relatively high risk of being victimized.

Introduction

Criminal victimization has been a prevalent social phenomenon. It is always a devastating experience for the victims suffering from physical or mental abuse, like anxiety, fear and frustration. Since large adverse effects are outstanding, studies of Canada's victimization have been conducted, and they deliver meaningful results to the public. The Canadian General Social Survey(GSS) conducts surveys on victimization to understand how Canadians perceive crimes and their victimization experiences. The survey in 2014 collects responses from non-institutionalized individuals national-wise and ends up with large sample size. According to Statistical Canada, the GSS on victimization is considered a comprehensive survey of self-reported victimization. Hence, the 2014 GSS survey on victimization is considered a reliable data source for studying potential risks. The paper has five main parts. Part II focuses on data collecting and cleaning. Part III explains the model we chose, including benefits, drawbacks and justification of the model. In Part IV, results from the model output are displayed. Finally, part V delivers an extensive discussion on the study results, including result interpretation, validity check, limitations, and the output's contribution to the original goal.

Note

All relevant code for this report can be found in the following link: https://github.com/QingWen-0310/The-Study-on-Factors-Corresponding-to-Potential-Risks-of-Victimization

Data

i.Survey data

The general social survey on victimization 2014 is conducted over Canada from August 1st, 2014 to January 17th, 2015. The survey and the data are collected through the Computer Assisted Telephone Interviewing (CATI) system. The survey is conducted voluntarily through phone calls, and the interviewers of the survey department tried to make phone calls more than once and explained the importance of the survey to reduce

the non-response rate. According to the documentation book, the response rate was about 52.9%. The target population for the victimization survey is "all persons 15 years of age and older in Canada," excluding people in Yukon, Northwest Territories, and Nunavut, and "full-time residents of institutions." Using the data from Statistics Canada, the population size is 35423701. The sample is chosen using stratified sampling and simple random sampling without replacement(SRSWOR), where the target population is first divided into small groups called strata by different locations. Then the final sample is formed by randomly selecting units within each stratum. The use of SRSWOR tends to give a representative sample of the population. The sampling frame includes all the telephone numbers available to Statistics Canada, where the target frame size comes to 39674, while the actual sample ends up being 33089. With more than 16 sections of questions, there are a total of 790 variables, both categorical and numerical.

ii.Questionnaire

The questionnaire is well-organized. It provides detailed background information and is entirely voluntary so that respondents are more likely to give out their real thoughts. Different types of questions (e.g. scaling and yes-no question) help collect responses from both subjective and objective sides. The survey format and questions are being improved each time according to the feedback received from the respondents. On the other hand, the survey's downside is that it is both time and money-consuming, as the survey is conducted national-wise. Moreover, confidentiality cannot be guaranteed, as some people may not choose to give a natural response during an in-person or telephone interview. However, drawbacks are hard to overcome, as the adjustments made may result in spending more money and time.

iii.Data for Study

As our logistic regression model focuses on the probability of being a victim, explanatory variables that seem to be plausibly influential in determining the risk of victimization include: *age group*, *number of evening activities per month*, *number of hours worked per week*, *carry things to defense*, *self-rated mental health*, *number of victimization*. A structure of the data is presented as below:

Victimization	Age group	Contact police	Defensive carryings	Safe route Planning	Stay home at night
0	25_to_34	somewhat likely	Yes	No	No
0	25_to_34	very likely	No	Yes	No
0	25_to_34	somewhat likely	No	No	No
0	55_to_64	very likely	No	No	No
0	55_to_64	somewhat likely	No	No	No
1	65_to_74	very likely	No	No	No

The variable *age group* is chosen as people of different age groups have different lifestyles and habits. It is categorized into seven categories, each representing an age interval. Since most victimization cases happen at night, a person's evening activities and one's working schedule can help investigate the probability of victimization. The variable *evening activity per month* is used in the model. It is a general variable measuring the average number of evening activities a respondent has in a month. The *number of hours worked per week* variable is numerical and records how many hours the respondent spends working every week. The variable *carry something for defense* records whether the respondent carries something for defense. Furthermore, according to Statistics Canada, more than 1 million Canadians suffer from mental health issues; therefore, mental health could also be a good factor. The *self-rated mental health* is a categorical variable with five categories, each representing the level of one's mental health degree, from poor to excellent. The responsive

variable here is whether the observation is *victimized*, stating whether there is victimization happening. This variable is mutated from the variable *number of victimization* and has a value of 0 if the individual has never been victimized and a value of 1 otherwise. Given over 30,000 observations, some respondents choose not to answer some questions in the survey due to the sensitivity of the topic(which is recorded as value 7 representing a good skip). The missing values are removed in the analysis as they provide meaningless information.

Model

i : Model structure

Regarding the importance of public safety and experience of victimizations shown in the survey, we look at the risks of victimization associated with different groups of people in the model below, which becomes the focus of our study. To fit our interest of predicting the risk of victimization, it is plausible to choose a logistic regression model since our response variable is bivariate. It may be questionable whether to adopt a multiple linear regression model(MLR) as an alternative. If an MLR model is adopted, it is helpful to look at how many times one was victimized instead of being victimized or not. Hence, an MLR model may not produce desirable results for our study purpose. The model adopted to the chosen response and predictor variable is written as

$$log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0 + \hat{\beta}_1 * x_{AgeGroup} + \hat{\beta}_2 * x_{NumHoursWorkedPerWeek} +$$

$$\hat{\beta}_3 * x_{EAPerMonth} + \hat{\beta}_4 * x_{CrimeSafetyCarrySthDefend} + \hat{\beta}_5 * x_{SelfRatedMentalHealth}$$

The log odds function on the left side of the equation represents the log changes in probability of victimization. The true parameters to look for for the model is $\hat{p}$, which can be algebraically manipulated after the log odd is predicted.

$\hat{p}$: the predicted probability of being victimized

$\hat{\beta}_0$:intercept of log odds when all responsive variables are 0

$\hat{\beta}_1$: the average difference in log odds of victimization between age groups coded as 0 and coded as 1

$\hat{\beta}_2$: the expected change in log odds of victimization when number of working hours per week increase by one

$\hat{\beta}_3$: the expected change in log odds of victimization when evening activity hours increase by one

$\hat{\beta}_4$: the average difference in log odds of victimization between a person carrying something to defend or not

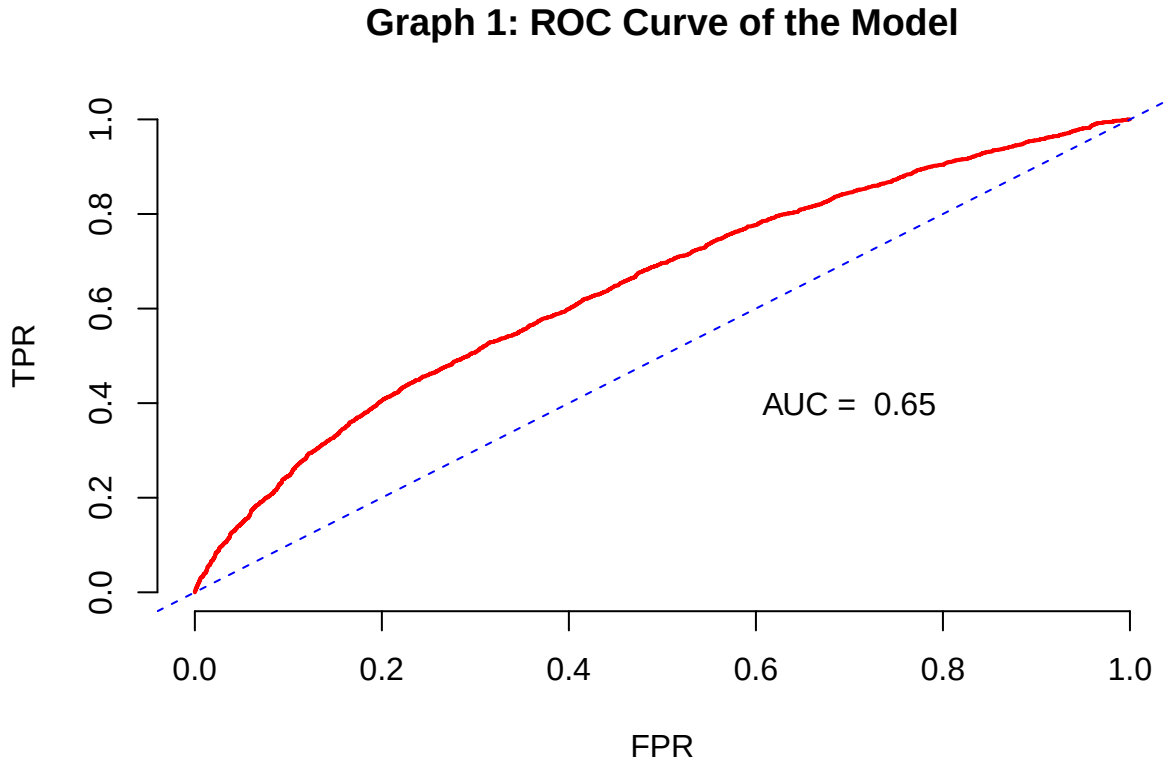
$\hat{\beta}_5$: the average difference in log odds of victimization between people of different self-rated mental health levels

Among all the variables of interest, two of them enter the model numerically(num_hours_worked_per_week, ea_per_month) as they are themselves numerical. The rest three variables are kept as categorical. When incorporating into the model, these variables are treated as having different levels, with each level corresponding to a value the variable takes. This way, it can be shown how the predictor variable affects the log odds but also how each **level** the predictor affects the log odds specifically. The model is built using the software R studio with the "glm()" function in the Survey package, which fits a logistic regression model with a given survey design. In the design, a "finite population correction"(fpc) is incorporated based on the population size to adjust for population variance, since we know the population is finite.

ii: Model validity

To check the goodness of the model, the method of AUC - ROC curve is applied (*Graph 1*). This check is based on the *test2_data*, which filters out all the NAs in relevant variables in the data list. AUC – ROC

curve is a performance measurement for the classification problem at various threshold settings. ROC is a probability curve and AUC, area under curve, represents the degree or measure of separability. The plot tells how much the model is capable of distinguishing between classes. The AUC has a range of 0 to 1. The closer the AUC is to 1, the better the model is. In other words, the higher AUC model has a better ability to distinguish between people being victimized and not victimized. In this case, the AUC of the logistic regression model is 0.65, which means the model is still able to be improved. When AUC is equal to 0.5, it means the model has no class separation capacity. Therefore, when AUC gets close to 0.5, the distinguishing ability is not as good as expected.



Results

Table 1 below shows the basic summary information of the logistic regression model. It contains the name of each coefficient, the estimate value, and the p_value.

Coefficient	Estimate	P_value
Intercept	-1.765497	< 2e-16
ea_per_month	0.022619	< 2e-16
num_hours_worked_per_week	0.007836	3.84e-06
as.factor(crime_safety_carry_sth_defend)Yes	0.692114	< 2e-16
as.factor(age_group)25_to_34	-0.094625	0.236180
as.factor(age_group)35_to_44	-0.262722	0.000861
as.factor(age_group)45_to_54	-0.516509	1.57e-10
as.factor(age_group)55_to_64	-0.674416	1.47e-15
as.factor(age_group)65_to_74	-1.008986	7.07e-13

Coefficient	Estimate	P_value
as.factor(age_group)over_75	-0.650995	0.103207
as.factor(self_rated_mental_health)Fair	0.896911	< 2e-16
as.factor(self_rated_mental_health)Good	0.228197	0.000303
as.factor(self_rated_mental_health)Poor	1.282224	2.67e-06
as.factor(self_rated_mental_health)Very good	0.146291	0.010159

Table 2 below shows the proportion value of different age groups in categorical variable *age group*. The bigger the number is, the higher the probability to be victimized.

age_group	proportion
15_to_24	0.3145161
25_to_34	0.2896053
35_to_44	0.2497657
45_to_54	0.2034043
55_to_64	0.1710587
65_to_74	0.1161290
over_75	0.1454545

Table 3 below shows the relationship between different levels of *self-rated mental health* and whether the individual is *victimized*. The higher the number, the higher the probability to be victimized. In this case, people who took this survey and rated themselves poor mental health have the highest probability to get victimized.

self_rated_mental_health	proportion
Excellent	0.2053809
Fair	0.4038055
Good	0.2376041
Poor	0.4603175
Very good	0.2292198

Figure 1 above shows the relationship between *age group* and *be victimized*. It shows both probability of being victimized and not being victimized. In this case, people from 35 to 44 years old are more likely to be victimized.

Figure 2 above shows the relationship between *self-rated mental health* and *be victimized*. It also shows both the probability of being victimized and not being victimized. In this case, even though most people are not likely to get victimized, it looks like people who rated themselves very good mental health get victimized most.

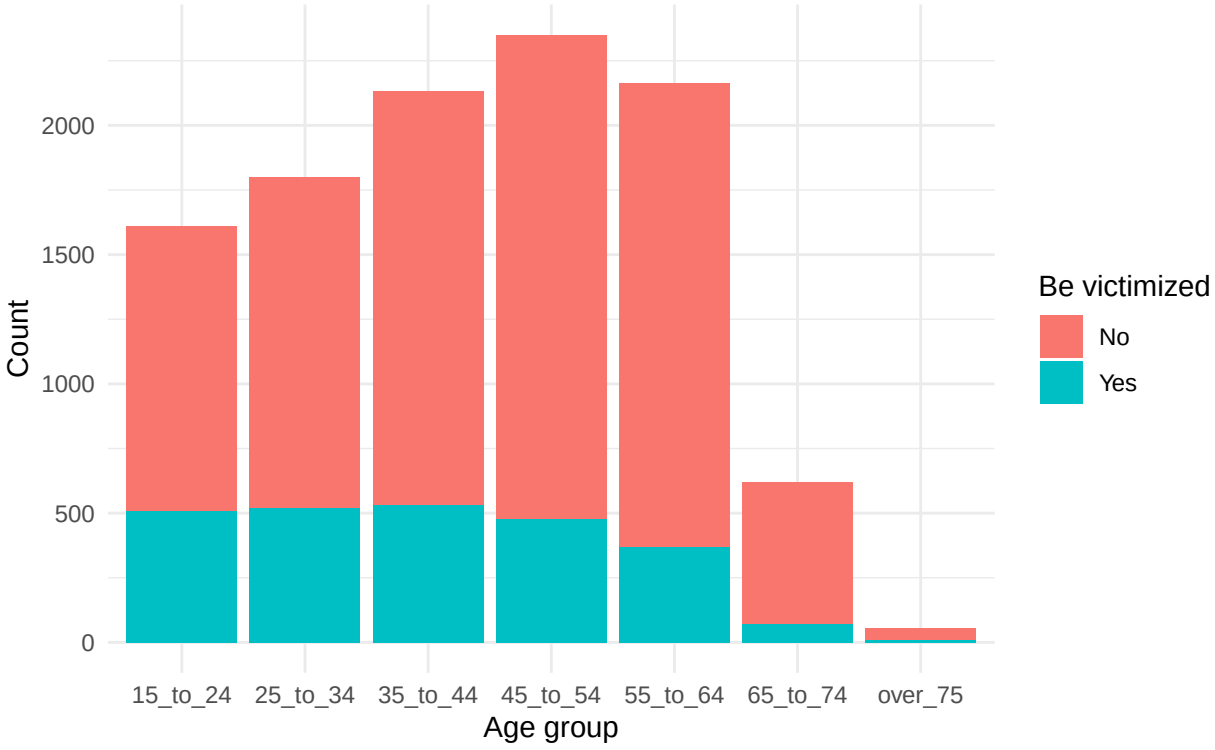


Figure 1: Figure 1: Barplot of relationship between age group and victimization

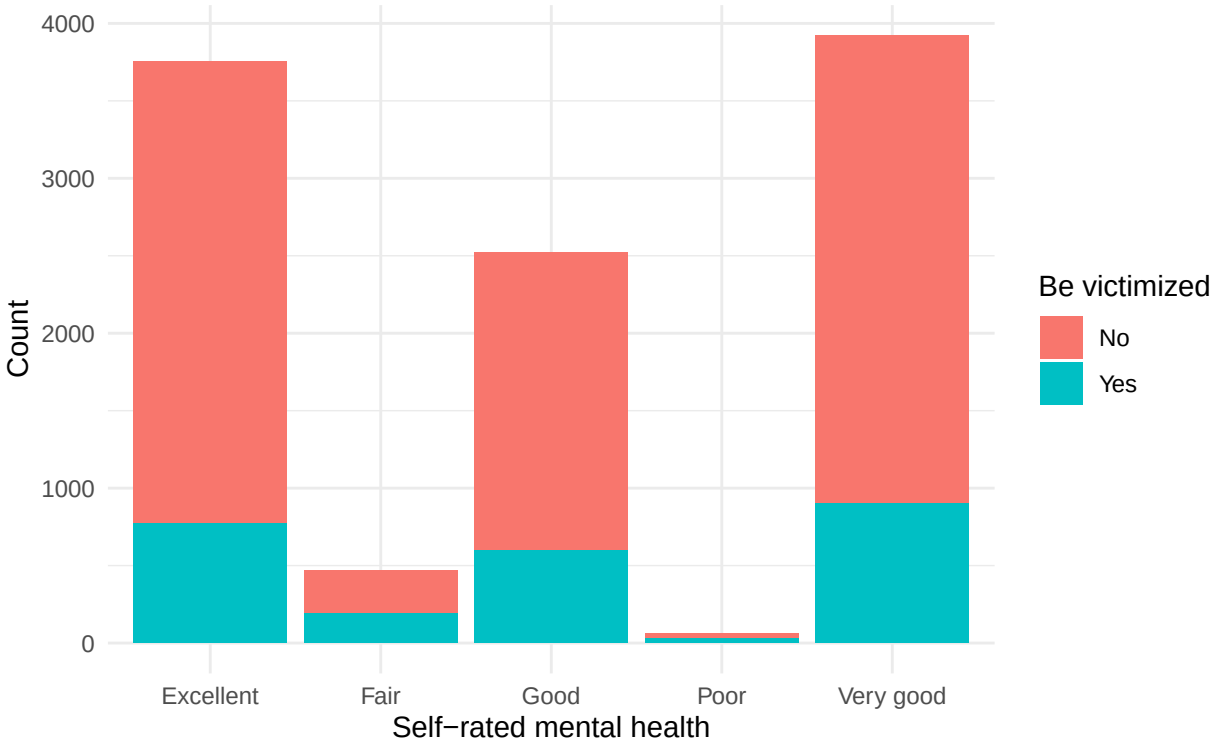


Figure 2: Figure 2: Barplot of relationship between self-rated mental health and victimization

Discussion

Output for Logistic Regression Model

Table 1 includes all the estimates for each predictor variable given by the logistic regression model simulated in R. The intercept estimate indicates that, on average, the log odds of the risk of being victimized is -1.765. However, extrapolating beyond the range of the predictor values is not practical. Due to the survey's nature, people below the age of 15 are not our target population, so it does not make sense to predict the risk for this group of people. For both the *evening activity per month* and *number of hours worked per week*, though their estimates seem to be statistically significant at any considerable significance level (having a p-value close to 0), the estimates suggest that these two factors may not play a huge role in determining the risk of victimization(one having an estimate of 0.02 and the other has an estimate of 0.007). Hence, though these two variables suggest that with a one-unit increase in the variable value tends to increase the log odds, the correlation does not seem to be significant. On the other hand, we can see that the estimates for different age groups are all negative, suggesting that people having age-inclusive to the survey population. The log odds for them would be decreasing, indicating a lower probability of being victimized. Moreover, the estimates seem to be decreasing as people grow older. To be more specific, people in the age group 25-34 tend to decrease the log odds by 0.09, while people in the age group 65-75 tend to decrease the log odds by one, which is significantly more than 0.09. This relation suggests that as people grow older, the risk of them being victimized is lowered. One exception is for people older than 75, in which the estimate(-0.65) does follow the decreasing trend of the estimates, but it still decreases the log odds more than the estimates for younger groups do. Finally, the estimates for different self-rated mental health levels make the two groups of people stand out. The estimate for poor mental health is 1.28, and the estimate for fair mental health is 0.9, which is a lot higher than the estimates for other mental health groups. It means that people with poor or fair mental health tend to be at a higher risk of being victimized. One unexpected result comes from the estimate for *carry something for defense*, which has an estimate of 0.69. It suggests that people who carry items for defense typically have a higher risk of victimization than people who do not.

Different Age Groups and actual Victims in each category

The barplot, *Figure 1* studies how different ages relate to the victims' number. The largest proportion of victims is in the 15-24 group. The second largest is the 25-to-34 group. The proportion of victims among the mid-aged and elderly adults (55+) is much lower than the other groups. Thus, though the elderly has more fear regarding the possibility of being victimized, the truely victimized number among the elderly group is much smaller than the ones in younger group. It may due to the fact that people in younger groups goes out more often at night and to a lot of places than the elderly does. The plot shows that 3 of the age groups 35-44, 45-54 and 55-64 are the most responsive, as they have the most valid responses. The response rate will also have to be considered, which means we cannot draw direct conclusions from the survey. To improve the result, we could control the number of people who take the survey. Therefore we could get the groups with about the same amount of survey takers.

Self_rated_mental_health and actual Victims in each category

Figure 2 depicts the relationship between respondents' self-rated mental health and their victimization experience. As most respondents rated their mental health as either Excellent or Very Good, only a small proportion of respondents rated themselves as having poor mental health. It can also be inferred from the plot that most of each group does not have a victimization experience except for the "poor" group. When looking closer to the proportion of victimized individuals in each group in *Table 3*, people with self-rated "poor" and "fair" mental health tend to have the highest victimization rate: 40% of people with fair mental health and 46% of people with poor mental health have been victimized before. Since no causation effect can be inferred from a non-experimental study, it suggests that people being victimized tend to have poor to fair mental health, which may guide the public to care more about the mental state of the members of their households.

Generalization & Limitation

As criminal victimization has been a prevalent social phenomenon every day in Canada, it is always a devastating experience for the victims suffering from the perpetrator's physical or mental abuse and the

trauma to be dealt with afterwards. According to the Canadian Resource Centre for Victims of Crime, the effects of being usually victimized last relatively long in victims' lifespan. Victims suffer mentally from anxiety, fear, shame, and frustration, along with physical reactions like increased heart rate, feeling of being frozen, and an enhanced "fight or flight" response. Due to the large negative effect, it has on human beings, many studies and investigations about victimization in Canada are conducted. They can deliver meaningful results to the public.

References

Burczycka, Marta. "Violent Victimization of Canadians with Mental Health-Related Disabilities". 2014. Statistics Canada: Canada's National Statistical Agency / Statistique Canada : Organisme Statistique National Du Canada, Government of Canada, Statistics Canada, 18 Oct. 2018, www.150.statcan.gc.ca/n1/pub/85-002-x/2018001/article/54977-eng.htm.

Canadian Resource Centre for Victims of Crime. "The Impact of Victimization", 2005. PDF file. "Domestic Violence/Intimate Partner Violence: Applying Best Practice Guidelines." ACE - Access Continuing Education - Domestic Violence/Intimate Partner Violence: Applying Best Practice Guidelines, www.accesscontinuingeducation.com/ACE4000LP-11/c5/index.htm.

General Social Survey - Canadians' Safety (GSS). www.23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey.

Government of Canada, Statistics Canada - tables and charts. "Population of Canada." Government of Canada, Statistics Canada - Tables and Charts, 27 Nov. 2015, www.150.statcan.gc.ca/n1/pub/12-581-x/2015000/pop-eng.htm.

Narkhede, Sarang. "Understanding AUC-ROC Curve". 2018. Towards data science, 27 June. 2018, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical, Computing, Vienna, Austria. URL https://www.R-project.org/.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, Müller M (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves." BMC Bioinformatics, 12, 77.

Social Challenges: Age, www.ccsd.ca/resources/CrimePrevention/c_age.htm.

Swaminathan, Saishruthi. "Logistic Regression - Detailed Overview." Medium, Towards Data Science, 18 Jan. 2019, www.towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.

T. Lumley (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1): 1-19

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686