Investigation of the Fertility Rate in Canada

Ying Xiong(1004795885),Yixin Liang(1005549998),Huiyi Lu(1004754615),Deyu Meng(1004739991)

10/19/2020

Contents

1 Abstract **2**

2 Introduction **2**

3 Data **2**
3.1 Survey Data . . . 2
3.2 Methodology in Data . . . 3
3.3 Strength & Weakness (Data) . . . 3
3.4 Cleaned Raw Dataset . . . 3
3.5 Data for Modelling . . . 3

4 Model **4**
4.1 Final Model . . . 4
4.2 Analysis on Predictors . . . 5
4.3 Model Diagnostics . . . 5
4.4 Alternative Models . . . 5

5 Results **6**
5.1 Summary Table (Numerical Variables) . . . 6
5.2 Summary Table (Model) . . . 6
5.3 Bar plots . . . 8
5.4 Box plots . . . 9

6 Discussion **10**
6.1 Data set . . . 10
6.2 Data Result Explanation . . . 10
6.3 Model Result Explanation . . . 10
6.4 Weaknesses and Caveat . . . 11
6.5 Next Steps . . . 11

7 Appendix **12**
7.1 Data - Cleaned Raw Dataset . . . 12
7.2 Model - Linearity Assumption . . . 13
7.3 Model - Influential Observations . . . 13
7.4 Model - Multicollinearity . . . 13
7.5 Alternative model . . . 14

8 References **16**

1 Abstract

The diminishing fertility rate is a serious issue that should be emphasized since the consequent aging and shrinking population problem is detrimental to the overall development of society. In this study, we retrieved, processed, and applied the logistic regression model on the data from the 2017 General Social Survey (GSS) on the Family. In particular, we analyzed five predictor variables (family income, education level, the prosperity of residence, age, and feelings of life) that would potentially affect the fertility rate. From the model, we can predict the probability of a person having a fertility rate higher than the replacement level and conclude that family income, age, and feelings of life have a positive impact while the level of education and the prosperity of residence have a negative influence on the fertility rate.

2 Introduction

Over the past 150 years, with society's rapid development, the baby boom has gradually become history. Conversely, the low fertility rate becomes an emerging issue in Canada, bringing potential pernicious problems. Canada's fertility rate has dropped to 1.509 in 2020 (Macrotrends), which is far from the replacement rate, 2.1 (Statistics Canada, 2018).[1] In this way, the population size will shrink, with more elders and fewer younger appearing in the demographic trend. Aging society forces the government to be more reliant on immigration. Therefore, the variables, which cause the fertility rate decreasing problem, are vital to explore to maintain a sustained population.

This study aims to generate a model that can predict the probability of a person having a fertility rate higher than the replacement level. From the model, we will explore how the variables affect the fertility rate. The result of the study will provide a guideline to the government on possible aspects that may increase the average birth rate, which ultimately will lead to a sustainable population. In the study, we will investigate and focus on five potential elements: family income, education level, the prosperity of residence, age, and feelings of life.

The report will provide a reflective analysis through different sections. In the data section, we will analyze the data set (2017 General Social Survey) in aspects of the questionnaire and the methodology. We will present the cleaned data we used and our final model with discussions in the model section and the result section. Finally, thorough and constructive explanations of the outcomes will be covered in the discussion section. Supplemental information is included in the Appendix for the reader's best understanding.

3 Data

3.1 Survey Data

The data used throughout this paper is retrieved from the 2017 General Social Survey (GSS) on the Family via the CHASS data center. The GSS was conducted from Feb. 2 to Nov. 30 in the year of 2017 through telephone interviews with its selected respondents.

The survey data set of 20602 observations and 461 variables are collected from non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. It provides information on specific social policy issues of current or emerging interests and helps monitor changes in the living conditions and well-being of Canadians. Variables can be grouped into core content and classification variables. Core content measures change in society related to living conditions and well-being, while classification variables help define population groups for the analysis of core content.

The **target population** of the survey is all non-institutionalized people who are older than 15 and live in 10 Canadian provinces. The **frame population** is people on the list of telephone numbers collected in 2013 by Statistics Canada and the Address Register, while the **sample population** is people who answered the calls.

[1] The replacement rate stands for the replacement-level fertility rate. A replacement rate of 2.1 children per woman implies couples had enough babies to replace themselves on average in the current population. (Statistics Canada, 2018)

3.2 Methodology in Data

The survey uses **stratified sampling** and **simple random sampling without replacement** within each stratum. It sets strata based on provinces, while more densely-populated areas are considered as separate strata. With geographical stratification, 27 strata are formed in total. For each eligible household in each stratum, a respondent is randomly selected for the interview. The stratified sampling approach ensures that each stratum will be included in the sample. In the 2017 GSS survey, each subgroup reflects the population in a specific province in Canada. This increases the possibility of accurately representing the target population. However, the sampling approach is laborious since it requires research on population density. Unlike the stratified sampling approach, simple random sampling is easier to conduct and requires less effort. It also contains less subjective matters.

Non-responses are divided into 3 levels based on the amount of information received from them to adjust for the weight and calculate the response rate. Then the dataset drops all the non-responding phone numbers.

3.3 Strength & Weakness (Data)

- Strength: The data can be used to answer many family-related questions as families are becoming increasingly diverse in Canada. It is recognized to be useful and credible for trend analysis, and it is capable to test and develop new concepts. Most questions in the questionnaires are concise, and the choices of answers provided to the respondents are quite specific (i.e. a range is given).
- Weakness: Data collected from surveys are subjected to both sampling and non-sampling errors. The total number of questions is relatively large with some questions being stated in a somewhat general or vague manner. Respondents may have various interpretations and perceptions of them. The potential miscomprehension and the survey length may lead to failure in attaining the objectives of the designed questions as well as non-response. Also, there might be mistakes during the answer recording process. Thus, some variables that contain too many NAs cannot be contributed to the analysis and may cause "garbage in, garbage out".

3.4 Cleaned Raw Dataset

The cleaned raw data consists of 20602 observations and 81 variables. The majority of the data are categorical while some are numerical. There are a significant number of missing values under some of the variables, for example, all responses under main_activity are NAs. These variables are useless for the purpose of analysis. Therefore, the raw data is messy and indigestible. The data visualization of the cleaned raw data is included in the Appendix 1 (Figure 3).

3.5 Data for Modelling

From the cleaned raw data set, we selected 6 variables and constructed a new data set, *age*, *income_family*, *feelings_life*, *education*, *pop_center* and *total_children*. We chose to remove all the observations with NA in all the variables we used in our model, both the independent and dependent ones. The models with and without NAs removed for predictors are the same as long as the NA of the response variable is removed. We found that the model will remove those NA observations automatically and there is no point to keep those observations if the model does not consider them. Also, in further analysis, the model diagnostics require the same amount of observations for the data we used as well as the model used.

- Dependent variable: *child_factor* constructed from *total_children*, based on the replacement rate, 2.1. Families with less than 3 children are noted as 0 and the rest is noted as 1 under child_factor. This classification helps determine who exceeds the replacement rate and who does not reach the replacement rate.
- Independent variable: *age*, *income_family*, *feelings_life*, *edu_factor*, *pop_center* We constructed *edu_factor* by grouping the choices under *education*. Less than high school diploma or its equivalent is classified into Primary Education; High school diploma or a high school equivalency certificate is classified into Secondary Education; Trade certificate or diploma, and College, CEGEP or other

non-university certificate or diploma, and University certificate or diploma below the bachelor's level, and Bachelor's degree(e.g. B.A., B.Sc., LL.B.) are classified into Post-secondary Education; University certificate, diploma or degree above the bachelor's level is classified into Post-graduation Education.

The reason we chose *income_family* instead of the variable *income_respondent* is that we believe the decision of having babies is not upon individuals. It should be decided by all family members and consider the family's living conditions.

4 Model

4.1 Final Model

To model our binary response variable "child_factor", we applied a logistic regression model in R. The following mathematical notation is the final model we got for analysis.

$$\log \frac{p_i}{1-p_i} = -0.243 + 0.0126X_{incomefamily,1} - 0.0474X_{incomefamily,2} - 0.0437X_{incomefamily,3}$$
$$-0.0217X_{incomefamily,4} - 0.0526X_{incomefamily,5} + 0.0408X_{edufactor,1} + 0.1132X_{edufactor,2} + 0.0677X_{edufactor,3}$$
$$+ 0.0807X_{popcenter,1} + 0.0841X_{popcenter,2} + 0.00756X_{feelingslife} + 0.0067X_{age}$$

p_i: the probability of the respondent being above the average fertility rate

$X_{incomefamily,1}$: = 1 if the respondent's total family income before tax is over $125,000

$X_{incomefamily,2}$: = 1 if the respondent's total family income before tax is between $25,000 to $49,999

$X_{incomefamily,3}$: = 1 if the respondent's total family income before tax is between $50,000 to $74,999

$X_{incomefamily,4}$: = 1 if the respondent's total family income before tax is between $75,000 to $99,999

$X_{incomefamily,5}$: = 1 if the respondent's total family income before tax is less than $25,000

- When all the conditions above are not met, then the $X_{incomefamily,i}$ would stay at 0, meaning the respondent's total family income before tax is between $100,000 to $124,999.

$X_{edufactor,1}$: = 1 if the respondent's level of education is Post-secondary (College/University)

$X_{edufactor,2}$: = 1 if the respondent's level of education is Primary (Didn't finish high school)

$X_{edufactor,3}$: = 1 if the respondent's level of education is Secondary (Finished high school)

- When all the conditions above are not met, then the $X_{edufactor,i}$ would stay at 0, meaning the respondent's level of education is Post-graduation (After College/University Graduation) .

$X_{popcenter,1}$: = 1 if the respondent lives on the Prince Edward Island

$X_{popcenter,2}$: = 1 if the respondent lives in Rural Areas or in Small Population Centers

- When all the conditions above are not met, then the $X_{popcenter,i}$ variable would stay at 0, meaning the respondent lives in Larger Urban Population Centers.

$X_{feelingslife}$: the value from 1 to 10 that represents the respondent's feelings about life as a whole

X_{age}: the respondent's age

With information implied by the above variables known, we can effectively predict a person's probability of having a fertility rate above the replacement-level fertility rate. The variables separate the information into different categories that allow new respondent's information to be classified precisely and give a relative correct prediction.

4.2 Analysis on Predictors

For the independent variables we used in our logistic model, we chose factors that may have impacts on the fertility rate from our perspective for simplicity.

We included 1 numerical variable (*age*) and 4 categorical variables (*income_family*, *edu_factor*, *pop_center*, *feelings_life*).

We did not separate ages into different groups because we believe numerical variables can achieve better predictions. When groups are used, there would be larger variation since a range of different information will be identified as the same and applied with the same coefficient. Thus, we chose to stick with the numerical values for the variable *age*.

feelings_life is a quantitative predictor, but we will treat it as an ordinal categorical predictor because the values have different meanings applied to each number. The numbers themselves do not have much meaning.

For the other categorical predictors, there are no other substitutes or appropriate ways to adjust them into meaningful numerical variables. Thus, we will treat them as qualitative factors for the remaining discussion.

4.3 Model Diagnostics

1. Linearity Assumption

We would check the linear relationship between continuous predictor variables and the log odds. In our final model, *age* is the only valid numerical predictor.

From the scatter plot in the Appendix 2 (Figure 4), we can see that the variable *age* violates the linearity assumption and may need some transformation. For simplicity and completeness of the model, we would ignore the violation here and still include *age* as one of the predictors.

2. Influential Observations

As we can see the amount of observations for leverage data points, there are 831 leverage points that are extreme in their x-values that may potentially deviate the regression line. After we used the cook's distance to calculate the influential observations, we found there are none. Thus, we don't need to remove any observations in the data set. This is shown in the Appendix - Model - Influential Observations section.

3. Multicollinearity

We checked the multicollinearity between the predictors to reduce the effect on the inference. A moderate correlation may change the inference and result in wrong interpretations. We used the Variance Inflation Factor. As a rule of thumb, we should remove one of the independent variables from the model if two or more predictors are correlated. As shown in the table in the Appendix 3 (Table 3), the VIFs for each predictor is about 1, which does not meet the cutoff. We don't need to remove any variable from our logistic model.

4.4 Alternative Models

As we redefined the numerical variable "total_children" into a binary variable "child_factor" and used it as the response, a logistic regression model is the best to fit the data. Notably, our research topic is to generate a model that can predict the probability of a person having above-average fertility rate and discuss factors that may influence the probability. One may argue that a multiple linear regression model can reproduce the work above. Using the trial and error approach, we did build a MLR model that would predict the number of children a person may have with the same independent variables. However, it comes to our attention that the MLR model has significant violations in the model diagnostic tests. It failed to achieve both the linearity and normality assumptions, which implies the model produced is invalid. The MLR model and its model diagnostics are included in the Appendix - Model - Alternative Models.

5 Results

5.1 Summary Table (Numerical Variables)

- Table 1:

```
##
## ===========================================================================
## Statistic       Min   Pctl(25) Median Pctl(75)  Max    Mean  St. Dev.
## ---------------------------------------------------------------------------
## total_children   0       0       2       3       7    1.672   1.486
## age             15.000  37.300  54.100  66.600  80.000 52.107  17.732
## feelings_life    0       7       8       9       10   8.093   1.642
## ---------------------------------------------------------------------------
```

According to the summary table for the numerical variable (Table 1) , three numerical variables are analyzed, which are total children that the respondents have, age and the feeling of life.

Firstly, the mean of the total children's variable is 1.672, which indicates the average fertility rate of the sample. This is close to the Canadian fertility rate, which is 1.509. The difference in the numbers may be due to the sample selection and the bias from the survey. The maximum number of children that the respondent has is 7 children and the minimum is 0. Secondly, the mean age of the respondents is about 52 years old. The oldest is 80 years old. Therefore, older respondents' answers may increase this study's fertility rate because they were during the baby boom period. This may also explain why the response's fertility rate is slightly higher than the Canadian current birth rate. Finally, the mean and the median of the respondents' feelings of life are 8.903 and 9, respectively.

5.2 Summary Table (Model)

- Table 2:

Table 1: Summary of Logit Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.243	0.021	-11.599	0.000
income_family$125,000 and more	0.013	0.011	1.166	0.243
income_family$25,000 to $49,999	-0.047	0.011	-4.248	0.000
income_family$50,000 to $74,999	-0.044	0.011	-3.863	0.000
income_family$75,000 to $99,999	-0.022	0.012	-1.843	0.065
income_familyLess than $25,000	-0.053	0.012	-4.281	0.000
edu_factorPost-secondary Education	0.041	0.011	3.882	0.000
edu_factorPrimary Education	0.113	0.013	8.908	0.000
edu_factorSecondary Education	0.068	0.012	5.880	0.000
pop_centerPrince Edward Island	0.081	0.016	5.039	0.000
pop_centerRural areas and small population centres (non CMA/CA)	0.084	0.007	11.225	0.000
feelings_life	0.007	0.002	3.723	0.000
age	0.008	0.000	44.794	0.000

According to the summary table of logistic regression results (Table 2), families with income that are higher than 125000 and families with income that range from $75,000 to $99,999 are statistically insignificant by

using 0.05 as the significant level. This represents that the logit odds ratio for families in these two income groups of having more than two children is similar to those on average than whose family income is between $100,000 and $125,000. Other variables are statistically significant, with a P-value that is lower than 0.05.

- $X_{incomefamily,i}$

According to the model results, if the respondent's total family income before tax is over $125,000 and more, his logit odds ratio of having more than 2 children will be 0.0126 higher on average than whose family income is between $100,000 and $125,000, holding everything else constant. If the respondent's total family income before tax is between $25,000 and $49,999, his logit odds ratio will be 0.0474 lower on average than whose family income is between $100,000 and $125,000, holding everything else constant. If the respondent's total family income before tax is between $50,000 and $74,999, his logit odds ratio will be 0.0437 lower on average than whose family income is between $100,000 and $125,000, holding everything else constant. If the respondent's total family income before tax is between $75,000 and $99,999, his logit odds ratio will be 0.0217 lower on average than whose family income is between $100,000 and $125,000, holding everything else constant. If the respondent's total family income before tax is less than $25,000, his logit odds ratio will be 0.0526 lower on average than whose family income is between $100,000 and $125,000, holding everything else constant.

- $X_{edufactor,i}$

If the respondent's level of education is Post-secondary, his logit odds ratio of having more than 2 children will be 0.0408 higher on average than who has a Post-graduation degree, holding everything else constant. If the respondent only receives Primary education, his logit odds ratio will be 0.1132 higher on average than who has a Post-graduation degree, holding everything else constant. If the respondent only receives Secondary education, his logit odds ratio will be 0.0677 higher on average than who has a Post-graduation degree, holding everything else constant.

- $X_{popcenter,i}$

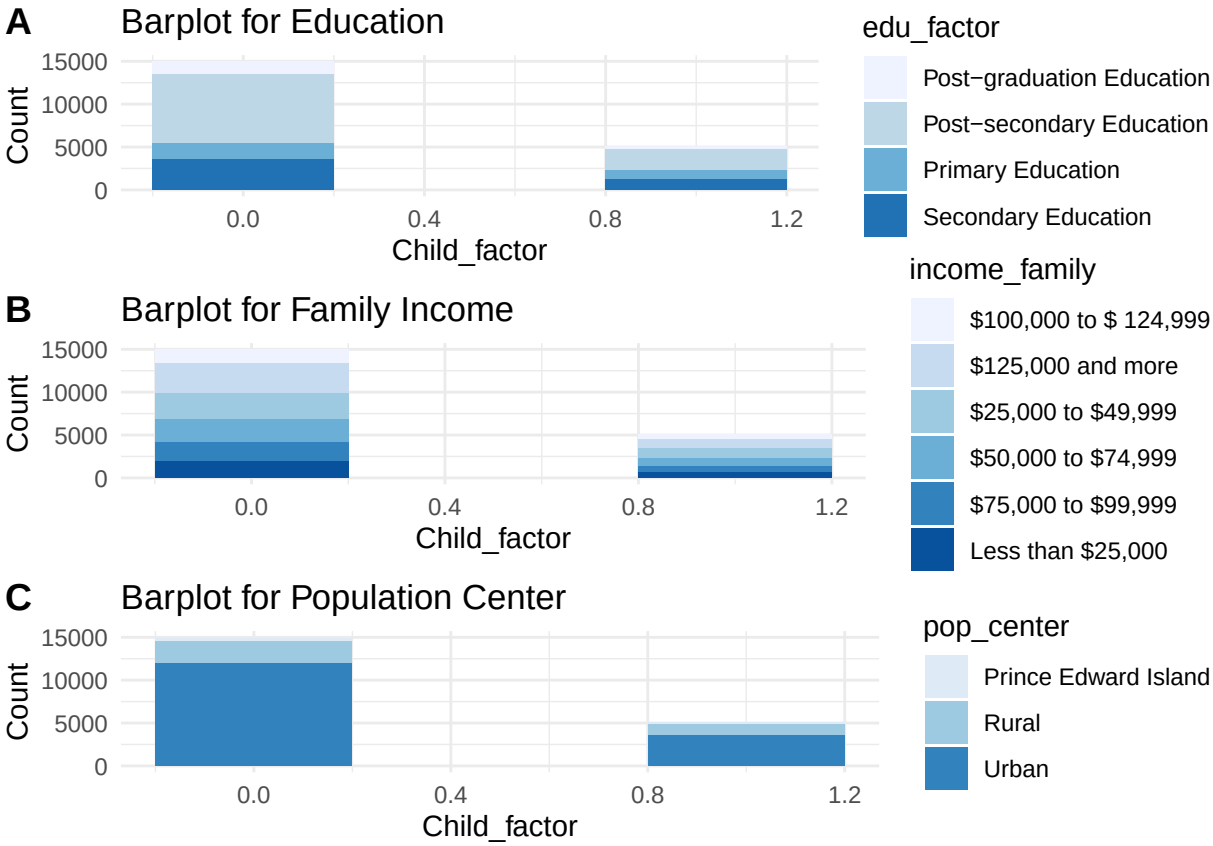
If the respondent lives on Prince Edward Island, his logit odds ratio of having more than 2 children will be 0.0807 higher on average than who lives in a larger urban population center, holding everything else constant. If the respondent lives in rural areas or in small population centers, his logit odds ratio will be 0.0841 higher on average than who lives in a larger urban population center, holding everything else constant.

- $X_{feelingslife}$ & X_{age}

Holding everything else constant, the average difference in logit odds ratio is 0.00756 and 0.0067 corresponding to the difference in rating of life and difference in age, respectively.

5.3 Bar plots

- Figure 1:

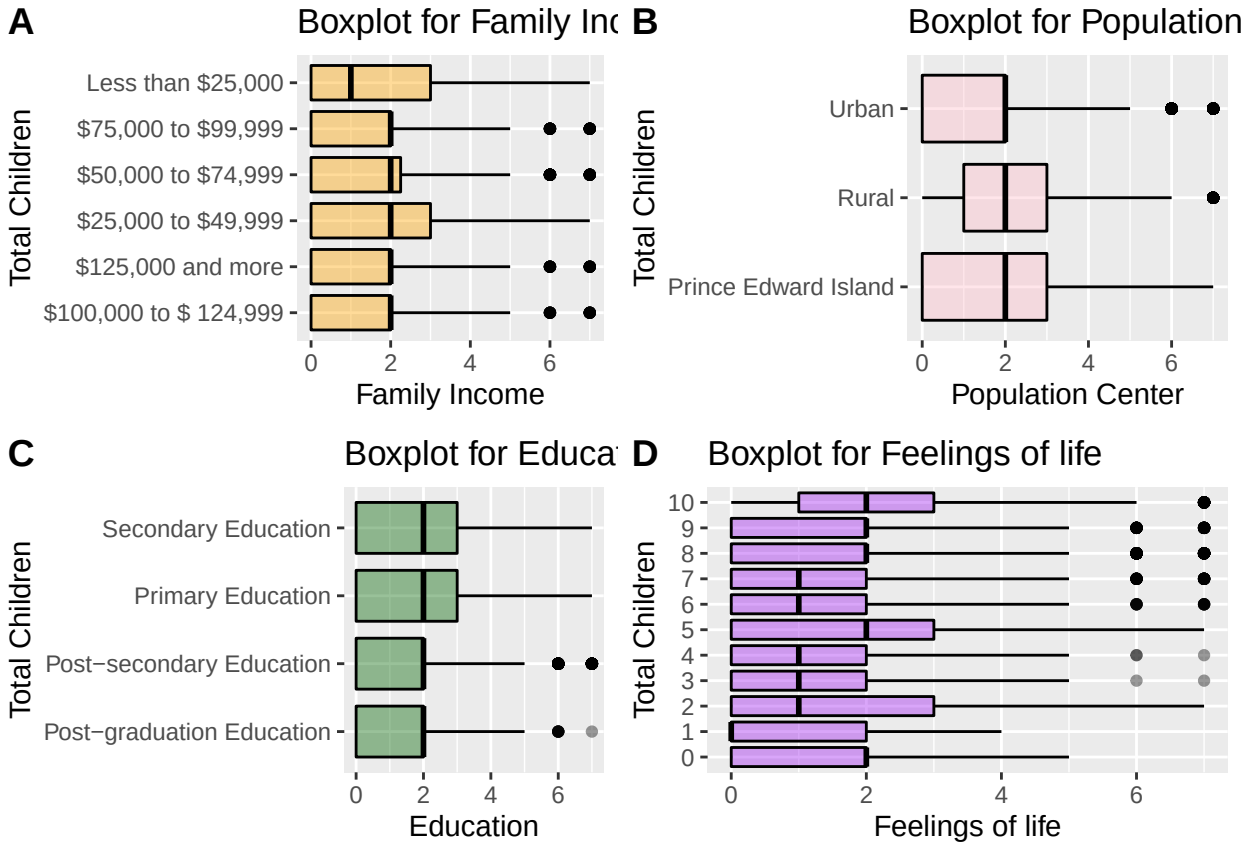


Referring to the previous analysis, the number of total children in the family is classified into two groups using the guideline of two children. According to the barplot (Figure 1), three separate barplots briefly provide the percentages that different variable classes account for in two children groups, respectively.

Firstly, most respondents have a high education level for the family with no more than two children. The bar plots show that post-secondary education and postgraduate education accounts for about 2/3 in the first group with fewer children. However, for families with more than 2 children, people with higher education only account for around 50%, less than 2/3. Therefore, most respondents with a high academic level may not choose to have more than two children. Secondly, there is no graphically significant difference among the family income in the second barplot. Each family income type seems to account for a similar percentage in respondents' with no more than 2 children. Individuals in the income class that is higher than 12500 dollars have the highest proportion. Also, within each income class, most respondents decide to have fewer children. Thirdly, there is a dramatic difference between the population center's proportion within the two groups. Urban population who have no more than 2 children is approximately three times as the remaining urban population, who have more than two children. Unlike the urban population, people from rural areas with less are two times as those with more children.

5.4 Box plots

- Figure 2:



According to boxplots (Figure 2), they demonstrate 4 categorical variables: family income, population center, education level and feeling of life.

For the family income, almost all income class types have a median of raising 2 children except for families with an income lower than 25000. Outliers exist in 4 ranges of family income, which are income from 75,000 to 99,999, from 50,000 to 74,999, above 125,000 and from 100,000 to 124,999. For the population center, all three areas face a median of having 2 children. For the urban population, the whisker box ranges from 0 to 2. However, the whisker plot for the rural population ranges from 1 to 3. This represents that the first quartile and third quartile of total children are higher for the rural population. For the education level, the first quartile and the sample median are the same for different education levels. However, respondents with lower education have a higher third quartile. Also, some outliers exist for samples that have post-graduation and post-secondary education. For the feelings of life, most respondents with higher feelings of life have a higher median for children raising.

6 Discussion

6.1 Data set

The data set used for analysis consists of *age*, *income_family*, *feelings_life*, *edu_factor*, *pop_center*, and *child_factor*. Human beings are not able to give birth until a certain age. Age is the most important factor when considering fertility. It becomes quite difficult to have babies after 40s. It takes money and time to raise a child from an infant to an adult. Financial condition becomes considerable when deciding to have children. Feelings about life represent one's satisfaction about his own life. "There is growing evidence that lifestyle choices account for the overall quality of health and life (QoL) reflecting many potential lifestyle risks widely associated with alterations of the reproductive function up to infertility(Lifestyle and fertility: the influence of stress and quality of life on female fertility, 2018)." Education affects people from every aspect. People's views about the ideal family are consistently improved. The size of the population center determines what people could do and their exposure to the new technology and theories.

We retrieved and cleaned the data set from the 2017 General Social Survey (GSS) on the Family via the CHASS data center. Variables *edu_factor* and *child_factor* are constructed by transforming education and total_children from the raw data set. Variables *age*, *income_family*, *feelings_life*, *pop_center* are kept. The resultant variables are finalized after removing the NAs. However, bias might occur when there is inaccuracy in the survey system, mistakes in data- recording process and respondents' miscomprehension towards the questionnaires. Unfortunately we are not able to adjust the systematic errors.

Data collected from surveys are subjected to sampling error, though it is unavoidable. A complete census can lead to different estimates, however, a census requires lots of time and money. Non-sampling errors can occur at almost every phase of the survey. Some variables that contain too many NAs cannot be contributed to the analysis and may cause garbage in garbage out. Another sampling approach that is less laborious compared to the stratified sampling approach is the simple random sampling but is constrained to less subjective matters.

6.2 Data Result Explanation

- Bar plots

Based on the analysis of bar plots (Figure 1), people who have higher education tend to have lower children. There is no essential conclusion drawn from the second barplot associated with family income types. Also, the size of the population center influences people's decision of baby birth. People who are living in a large urban center are more likely to raise fewer children.

- Box plots

According to the boxplots (Figure 2), people from lower family incomes tend to have fewer babies than high-income groups. Highly educated people from a bigger population center are more likely to have fewer children. Additionally, people with more happiness may have a higher fertility rate.

6.3 Model Result Explanation

- Income of a Family

Overall, the result shows that income has a positive effect on the logit odds ratio. People with higher incomes tend to have a higher logit odds ratio on average. One possible explanation would be the rich have less financial stress when considering having babies. However, the reality shows the contrary, that richer countries have lower fertility rates than poorer ones, and high-income families have fewer kids than low-income ones.

- Level of Education

Level of education has a negative effect on the logit odds ratio as people who receive a higher level of education have a lower logit odds ratio on average. More educated people have a higher opportunity cost of raising a child, as they have the option to spend the time of raising a child doing more meaningful things. Also, they might have higher expectations of their children and thus it takes them more time to educate a child than uneducated people. More educated people want their children to live a high-quality life, whereas less

educated people only want their children to survive. More educated people know more about childcare, so they do not have to worry about infant mortality. Consequently, more education causes a lower fertility rate.

- Size of Population Center

The size of the population center also has a negative effect on the logit odds ratio. People who live in larger urban population centers have a lower logit odds ratio on average than people who live in smaller urban population centers such as Prince Edward Island. Due to the uneven population size, usually, it is relatively more competitive in urban areas and thus there is a relatively better development prospect. It is highly possible that even though couples located in regions with a larger population size get married, they would postpone their plans of having a child for a better career or education development. Moreover, by taking the higher expenses in urban population centers into consideration, it is relatively more costly to raise a child and hence the lower fertility rate.

- Feelings about life

Life satisfaction has a positive effect on the logit odds ratio. Generally speaking, the main focus for people who are not satisfied with their current life is to improve their life quality rather than having children. By inference, they are probably not able to provide a good living environment along with a high-quality life for their children at the moment. Whereas those who find their life satisfactory have a higher chance of having children.

- Age

Age has a positive influence on the logit odds ratio. In a more general sense, as people get older, there is a higher probability of them having more children.

6.4 Weaknesses and Caveat

- We chose the variables by ourselves, which is full of subjectivity. There could be other factors affecting the probability we tried to research and they may add precision to our model if we include them.
- From the model diagnostics, we can tell there is a small non-linearity for the variable "age". This could harm the validity of our model.

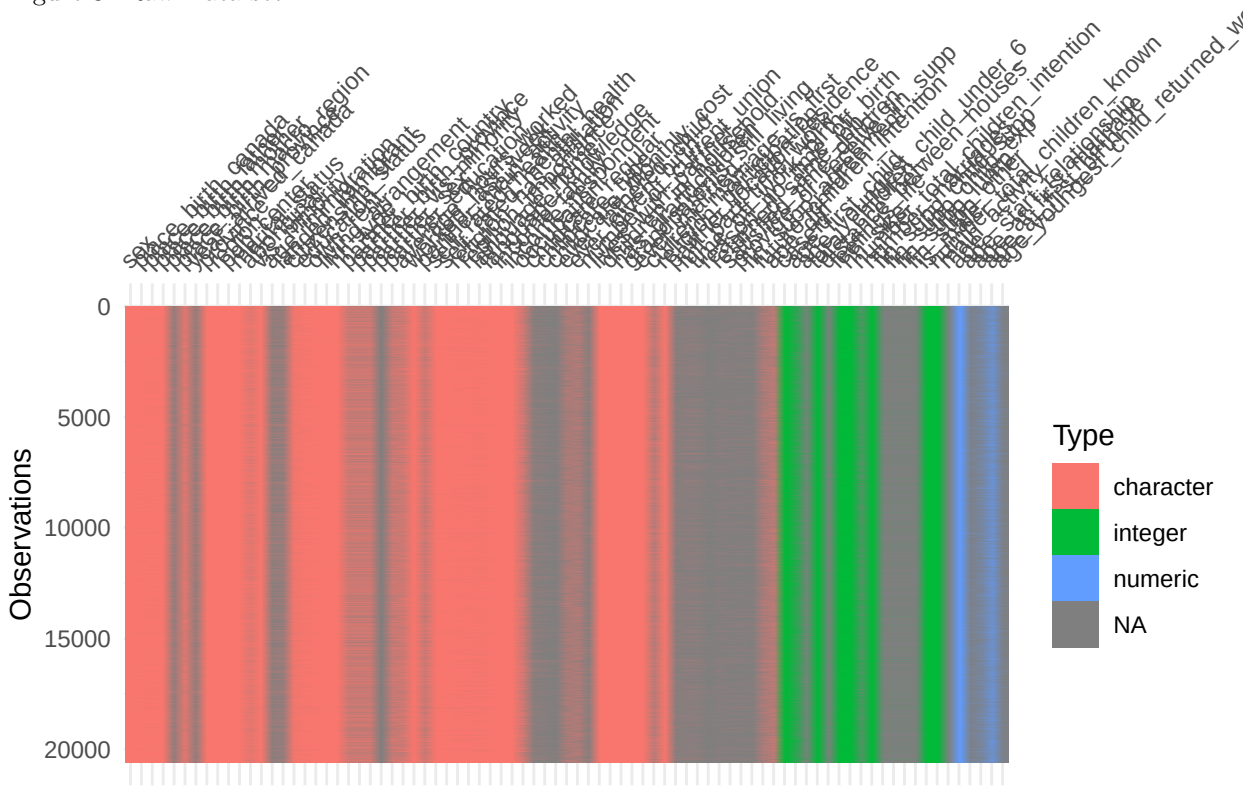
6.5 Next Steps

- To enhance the prediction accuracy of our logistic regression model, we can apply power transformation on the age variable to alleviate the non-linearity. However, this could make interpretations harder for this predictor.
- Since our primary purpose of the model is to make predictions, it is acceptable for us to include more predictors. Also, we need to avoid overfitting where this can be tested through cross-validation.
- We could apply more model/variable selection through trial and runs. Measures like adjusted R^2, AIC (Akaike's information criterion), and BIC (Bayesian Information Criterion) can help us select the best model. To compare models with the same number of predictors, we could calculate the residual sum of squares (RSS) and select the model with the smallest value.

7 Appendix

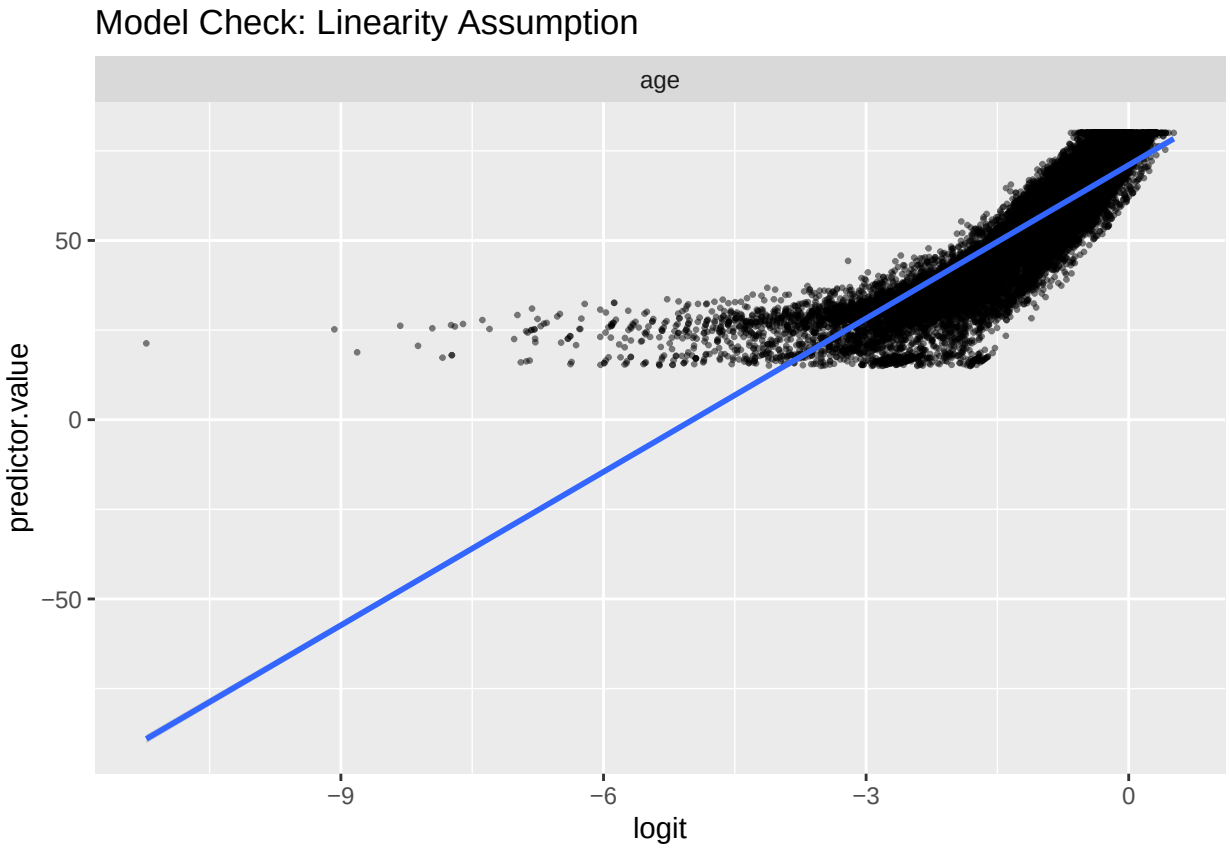
7.1 Data - Cleaned Raw Dataset

Figure 3: Raw Data set



7.2 Model - Linearity Assumption

Figure 4: Linearity Assumption Model Check



7.3 Model - Influential Observations

```
## [1] 831
```

```
## named integer(0)
```

```
## NULL
```

7.4 Model - Multicollinearity

Table 3: Multicollinearity

Table 2: VIFs for 5 Predictors

	GVIF	Df	GVIF^(1/(2*Df))
income_family	1.187	5	1.017
edu_factor	1.132	3	1.021
pop_center	1.032	2	1.008
feelings_life	1.046	1	1.023
age	1.066	1	1.033

7.5 Alternative model

Figure 5: Residual vs Fitted Plot

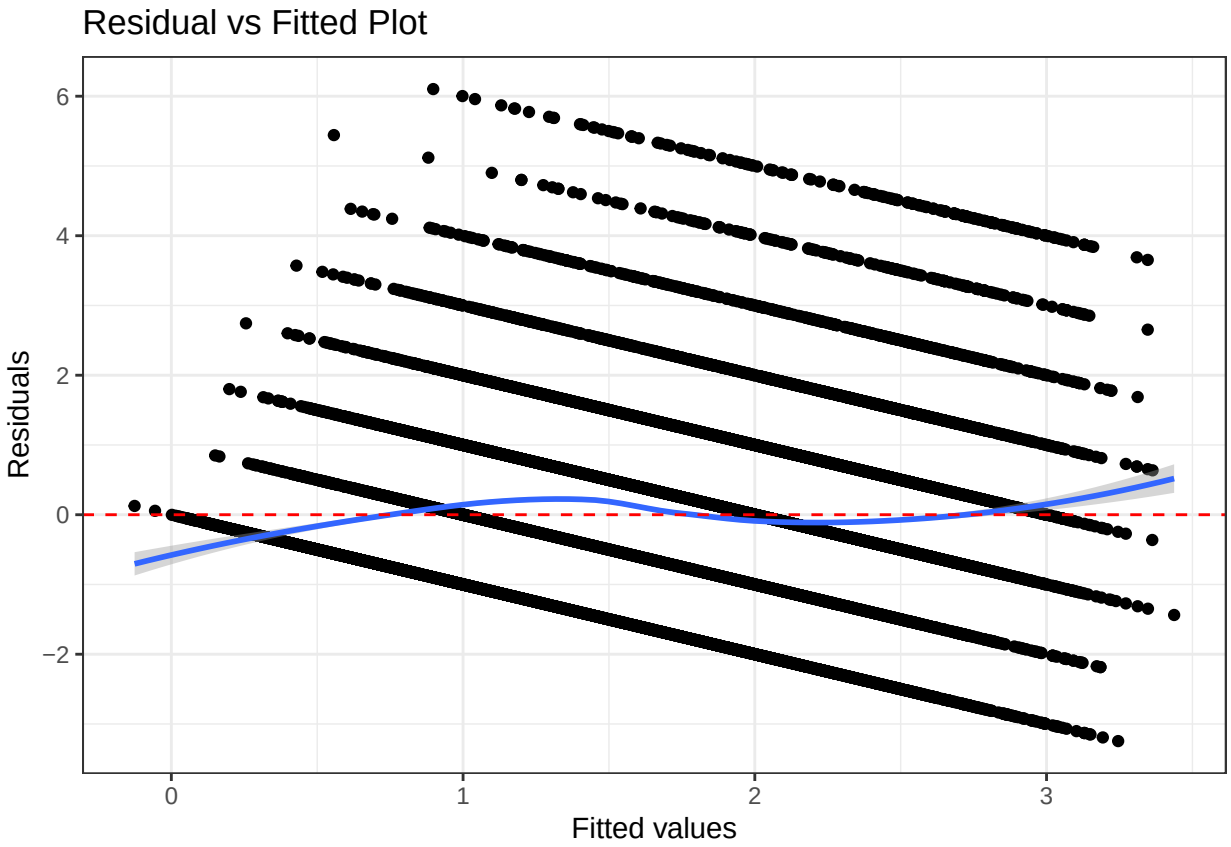
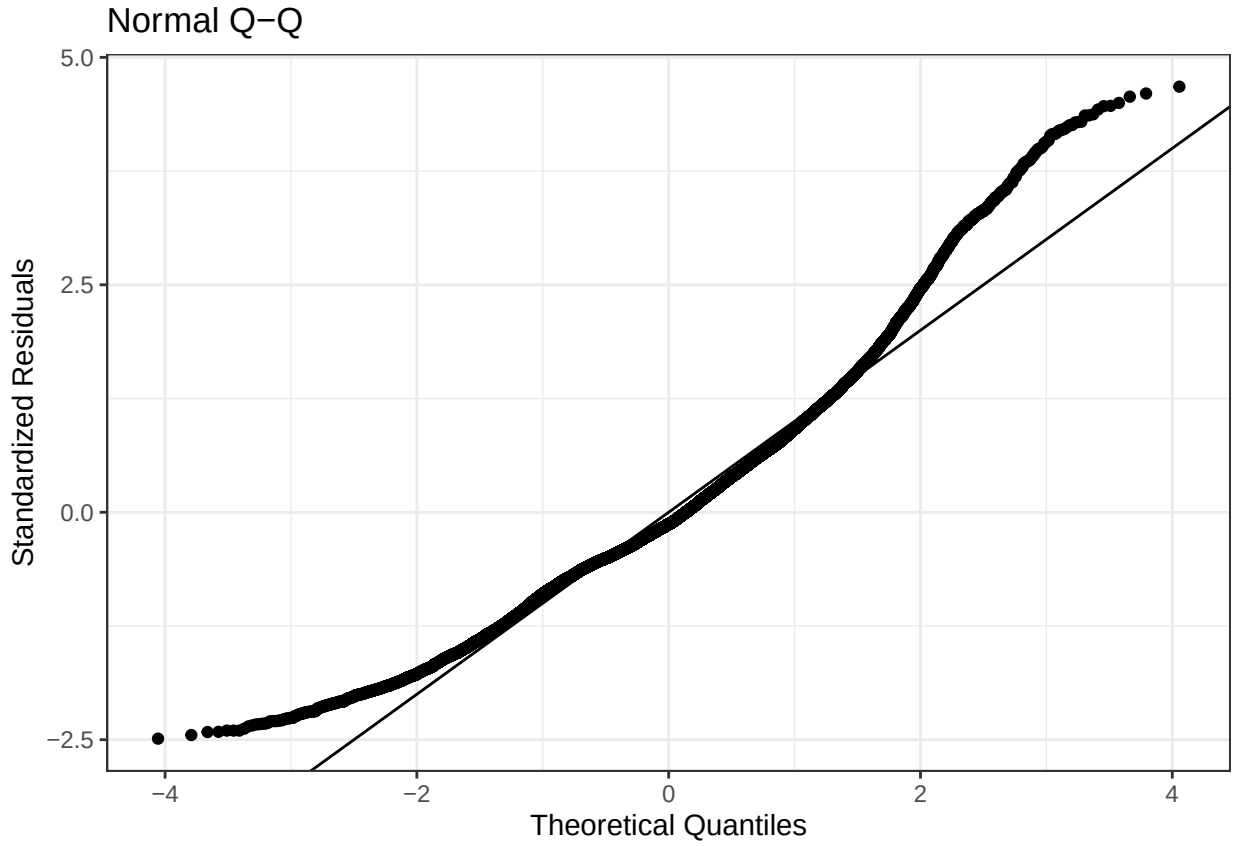


Figure 6: Normal Q-Q


Normal Q−Q
Standardized Residuals
5.0
2.5
0.0
−2.5
−4
−2
0
2
4
Theoretical Quantiles

8 References

1. Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr
2. Government of Canada, Statistics Canada. "Fertility: Fewer Children, Older Moms." Government of Canada, Statistics Canada, 17 May 2018, www150.statcan.gc.ca/n1/pub/11-630-x/11-630-x2014002-eng.htm.
3. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
4. Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
5. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
6. John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
7. Kassambara, and Michael U. "Logistic Regression Assumptions and Diagnostics in R." STHDA, 11 Mar. 2018, www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/.
8. Kim, Jungho. Female Education and Its Impact on Fertility. Feb. 2016, wol.iza.org.
9. Macrotrends. "Canada Fertility Rate 1950-2020." MacroTrends, www.macrotrends.net/countries/CAN/canada/fertility-rate.
10. Nicholas Tierney, Di Cook, Miles McBain and Colin Fay (2020). naniar: Data Structures, Summaries, and Visualisations for Missing Data. R package version 0.6.0. https://CRAN.R-project.org/package=naniar
11. Pradhan, Elina. "Female Education and Childbearing: A Closer Look at the Data." World Bank Blogs, 24 Nov. 2015, blogs.worldbank.org/health/female-education-and-childbearing-closer-look-data.
12. Price, Joseph. "How Income Affects Fertility." Institute for Family Studies, 8 Oct. 2013, ifstudies.org/blog/how-income-affects-fertility.
13. Rimal, Raju. "Playing with ggplot2." RPubs, 16 Nov. 2014, rpubs.com/therimalaya/43190.
14. Technology, Advancing Knowledge through. Computing in the Humanities and Social Sciences. www.chass.utoronto.ca/.
15. Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.
16. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686,https://doi.org/10.21105/joss.01686
17. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.
18. Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963
19. 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