Bayesian binary regression model: a link between childhood maltreatment and aduldhood victimization

Huining Qu

17/10/2020

Abstract

Childhood abuse and maltreatment have been determined to increase the risk of criminal behavior. In this report, a Bayesian binary logistic regression model with non-informative priors is developed, however, to predict violent crimes towards whom experienced childhood abuse or witnessed spousal violence. We found that people who experienced childhood mistreatment are more likely to be the victim of violent crimes. Such finding has been overlooked by social studies, yet remains its significance as more resources are needed for child maltreatment victims, to not only minimize the risk of future criminal behavior, but also to prevent them from being continuously victimized.

Introduction

Childhood maltreatment is known to have significant psychological impacts on adults. Many studies have linked childhood abuse to criminal involvement as an adult. However, the behavioral impact of childhood experience cannot be limited to just criminal involvement. In this paper, we use Bayesian binary regression model, hoping to inspect the impact of childhood maltreatment on the other side of the spectrum - adulthood victimization.

In this paper, we have explored 2014 General Social Survey: Canadians' Safety and Security, and developed a regression model using childhood assault and violence between parents, aiming to predict the probability of an adult being violently victimized. Our findings suggests that being abused as a child, physically or sexually, will lead to a 83% increase in probability of experiencing violent crimes. Moreover, adults who themselves had never been abused, but witnessed violence between parents as a child, also show a higher probability of violent victimization.

In summary, childhood maltreatment, regardless the extend, will increase the probability of violent victimization as an adult. In this report, we will further explore the data, model, and results, intending to provide a through explanation of such conclusion. The code can be accessed from: https://github.com/mackenziequ/regression/blob/main/r%20code. That being said, this report only demonstrates the existence of such phenomenon, we do not have enough background information to provide and behavioral or psychological reason behind this trend.

Data

General Social Survey

The data used in this report is from Canadian 2014 General Social Survey(GSS): Criminal Victimization. Upon downloading the data, we have chosen only the variables involved in this study, including Childhood Experiences, Crime Incident Report, and Demographic Derived Variables, aiming towards fitting our statistical model.

In the 2014 GSS, the target population is Canadian residents aged 15 or above - excluding Residents of the Yukon, Northwest Territories, and Nunavut; and Full-time residents of institutions.

GSS 2014 had received funding for an oversample. The sample size was 39,674 while the actual data collected was 33,089 after removing some confidential observations. To reach all the desired respondents, computer assisted phone calls are used to conduct the interviews. As in territories, the interviews are done in person or by telephone. Each interview takes approximately 45 minutes. Moreover, some data such as income are drawn directly from tax or other administrative files instead of questionnaires as "Statistics Canada began asking respondents for permission to link their survey information to additional data sources in 2014"(Statistics Canada).

The sample is taken with a stratified approach with probability sampling, which is done at the "province/census metropolitan area level."(Statistics Canada) Canada is first divided geographically into each province, and then each sample is randomly selected within each strata(ie. geographic locations). Census Metropolitan Areas for each province such as Toronto, Vancouver, Ottawa are counted as each separate strata. The remaining areas of each province are grouped to form 10 more strata. Within each strata, groups of telephone numbers associated with each dwelling are randomly selected proportional to the population, then one eligible potential respondent is selected within the identified household. Such two stage sampling design minimizes the repetition of data and maximizes the representation of the population. Stratification is also applied for the oversample of immigrants and youth. For more specific strata distributions, view Appendix[1].

The survey frame was created using lists of telephone numbers in use, combined with Address Register. The two were then combined to group all telephone numbers associated with the same address. The telephone numbers without a valid address are also included in the frame. The samples are representative for all households in Canada. Note that the sample frame for the oversample was altered, more detail is discussed in Appendix[2]. For non-respondent, multiple attempts were made to encourage their participation. The overall response rate was 52.1%, which is significantly higher than the adjacent years. Non-response was adjusted by weighting the responding interviews in each strata.

Weakness

In 2014 landline method was completely abolished, resulting in failure of reaching households without a telephone. However, Statistics Canada revealed that "in 2013, the proportion of households without any phone service was estimated at 1%"(Statistics Canada), indicating that the exclusion error caused by abolishing landlines are not as significant. Sampling error and bias, instead, are largely due to its sampling method. For instance, some key characteristics of the population are not thoroughly represented due to sampling error and lack of identification from random sampling. Secondly, the sample is limited by exclusion errors during the sampling process. Statistics Canada has mentioned that "data are not available for First Nations people living on or off reserve specifically"(Statistics Canada) because the sampling population is not large enough to estimate for the entire population living on reserves. Due to such sampling bias, some data has to be combined and weighted in order to obtain reliable data. Moreover, sampling error was derived from the non-response rate in 2014. Although it was lower than the adjacent years, 47.1% of non-response rate might cause limitation of generalizing the sample to its population, as sample became less representative, the margin of error might increase.

Survey

The 2014 General Social Survey focused on crime and victimization. In this report, we focus on the two subtopics: Crime Incident Report and Childhood Experiences. The questionnaire of these two related subtopics are well designed, as it broadly covers the topics in detail. Crime incident Report includes personal incidents and household incidents varies from assault to motor vehicle thefts. And the Childhood Experiences includes childhood victimization, such as physical and sexual abuse, relationship with the abuser, and spousal violence. The survey was in general, extremely detailed and clear. However, it is noticeable that the refusal rate for childhood victimization questions was high due to privacy or personal reasons. 334 respondents refused to answer if they experienced physical assault before the age of 15, 394 respondents reduced to answer if they experienced sexual assault, and 366 respondents refused to state if they witnessed violence between parents. Compared to 54 respondents who had refused to answer if they are confident in police in the following section, the refusal rate was significantly higher, which may give rise to possible bias. Some more weaknesses from the data which may influence our model will also be discussed below.

Raw Data

The three variables we used were abuse(figure 1), violence(figure 2), and victimization(figure 3). In the survey, multiple questions were asked about childhood experiences, topic including abuse, spousal violence, and incident report. We have chosen Childhood Assault more specifically as it includes both physical and sexual violence("assault"). Childhood maltreatment also includes witnessing violence between parents, we have chosen to include that as one of our variables("violence"). For Criminal Activities, our variable("victimization") includes only personal experience, excluding any household event because they have no effect on if respondents have experienced serious violent incidents. Note that respondents who have experienced serious violent incidents(ie. sexual assault, robbery, attempted robbery, assault) in figure 3 is an extremely small proportion, as it does not include violence and assault by spouse/partner. The topic is included in the questionnaire under "Physical and Sexual violence by spouse/Partner (PSP)", but not disclosed in the GSS data. This has a significant effect on our model, and will be shown and discussed in detail later on.

(1)

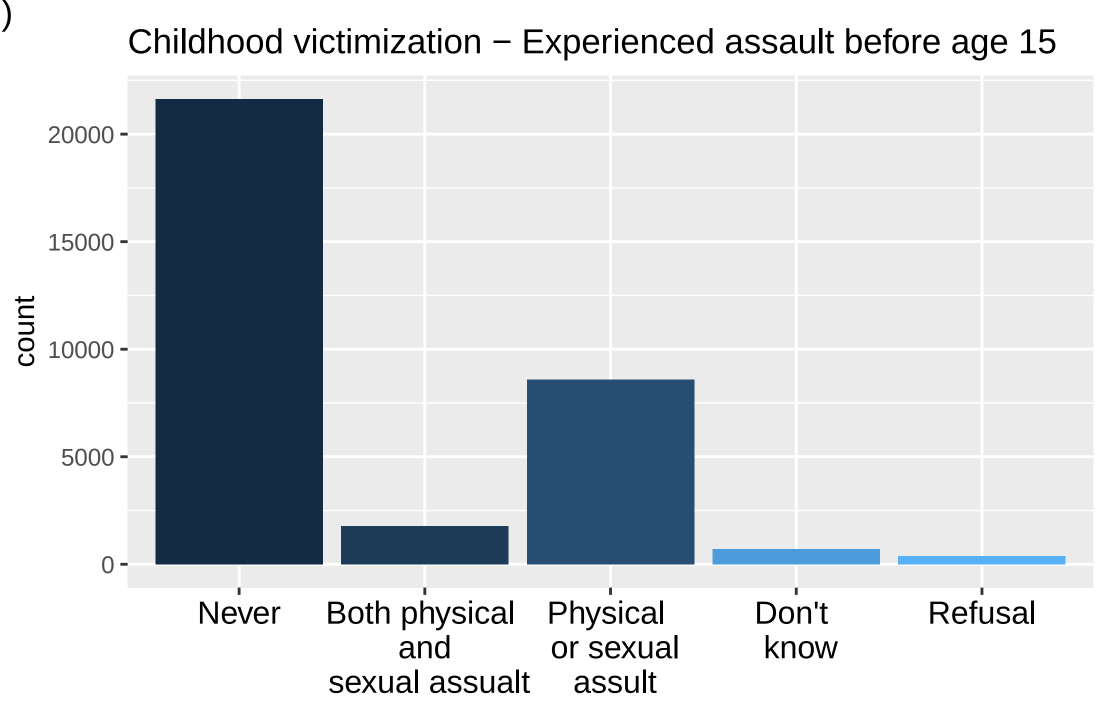


Figure 1 - Childhood victimization - Experienced assault before age 15

(2)

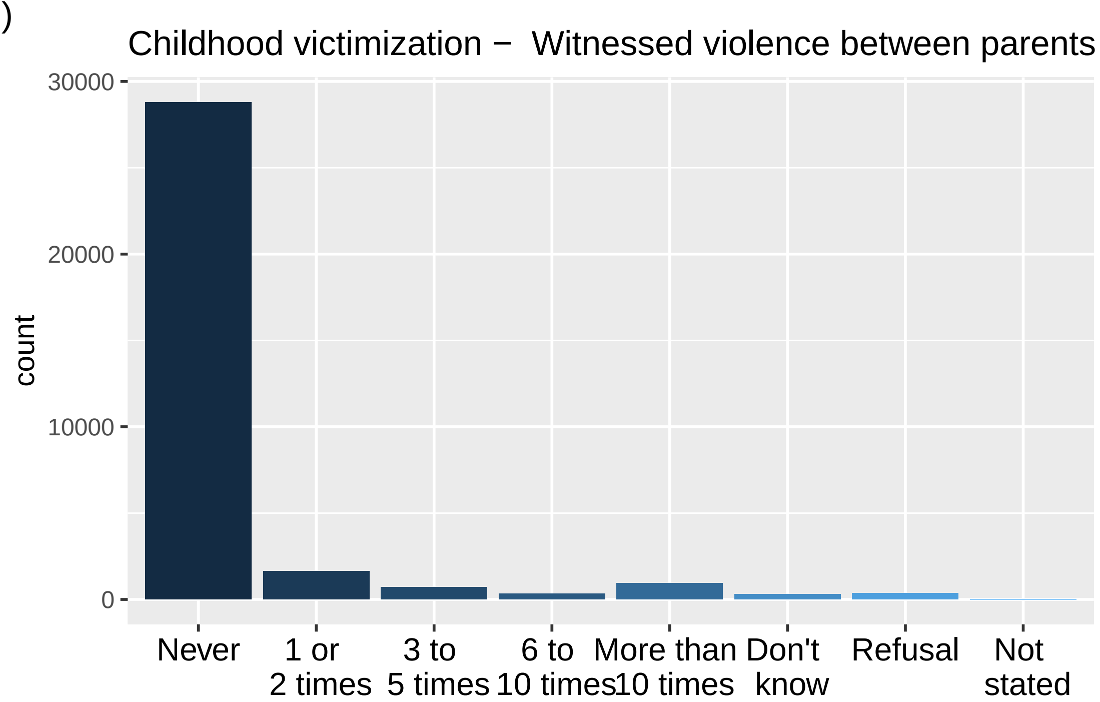


Figure 2 - Childhood victimization - Witnessed violence between parents

(3)

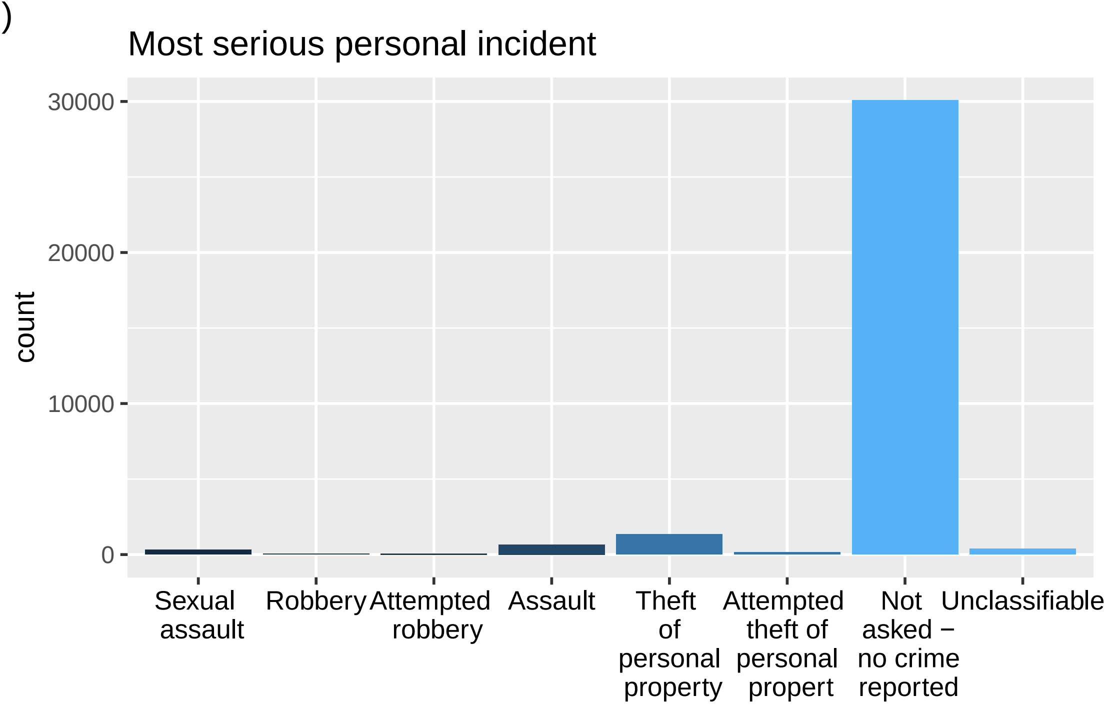


Figure 3 - Most serious personal incident

Prior to modeling the data, we have made some visible changes to the variables used to help it fit into our model. We have first categorized all "valid skip", "don't know" and "refusal" under N/A. Then we adjust the survey answer accordingly. For Childhood Assault, we have changed "Never" to numeric 0, and combined "Both physical and sexual" and "Physical or Sexual Assault" into numerical value 1. The same adjustment was done for "witnessed violence between parents" as we are looking for either yes or no for these two variables.

Upon examining “most serious personal incident”, we have categorized all personal incidents into violent incidents, and non-violent incidents. For respondents who answered “sexual assault”, “robbery”, “attempted robbery”, and “assault”, we have grouped them as 1, otherwise 0. This allows us to have fit our data into Bayesian Binary Logistic Regression.

Data Preview

Below(Table (1)) is a preview of out cleaned and altered data.

abuse	violence	victimization
1	0	0
0	0	0
1	0	0
0	0	0
0	0	0
1	0	0

Table(1) - Data Preview

Before establishing our model, we first observe the count of violent incidents for our two explanatory variables - violence and abuse in Table(2) and Table(3). From the observations, it seems that number of people experienced serious incidents differs a lot between if they have witnessed violence between parents(Table (2)), and more people experienced childhood assault are victims of incident crime(Table 3). This observation indicates that violence between parents and experienced childhood assault might be predictive of violent victimization.

violence	victimization
0	855
1	235

Table(2) - Victimization ~ Violence

abuse	victimization
0	502
1	588

Table(3) - Victimization ~ Abuse

Model

Our main focus is to estimate if a person will be the victim of violent crime(ie. sexual assault, robbery, physical assault). According to the summary of GSS 2014(cdeunodc), "People who suffered child maltreatment were more likely to be victims of a violent crime", we are hereby to verify such a statement. We use the Bayesian Binary Regression Model(Collet, 1994) with two predictors - violence, and abuse to explain the probability of a person experiencing violent victimization based on their childhood experiences.

We have used a binary model due to the response. Let Y_i denote the i-th response, $Y_i = 1$ when the person has experienced violent crime, and $Y_i = 0$ otherwise. Each individual experiencing violent crime is modeled as a response Y_i with a Bernoulli distribution, where θ_i is the probability of violent victimization:

$$Y_i|\theta_i \sim Bernoulli(\theta_i), i = 1, 2, 3..., n \qquad (1)$$

The Bayesian model suggests we an use the following equation to find the posterior distribution:

$$p(\theta|\beta) = \frac{p(\beta|\theta)p(\theta)}{p(\beta)} \qquad (2)$$

Let β represent a vector of unknown parameters. $p(D|\theta)$ in equation (2) is the likelihood, and $p(\theta)$ is the prior probability. More specifically, equation (2) can be interpret as:

$$posterior \propto likelihood \times prior \qquad (3)$$

In a full Bayesian approach, we would first prostate prior distribution on the regression coefficient. Little historical data or related literature was found to establish a strong informative prior. The Bayes/Laplace postulate stated that, "when nothing is known about θ in advance, let the prior $\pi(\theta)$ be a uniform distribution, that is, let all possible outcomes of θ have the same probability". Thus, in this study, we are using a non-informative prior with Uniform distribution. Non-informative prior is not the best approach, however, due to the large sample size of 31,672, the likelihood overwhelms prior. Therefore the prior will not cause significant different.

The logistic equation for our model is:

$$logit(\theta_i) = log(\frac{\theta_i}{1 - \theta_1}) = \beta_0 + \beta_1 x_{violence} + \beta_2 x_{abuse} \qquad (4)$$

$x_{violence}$ and x_{abuse} are our explanatory variables, each will have either "Yes"(1) or "No"(0) as an answer. And β_0, β_1 are unknown regression parameters. We can rearrange equation (4) to get the probability of response:

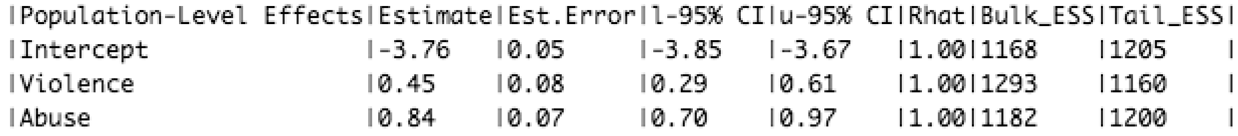
$$\theta_i = \frac{exp(\beta_0 + \beta_1 x_{violence} + \beta_2 x_{abuse})}{1 + exp(\beta_0 + \beta_1 x_{violence} + \beta_2 x_{abuse})} \qquad (5)$$

Equation (5) ensures the probability of $i - th$ individual experiencing violent crime lies in the interval $[0, 1]$.

We then used the "brm" function from "brms"in R to perform Bayesian model. In the argument, family is specified to be Bernoulli, aiming to achieve the result. Our model in R would be represented by:

$$serious\_incident \sim violence\_between\_parents + experienced\_childhood\_assult$$

Below is the summary of the model:

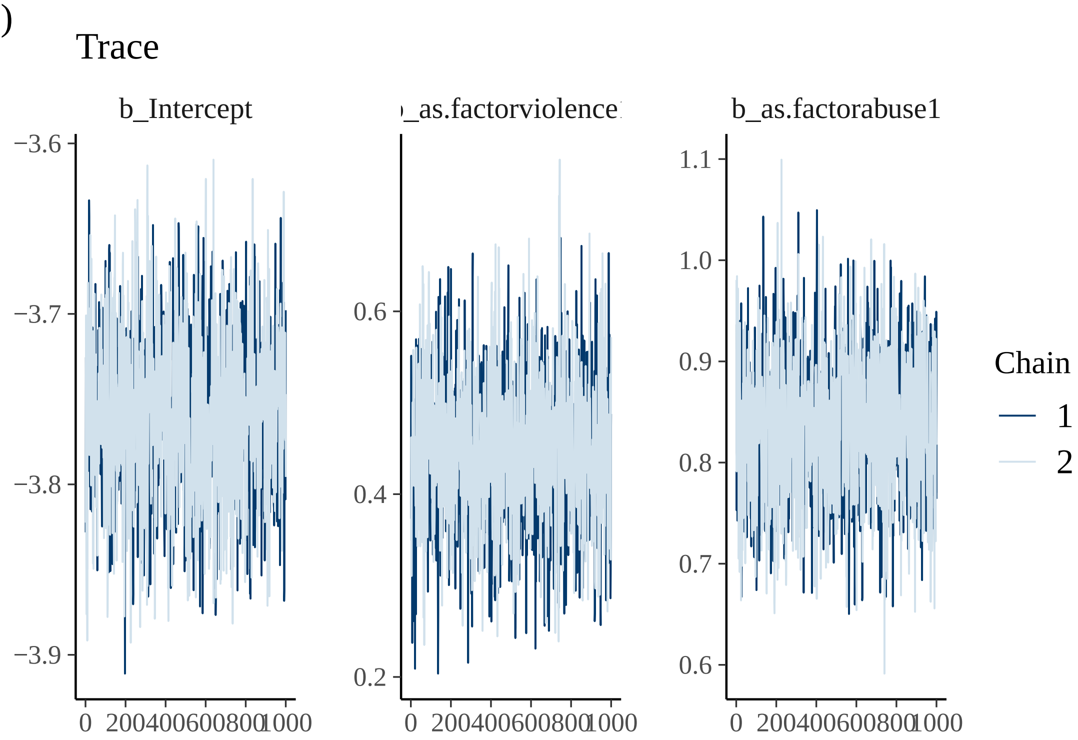


```
|Population-Level Effects|Estimate|Est.Error|l-95% CI|u-95% CI|Rhat|Bulk_ESS|Tail_ESS|
|Intercept               |-3.76   |0.05     |-3.85   |-3.67   |1.00|1168    |1205    |
|Violence                |0.45    |0.08     |0.29    |0.61    |1.00|1293    |1160    |
|Abuse                   |0.84    |0.07     |0.70    |0.97    |1.00|1182    |1200    |
```

Table (4) - Bayesian Model Summary

Upon setting our model, we would like to check if the model chosen is applicable. We would first check whether an evidence for non-convergence for the two chains exists.

(4)

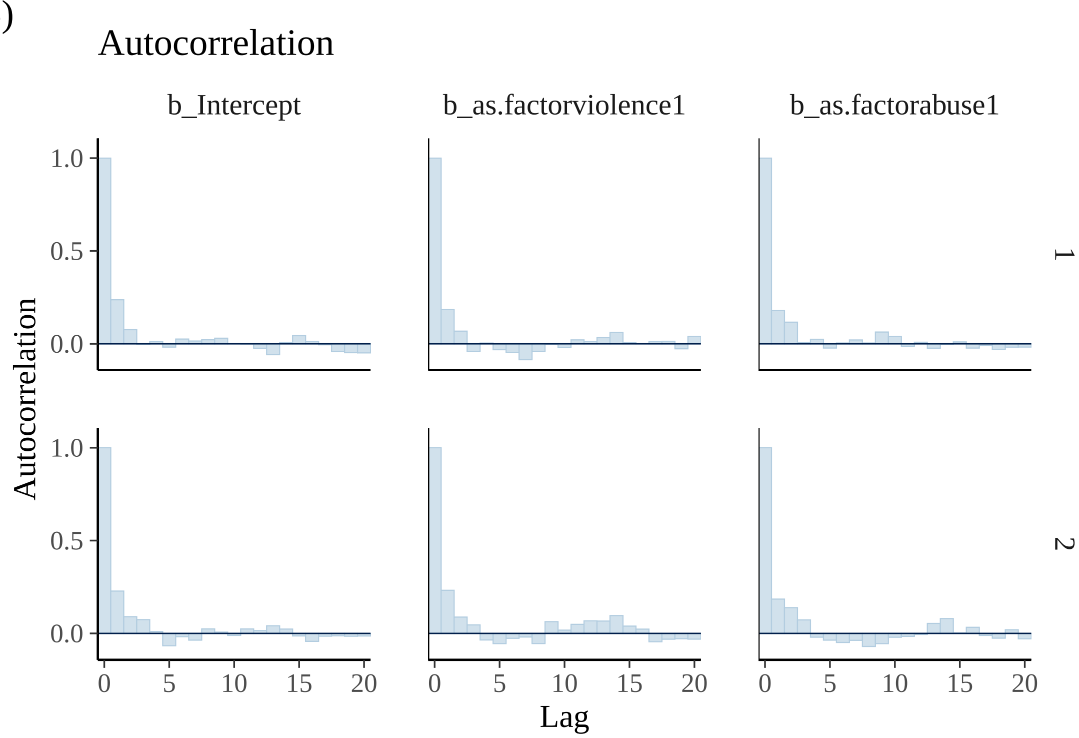


Figure(4) - Model Convergence

The two chains mix well in all the parameters in figure (4), making it possible for us to conclude that the two chains do not diverge.

We can also look at the autocorrelation plot in figure (5), in which the chain's correlation with successive lags of the chain are plotted. A strong autocorrelation would produce bias in variance estimates.

(5)



Figure(5) - Autocorrelation

The autocorrelation parameters for both chains starts at 1, then quickly drops down to around 0, meaning that there is no evidence suggesting autocorrelation exists. Below is a quick summary of the Bayesian binary model. We have also used Confusion Matrix to calculate the correct classification rate, which is 96.56%. A high correct classification rate suggests that the model is well fit for the data. However, upon further examination of the confusion matrix, we have found that the model predicts all the observations to be 0. Table (5) gives an explanation of such occurrence. Given that majority of the population was not victimized, the model is simply suggesting that no observations will bee victimized. Our data source, the survey(GSS 2014), asked about victimization excluded crimes committed by (ex)spouse/partner. Thus the response is limited. We would observe more people experiencing violent crimes if we had access to a full data, and our result may vary.

	0	1
0	30582	1090

Table(5) - Victimization

Discussion and Result

Data

The variables we used is from Canadian General Social Survey 2014. Upon clearing the data to apply Bayesian regression model, we have made some changes to the observations. First of all, the response variable "serious incident" is categorized into "violent incident" and "non-violent incident", which respectively have the value of 1 and 0. Then we have identified the explanatory variable for our issue of discussion, which is "witnessed violence between parents" and "experienced childhood assault". The two variables covers two types of maltreatment in childhood.

Result

From the Bayesian binary logistic regression model we have applied, figure (6) and figure (7) contains the posterior probabilities given various childhood experiences. Figure(6) shows the densities of parameter estimates, where the vertical line shows its point estimate, and light blue area is the 95% confidence interval. Both violence between parents and experienced childhood assault are explanatory toward the response since their credible interval does not include 0. And the estimators positively predicts the probability of a person being the victim of violent crimes as their point estimate and point interval are both positive. More specifically, witnessing violence between parents as a child increase the probability of being criminally victimized by 45%, and experiencing either physical or sexual assault will increase the probability by 83%. The curves in figure(6) have very narrow shapes, meaning that the confidence interval is relatively small and that there is 95% of chance that the population falls within.

Figure (7) further demonstrates the probability of a person experiencing violent crime. For an adult who had never experienced childhood maltreatment, the probability of them being a victim of violent crime is around 0.02. In contrary, an adult who was both a victim of childhood abuse and a witness of spousal violence, has a probability of 0.08 for being violently victimized. In general, both explanatory variables increases the probability of being violently victimized. However, the impact of being a direct victim of violence in childhood is more significant compared to witnessing violence between parents. The two width in figure (7) represents both 95% credibility intervals(thick line) and 68% credibility intervals(thin line). This gives us a rough idea of where the estimate lies.

The findings in our study is significant for reevaluating the impact of childhood maltreatment. Over the years, studies has been done about childhood experience as a risk factor for criminal activities in the future. However, little has been done addressing childhood abuse victims are likely to become victims of violent crimes as an adult.

(6)

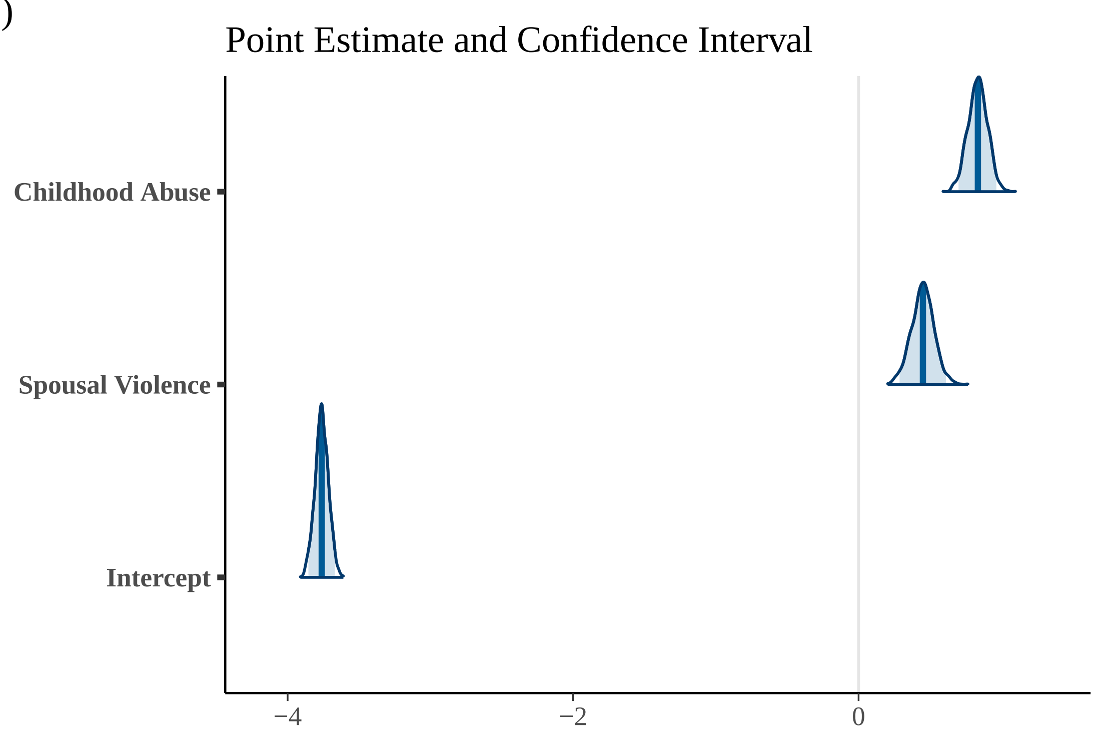


Figure (6) - Point estimate and confidence interval

(7)

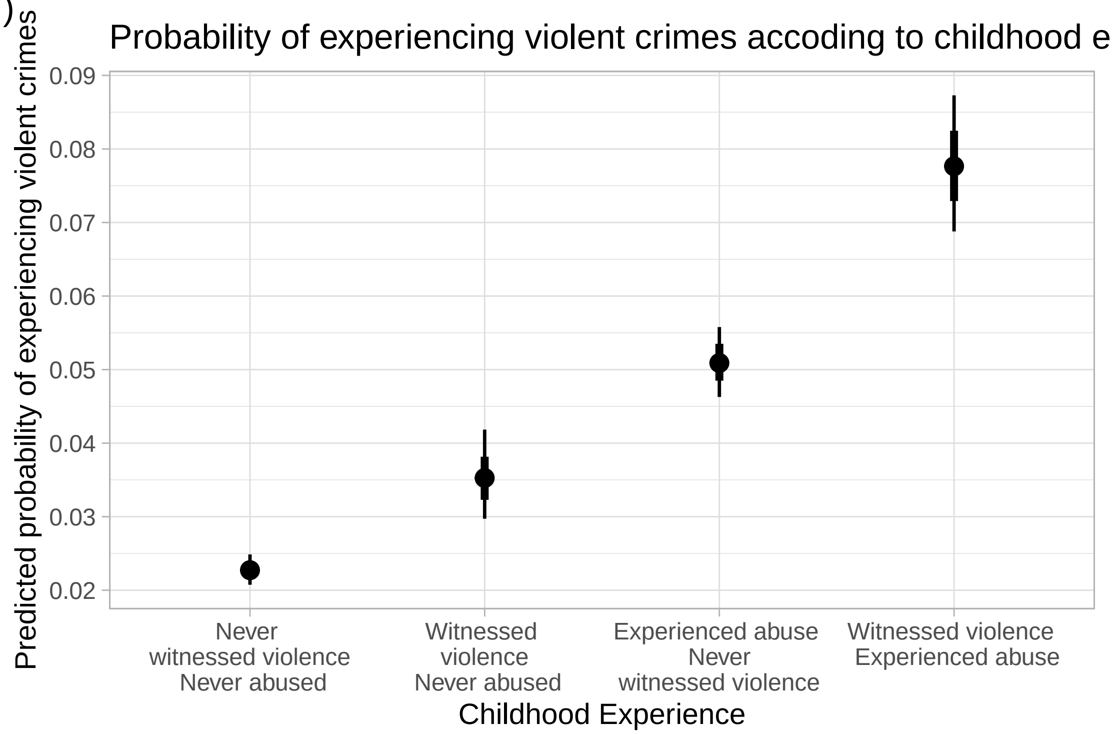


Figure (7) - Estimated probabilities of experiencing violent crimes

Weakness and Future studies

In the report, we have concluded that victimization in childhood increases the probabilities of being victimized towards violent crimes as an adult. However, we have encountered three main errors that might result in possible bias. The General Social Survey, to begin with, has its limitations. Regardless of the error due to non-response, the refusal of some specific questions may also cause errors. The variables we have used, are all personal and sensitive information that some individuals may not feel comfortable discussing. We could assume a significant proportion of refusals is counted towards positive response. As a result, the distribution of our variables could be altered. Secondly, our model design is by no means exhaustive. We have used Bayesian binary logistic regression model to predict response, yet our prior is non-informative. Kass and Wasserman (1996) stated: "Non-informative priors are formal representations of ignorance"(stat columbia). As we have little to know information of the priors, we have assumed it to be uniform so that it will have little to no impact on the posteriors distribution. Thus our result came largely from the dataset itself, making the model seemingly useless. Moreover, our conclusion merely indicates that childhood victimization increase the probability of experiencing violent crimes. Our result cannot explain the reason behind such phenomenon.

In future studies, we can further examine the model and develop a prior distribution that would be a better fit for the data. Also in a social aspect, it is important to find out the reason behind our findings. We could have more survey questions related to childhood victimization aspect and determine what directly caused such issue.

Citations

(1) JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3.9. URL https://rmarkdown.rstudio.com.

(2) Yihui Xie and J.J. Allaire and Garrett Grolemund (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC. ISBN 9781138359338. URL https://bookdown.org/yihui/rmarkdown.

(3) Yihui Xie and Christophe Dervieux and Emily Riederer (2020). R Markdown Cookbook. Chapman and Hall/CRC. ISBN 9780367563837. URL https://bookdown.org/yihui/rmarkdown-cookbook.

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

(5) Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

(6) Hadley Wickham, Jim Hester and Winston Chang (2020). devtools: Tools to Make Developing R Packages Easier. R package version 2.3.2. https://CRAN.R-project.org/package=devtools

(7) Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01

(8) Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017

(9) Hadley Wickham (2020). modelr: Modelling Functions that Work with the Pipe. R package version 0.1.8. https://CRAN.R-project.org/package=modelr

(10) Kay M (2020). tidybayes: Tidy Data and Geoms for Bayesian Models. doi: 10.5281/zenodo.1308151 (URL: https://doi.org/10.5281/zenodo.1308151), R package version 2.1.1, <URL: http://mjskay.github.io/tidybayes/>.

(11) Hadley Wickham (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr

(12) Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). "ROCR: visualizing classifier performance in R." Bioinformatics, 21(20), 7881. <URL: http://rocr.bioinf.mpi-sb.mpg.de>.

(13) Fang, Qixiang, and Rens van de Schoot. "Generalised Linear Models with Brms." Rens van de Schoot, 4 Oct. 2019, www.rensvandeschoot.com/tutorials/generalised-linear-models-with-brms. Accessed 20 Oct. 2020.

(14) Government of Canada, Statistics Canada. "General Social Survey - Victimization (GSS)." Www23.Statcan.Gc.Ca, 8 Jan. 2014, www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=148641#a3. Accessed 20 Oct. 2020.

(15) Government of Canada, Statistics Canada. "The General Social Survey: An Overview." Www150.Statcan.Gc.Ca, 20 Feb. 2019, www150.statcan.gc.ca/n1/en/catalogue/89F0115X. Accessed 20 Oct. 2020.

(16) "GSS Cycle 28, 2014." Login.Library.Utoronto.Ca, sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gs
Accessed 20 Oct. 2020.

(17) Perreault, Samuel. Criminal Victimization in Canada, 2014. 23 Nov. 2015.

(18) SYVERSVEEN, ANNE RANDI. NONINFORMATIVE BAYESIAN PRIORS INTERPRETATION AND PROBLEMS WITH CONSTRUCTION AND APPLICATIONS. www.ime.unicamp.br/~veronica/MI402/Randi219
Accessed 19 Oct. 2020.

(19) Collet, D. (1994). Modelling Binary Data. Chapman & Hall, London

(20) Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra

(21) Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

(22) Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

(23) Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Appendix

[1]"Only specific geographies were targeted for the oversample of immigrants and youth. For the oversample of immigrants, only the Census Metropolitan Areas of Montreal, Toronto and Vancouver were considered as separate strata. The remaining geographic areas of the ten provinces were grouped together to form the last stratum. For the oversample of youth, only the Census Metropolitan Areas of Halifax, Montreal, Ottawa, Toronto, Winnipeg, Calgary, Edmonton and Vancouver were covered by the frame and considered as separate strata. Also, for each geographic area, two strata were formed: one with the households including at least one immigrant between the age of 15 and 24 and one with the households including at least one nonimmigrant between the age of 15 and 24." (Statistics Canada)

[2] "Separate frames were created for the oversample of immigrants and for the oversample of youth. For the oversample of immigrants, the survey frame created for the regular sample was used first. A flag from an administrative source identifying the households with at least one immigrant was then added and only households flagged as having at least one immigrant were kept on the frame. The same principle was used to create the frame for the oversample of youth, but in this case, the flag identified households with at least one person between the age of 15 and 24 years old. This frame was also stratified to separate households with at least one immigrant between the age of 15 and 24 years old from the households with at least one non-immigrant (born in Canada) between the ages of 15 and 24 years old." (Statistics Canada)

[3] Link to code: https://github.com/mackenziequ/regression/blob/main/r%20code