A statistical study on effects of factors on fertility intention

Lan Cheng, Liyuan Cao, Shabier Zulihumaer

2020-10-19

Abstract

This study mainly applied a logistic model to investigate effects of various factors on couples' fertility intention based 2017 Canadian General Social Survey (GSS). It was found young couples with high life feeling score and few total number of children have more willingness to have children in future. For individuals, single males have good body and mental health status with high education level above bachelor and work less than 30 hours per week have higher fertility intentions. However, it was also found individuals with rented house and low family income tend to be more willing to have children in future. The findings are important because it described some key characteristics of groups of people who have high fertility intention as well as groups of people who have low fertility intention, and this could help policymakers to improve the conditions and fertility willingness of people in dealing with delayed childbearing problem better.

Introduction

Delayed childbearing becomes one of the critical characteristics of fertility trends in recent years. The problem is especially more severe in developed countries than others. It is essential to investigate fertility intentions because it plays a crucial role in explaining fertility trends. And various factors could affect people's fertility intentions.

Kariman N et al. (2016) pointed out in their study that these factors mainly including personal factors, family factors and social factors. For personal factors, the most important ones are age at marriage and quality of life. For family factors, the most important one is marital satisfaction, and for social factors, social support is most important. Roberts, E. (2011) claimed that factors like financially secure, interests of parents are essential in determining fertility intention. Testa, Maria Rita (2014) claimed young people are more willing to have children, based on a study of students, about 77% of them wanted to have children; the study pointed outage is a crucial factor in determining fertility intention.

Under this background, this study is also aimed to investigate the problem of the effects of factors on fertility intention. However, this study examined a more comprehensive range of factors, including age, gender, income, health status, education level, work status, marital status, and so forth. This study mainly applied logistic regression to investigate the effects of these factors on the binary outcome response fertility intention to have children in future or not. And based on the inferences from the logistic model, this study described the key characteristics of groups of people who have high fertility intention as well as groups of people who have low fertility intention. It is hoped that the findings of the study could be helpful for policymakers to improve the conditions and fertility willingness of people.

The structure of the report is organized as following: Firstly, an introduction of whole study was given. Then the data used in this study was discussed followed by a description of logistic model used in the model section. At last, the main results and discussions are presented. A link to the source of the report file could be found in https://github.com/tong304/GSS/tonga3.pdf.

Data

The data used in this study comes from the Canadian General Social Survey (GSS), Cycle 31, 2017. The 2017 GSS data includes lots of observations and attributes in various aspects. And this study used a cleaned version provided by Rohan Alexander and Sam Caetano (2020). We did not use all of the data; instead, we chose some of the interesting ones. After variables selection and data cleaning, the final data used includes 6475 observations with 12 variables. We described in details as following: 1. age - age in years; 2.totalchildren - total number of children; 3.lifefeeling - life feeling score, scaled 1-10; 4.sex - gender, female and male: 5.education - education level, above or below bachelor degree; 6.health - body health status, good or poor; 7.mentalhealth - mental health status, good or poor;8.house - owned or rented;9.evermarried - single or married;10. work - average hours worked, less or more than 30 hours; 11.familyincome - lower or higher than 74,999 dollars; 12.intention - response variable, 1 = intention to have children in future, 0 = intention not to have children in future.

For the questionnaire of the survey, it is good because lots of significant effort was made to minimize bias. The questionnaire used was a well-tested one with strict quality control. However, it is terrible to have too many questions in the questionnaire, which leads to lots of non-responses.

For the methodology of the survey, the target population of the 2017 GSS is all of the non-institutionalized people who are living in the ten provinces of Canada and no younger than 15 years old. The frame in the survey is the combination of Census' landline and cellular telephone numbers as well as various sources with dwelling frame. The original samples collected in the survey are over 40,000 units. The survey took a stratified two-stage sampling design, at the first stage, the sampling units are the groups of telephone numbers, and in the second stage, the sampling units are individuals in the household associated with the telephone number.

There are some of the trade-offs in the survey, and there are some non-responses in several stages. There might be non-responses at the household level and individual level. The overall response rate is about 52.4%. As there are lots of questions in the questionnaire, some questions could have non-answers such as the income has a low response rate. Because the other information is valuable, the survey deals with some non-responses by adjusting survey weights and apply estimates, for example, based on characteristics of non-response households, the 2016 census data was used to adjust non-responses. At last, because the 2017 GSS survey is a national wide one and it was created using lots of other linked sources such as census data, tax data, etc. So the cost is enormous; however, it is widely used in lots of areas, the survey is reliable, and it worths the price.

Model

The study mainly uses Logistic Regression to model the response of the intention to have children in future. Logistic Regression is a well-known generalized linear model that models the logit-transformed probability as a linear relationship with the interested factors. The form of the logistic model is as follows:

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

Or using the probabilities form:

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon)}{1 + exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon)}$$

where p is target probability, $x_1 - x_p$ are related p interested factors, ϵ is the error term. Unlike linear regression model, in logistic model, ϵ does not required to be normal distributed.

For this study, we chose the logistic regression model because the response intention to have children in future is a binary outcome 1 or 0 indicating yes or no which is suitable for using a logistic regression model.

Also, the logistic model is important because lots of factors interested in this survey and study could be well interpreted by the logistic model, so the logistic model could be applied to deal with the survey data.

For alternative models, because the response is binary, linear regression models are not appropriate. Probit model is appropriate for modeling binary outcome response, but it is harder to interpret the results other than using a logit link as logistic model. More advanced models such as bayes models are also not considered in this study mainly due to the large sample size which would cost a huge running time.

We used the R programming software to run a logistic model. In practice, there are few issues about model convergence, and in this study, the model convergence is also satisfied. For model checks and diagnostic issues, this study mainly applies formal statistic tests such as goodness of fit test as well as model diagnostics plots to verify it.

Results

The results of exploring descriptive analysis are shown in figures 1-3. The fitted logistic model is shown in tables 1-2. The model diagnostics are shown in figures 4-7.

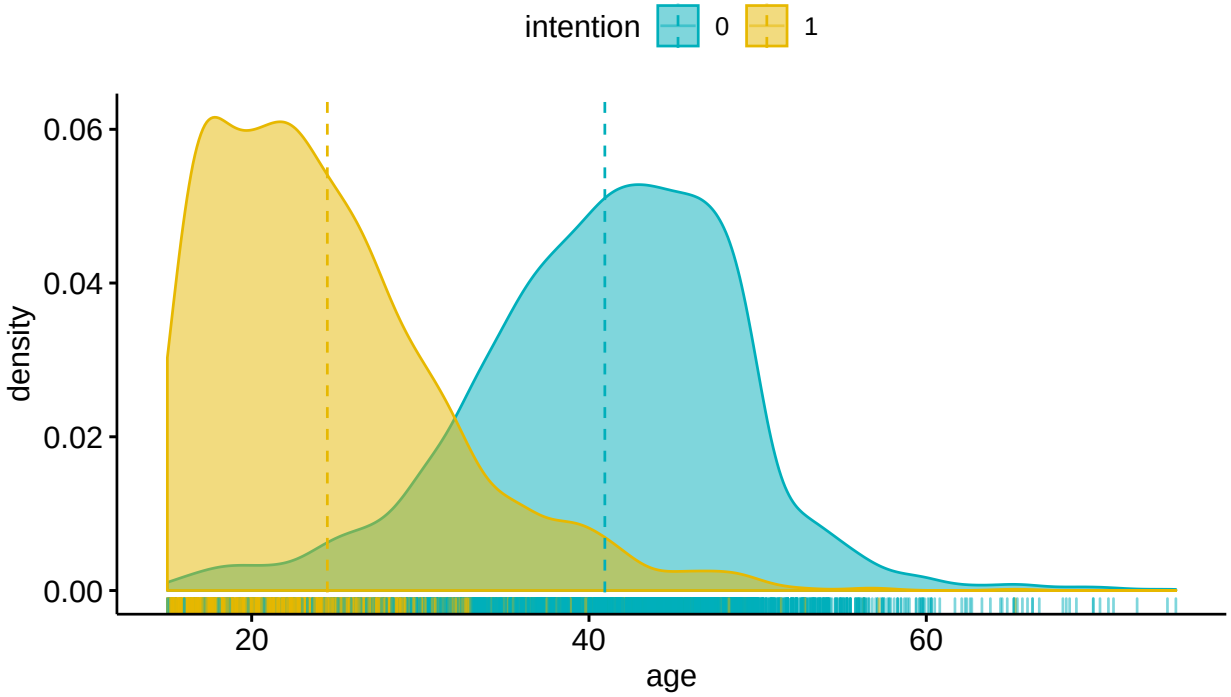

Figure 1: Density distributions of age grouped by whether intention to have children in future

Figure 1 illustrates the density distributions of age grouped by whether intention to have children in future. The graph clearly shows that the two groups have different density distributions, it indicates older people tends to be more likely not to have children in future while young people tend to be more likely to have children in future. This indicating age might be an essential factor in modelling the intention of having children in future.

Figure 2 illustrates the side-by-side boxplots of the total number of children and life feelings' score scaled 1-10 grouped by whether intention to have children in future respectively. The left panel of the figure clearly shows that the average level of the total number of children is much higher in the group which is intended not to have children in future than that of the group which is an intention to have children in future. However,

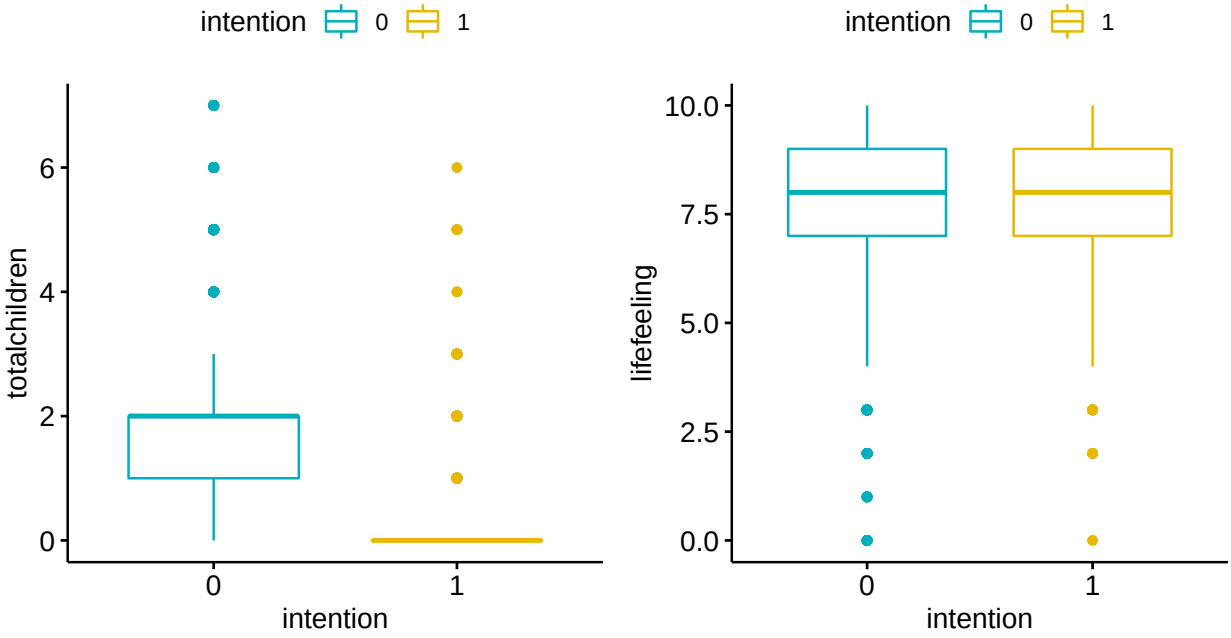

Figure 2: Side-by-side boxplots of total number of children and life feelings's score scaled 1-10 grouped by whether intention to have children in future

the right panel of the figure shows the average level of life feelings' score is very close between the two groups. This figure is indicating that the total number of children might be necessary for modelling the intention and having children in future while life feelings' score might not be significant.

Figure 3 illustrates the percentage of intention to have children in future grouped by various factors such as sex, education level, health level, income, etc. The plots show that there are differences in the percentage of intention to have children in future between most of the exciting factors except mental health level. So it means all of the stimulating factors except mental health level might help explain intention to have children in future.

Table 1 shows the logistic model estimates for various factors relating to the response intention of having children in future. All of the elements included in the model are significant at a 5% significance level as the p values are less than 0.05. As the regression model is a logistic model, table 2 shows the Odds-ratio of various factors based on the fitted logistic model.

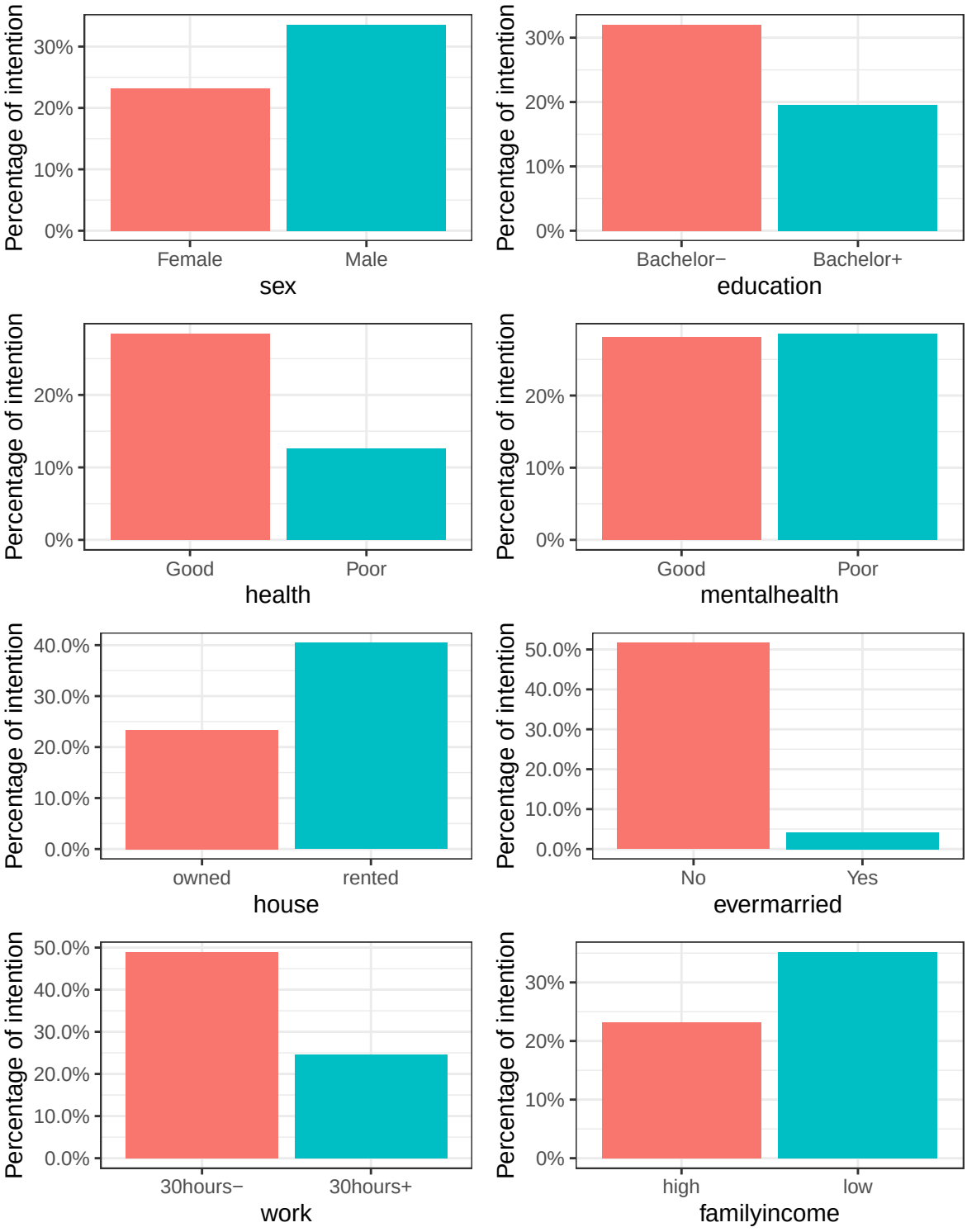

Figure 3: Percentage of intention to have children in future grouped by various factors

Table 1: Logistic model estimates for various factors relating to response intention of having children in future

	Dependent variable:
	intention
age	-0.173^{***}
	(0.007)
totalchildren	-0.886^{***}
	(0.065)
lifefeeling	0.154^{***}
	(0.031)
sexMale	0.816^{***}
	(0.094)
educationBachelor+	0.216^{**}
	(0.103)
healthPoor	-0.582^{*}
	(0.349)
mentalhealthPoor	-0.693^{**}
	(0.290)
houserented	0.317^{***}
	(0.102)
evermarriedYes	-0.716^{***}
	(0.132)
work30hours+	-0.270^{**}
	(0.129)
familyincomelow	0.175^{*}
	(0.100)
Constant	3.998^{***}
	(0.318)
Observations	6,475
Log Likelihood	$-1,608.725$
Akaike Inf. Crit.	3,241.449
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Odds-ratio of various factors based on fitted logistic model

variable	OR
age	0.84
totalchildren	0.41
lifefeeling	1.17
sexMale	2.26
educationBachelor+	1.24
healthPoor	0.56
mentalhealthPoor	0.50
houserented	1.37
evermarriedYes	0.49
work30hours+	0.76
familyincomelow	1.19

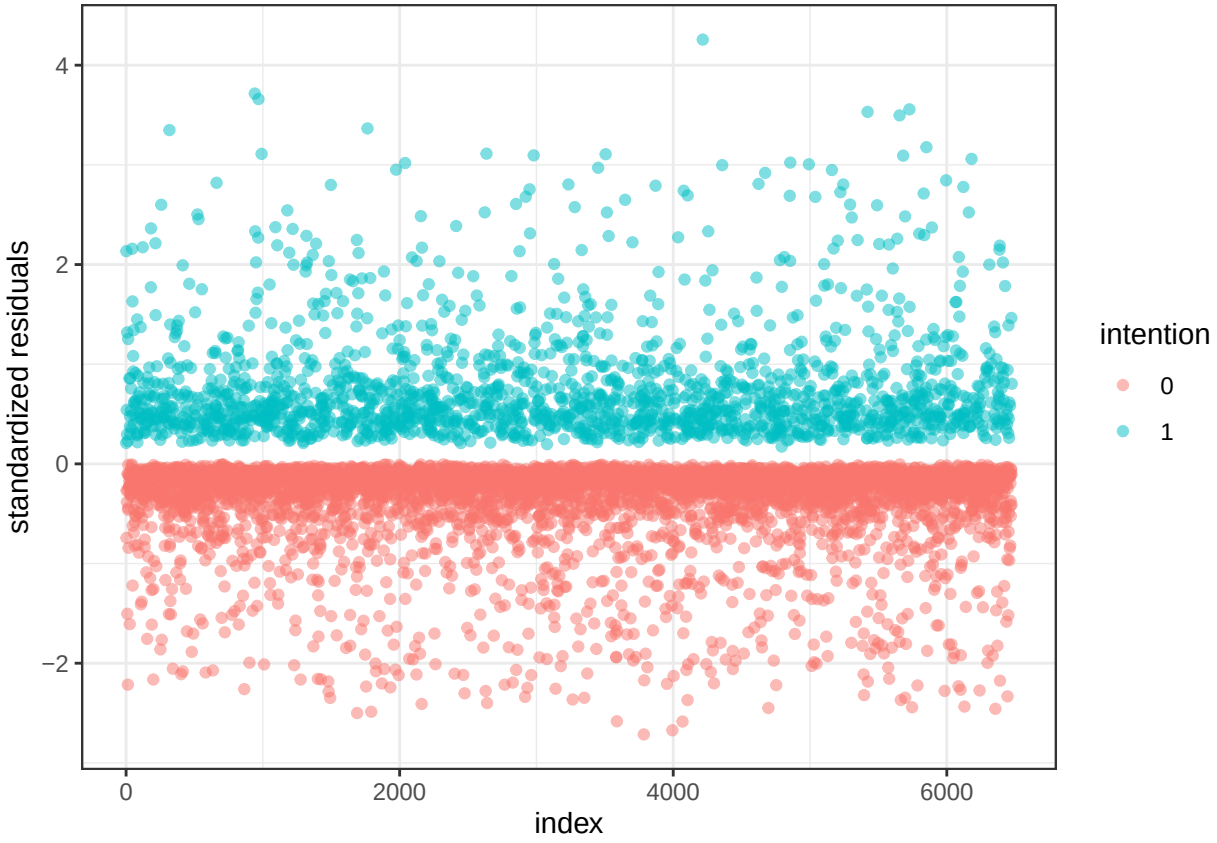

Figure 4: Model diagnostics standardized residuals plot of logistic model

Figure 4 illustrates the model diagnostics standardized residuals plot of the logistic model. Although the plot shows there might be some outliers with absolute standardized residuals larger than 2, there is no particular pattern in the plot indicating there is no big issue of the logistic model. The goodness of fit test shows a p-value close to 1, meaning there is no substantial evidence that the model lacks fit. The likelihood ratio test compared the fitted logistic model with a null model, and the test shows the fitted model is significant as the p-value is close to 0. More details of diagnostics could be found in the Appendix.

Discussion

This section mainly focuses on the effects of the exciting factors based on the estimated odds ratios in table 2; the interpretation of results is listed below:

1. The estimated odds ratio for age is 0.84 which suggests that for every unit increase in the age of years the odds of intention to have children in future would be about 16% lower holding all other factors constant.
2. The estimated odds ratio for the total number of children is 0.41 which suggests that for every unit increase in the total number of children the odds of intention to have children in future would be about 59% lower holding all other factors constant.
3. The estimated odds ratio for life feeling score is 1.17 which suggests that for every unit increase in the life feeling score the odds of intention to have children in future would be about 17% higher holding all other factors constant.
4. The estimated odds ratio for sex is 2.26 which suggests that the odds of intention to have children in future for male are about 126% higher than the odds for female holding all other factors constant.
5. The estimated odds ratio for education level is 1.24 which suggests that the odds of intention to have children in future for education level above bachelor are about 24% higher than the odds for education level under bachelor holding all other factors constant.
6. The estimated odds ratio for health level is 0.56 which suggests that the odds of intention to have children in future for poor health level are about 44% lower than the odds for good health level holding all other factors constant.
7. The estimated odds ratio for mental health level is 0.50 which suggests that the odds of intention to have children in future for poor mental health level are about 50% lower than the odds for good mental health level holding all other factors constant.
8. The estimated odds ratio for "house" is 1.37 which suggests that the odds of intention to have children in future for people with the rented house are about 37% higher than the odds for people with owned house holding all other factors constant.
9. The estimated odds ratio for "ever married" is 0.49 which suggests that the odds of intention to have children in future for people who have ever married are about 51% lower than the odds for people who have not ever married holding all other factors constant.
10. The estimated odds ratio for work is 0.76 which suggests that the odds of intention to have children in future for people who have above 30 hours work are about 24% lower than the odds for people who have less than 30 hours work holding all other factors constant.
11. At last, the estimated odds ratio for family income is 1.19 which suggests that the odds of intention to have children in future for people who have low family income are about 19% higher than the odds for people who have high family income holding all other factors constant.

If the above factors are combined together, then several groups could be found related with willingness to have children in future. For example, the group of young couples with high life feeling score and few total number of children have more willingness to have children in future compared with the group of old couples with low life feeling score and more total number of children. For individuals, the group of single males have good body and mental health status with high education level above bachelor and work less than 30 hours per week has higher fertility intentions than the group of single females who have poor body and mental health status with low education level and work more than 30 hours per week. Also, there is a group of individuals with rented house and low family income tend to be more willing to have children in future compared with individuals with owened house and high family income.

The findings are very important because policymakers are faced with a serious delayed childbearing problem, they need to understand the situations and conditions of people to make better decisions to encourage those

couples with low fertility intention to be more willingness to have children in future. The findings of this study described those groups which might be very helpful for the policymakers. Although the findings are based on the Canadian General Social Survey, this study could be performed similarly for other countries in the world.

Finally, there are some weaknesses in this study. First, as the study is based on the 2017 Canadian General Social Survey data, and one of the interested factors - family income used in this study could be seriously biased because there were lots of non-responses in the question family income in the survey and most of the information of family income in the 2017 GSS data were estimated and adjusted not orignally responses. This could introduce a serious biasness problem into the logistic model in this study. Second, this study used only about 30% of the cleaned GSS data, there might be also biasness in the results obtained due to smaller sample size. At last, in this study, only a logistic model was investigated limitted to interested factors, there might be other important factors, also, the model only included first order terms, there were no interaction terms or high order terms. Thus, in future work, a more reliable data set with a more wider ranger of factors as well as models could be considered to improve the results obtained in this study.

References

1. Alboukadel Kassambara (2019). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2.4. https://CRAN.R-project.org/package=ggpubr
2. Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-1. https://CRAN.R-project.org/package=arm
3. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
4. David Robinson and Alex Hayes (2020). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.4. https://CRAN.R-project.org/package=broom
5. Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr
6. Hadley Wickham (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
7. Hadley Wickham, Romain Franois, Lionel Henry and Kirill Muller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.Rproject.org/package=dplyr
8. Hadley Wickham and Dana Seidel (2019). scales: Scale Functions for Visualization. R package version 1.1.0. https://CRAN.R-project.org/package=scales
9. Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
10. Julian Faraway (2016). faraway: Functions and Datasets for Books by Julian Faraway. R package version 1.0.7. https://CRAN.R-project.org/package=faraway
11. Kariman N, Amerian M, Jannati P, Salmani F. Factors influencing first childbearing timing decisions among men: path analysis. Int J Reprod BioMed. (2016);14(9):589-596. doi: 10.29252/ijrm.14.9.589
12. 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/.
13. Roberts, E. , Metcalfe, A. , Jack, M. , & Tough, S. C. (2011). Factors that influence the childbearing intentions of Canadian men. Human Reproduction, 26(5), 1202-1208. 10.1093/humrep/der007.
14. Rohan Alexander and Sam Caetano (2020). 2017 GSS data cleaning source code gss_cleaning.R (Version 1.0).

15. Statistic Canada(2019). General Social Survey - Family (GSS), Detailed information for 2017 (Cycle 31). https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816

16. Testa, Maria Rita. (2014). "On the positive correlation between education and fertility intentions in Europe: Individual- and country-level evidence," Advances in Life Course Research 21: 2– 42.

17. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.27.

Appendix

The Github repo link of source codes in a rmd format is: https://github.com/tong304/GSS.

Figures 5-7 shows more diagnostics plots for the fitted logistic model. Figure 5 shows the binned residuals plot of the logistic model. Figure 6 shows the Halfnorm plot of the logistic model. Figure 7 shows Cook's Distance plot of the logistic model. Figure 5 shows there are many points outside the bins indicating the model needed to be improved. Figure 6 indicates there are some outliers, while Figure 7 means there are some strong influence points with high Cook's distance.

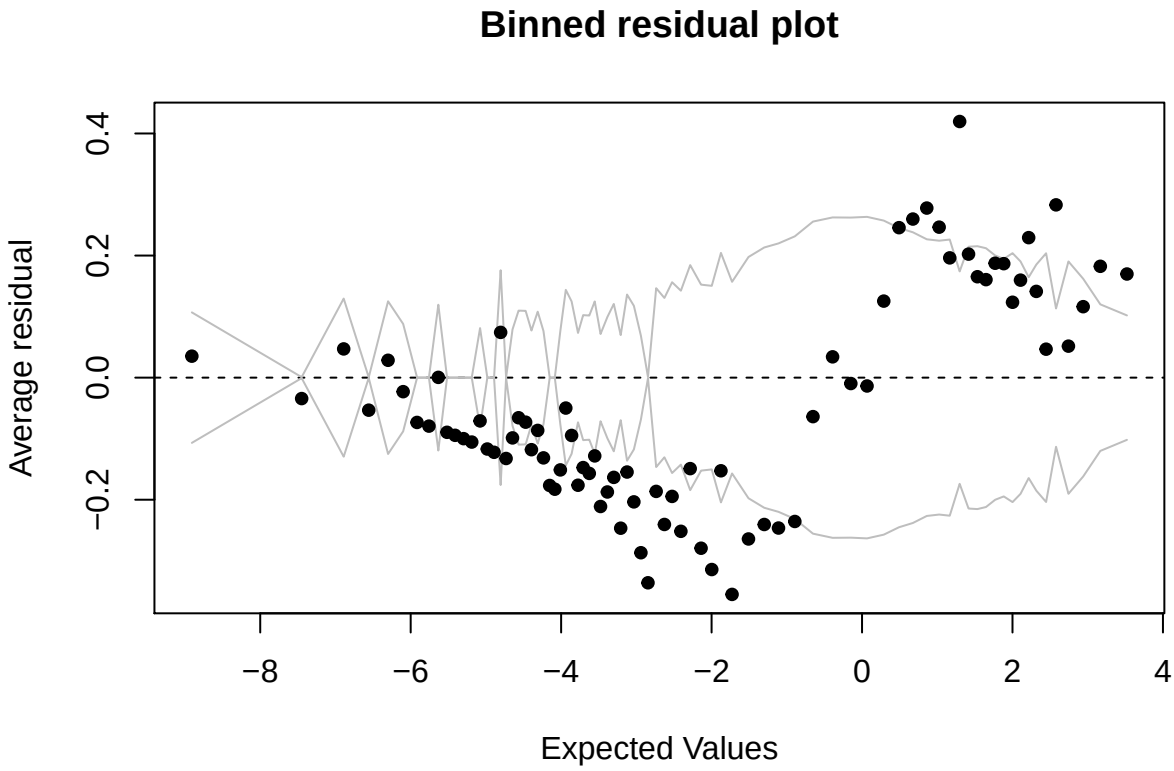

Figure 5: Binned residuals plot of logistic model

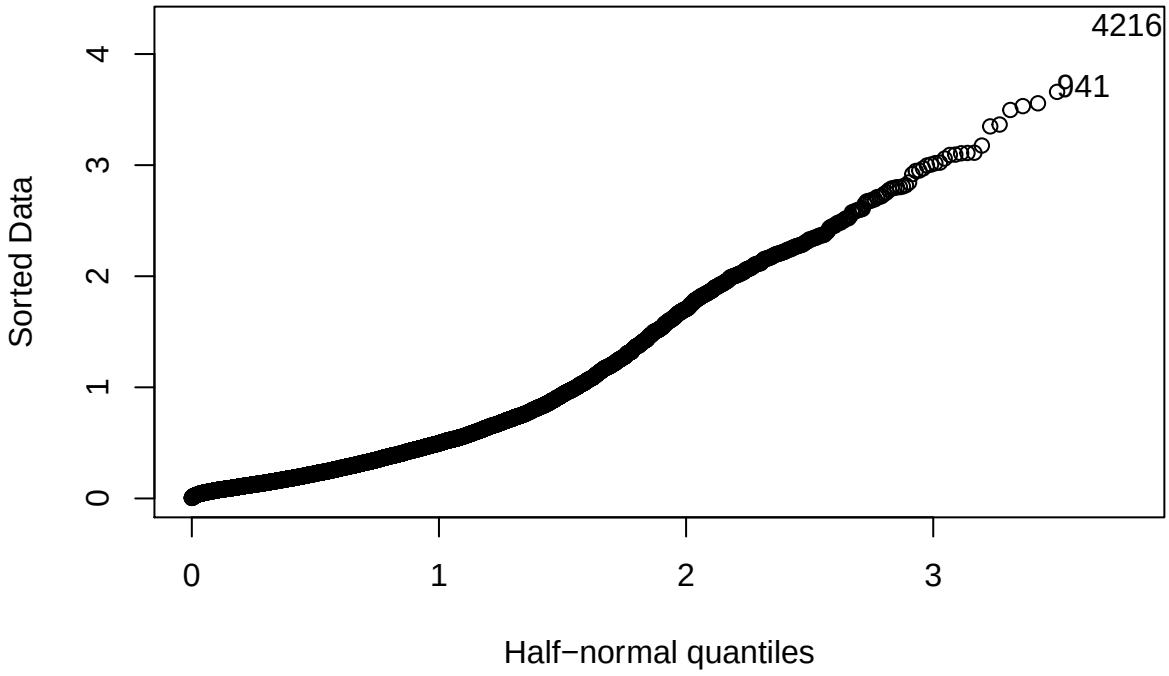

Figure 6: Halfnorm plot of logistic model

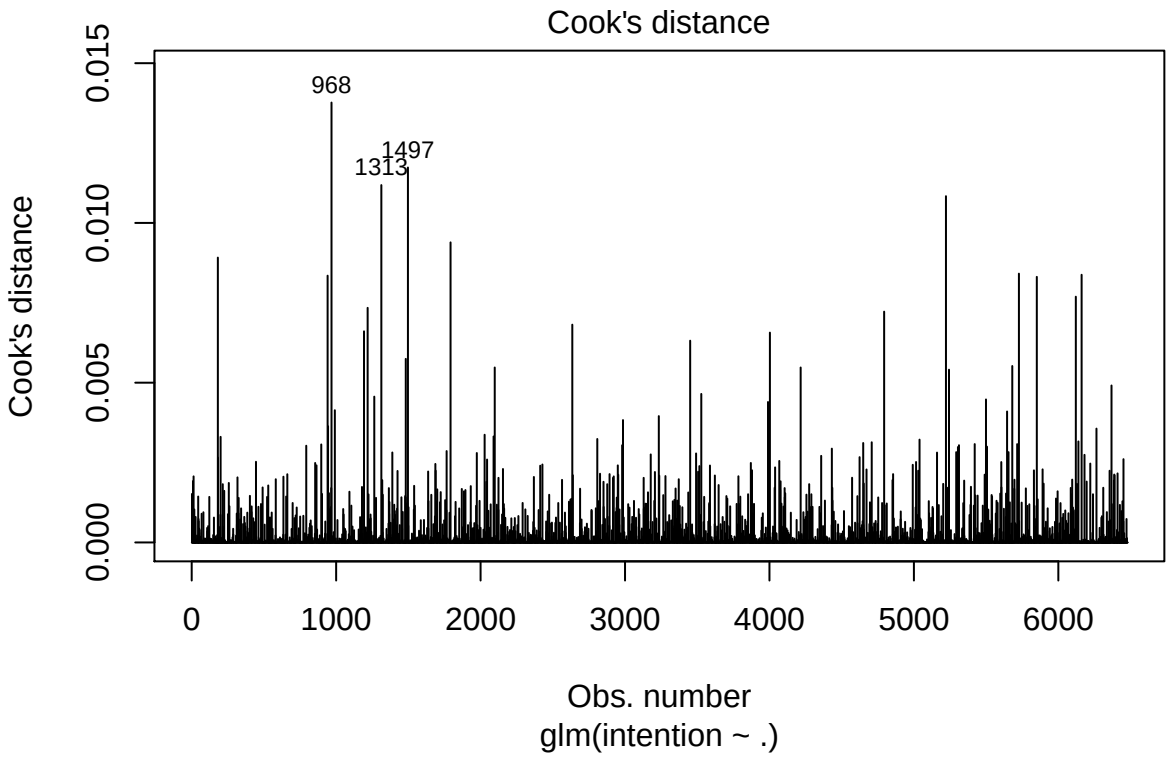

Figure 7: Cook's distance plot of logistic model