The importance of Age, Family Income, Education Level and Population Centre On When a Woman Decides to Have a Child Is Multifaceted

Arjun Dhatt, Benjamin Draskovic, Yiqu Ding, Gantavya Gupta

10/19/2020

Abstract

In this report, we will be looking at the different factors that affect a woman's choice of having children. We filtered the 2017 GSS (General Social Survey) dataset to include variables related to our studies, such as age, family income, education level, and population center. Then we ran a logistic regression-based model to identify the factors that affect a woman's choice to have children. After completing our analysis, we concluded that age, income, education, and region do affect a woman's choice to have children. Our analysis results are essential because they help find factors that influence birth rates, which can aid in determining population demographics.

Introduction

Many factors can influence a woman's choice of whether or not to have children and the information can be used in many ways. This poses an interesting question — what factors influence a woman's choice to have children?

In this report, we decided to further investigate this question by deeply analyzing the different factors that influence a woman's choice to have children. Using the 2017 GSS (General Social Survey) data set, we began by identifying various variables that we thought could influence our analysis. We decided to inspect the following variables:

- Age
- Family Income
- Education Level
- Population Center

Because we are looking at factors that influence a woman's choice to have children, we decided to look at only a subset of the data — only female respondents.

To conduct the analysis, we decided to use logistic regression. In our study, variables such as family income, education level, and population center are categorical variables; to accommodate these variables, we used dummy variables to incorporate them in the logistic regression.

After looking at the results from our logistic regression and using a significance level of 0.05, we inspected the p-values of different variables to determine if the variables significantly influence the probability a woman has children. Age was determined to be a significant factor; income for those making an income less than 100,000 was determined to be influential; Population centered at PEI, rural and small population centers were also determined to be substantial.

The results from our analysis could be useful in many different ways. Information about factors that influence a woman's choice to have children can be used for information related to population demographics. Countries with low birth rates can also find the results of our analysis useful as it can help them determine factors that can raise the birth rates.

In this report, we will begin by discussing the data set and then model our results.

Data

This analysis uses the data collected by the 2017 General Social Survey (GSS). A survey done by the Social and Aboriginal Statistics Division of Statistics Canada. Their focus for that year of data collection was Family and the different forms family takes across Canada. This involved collecting data on current living arrangements, child care, household chores splits and many more. However, for the purpose of this analysis the key pieces of collected data are whether or not a female respondent has had children, what her reported education level is, how religious they are, how old they are, what their current family income is, and whether they live in a rural or urban community. Table 1 displays the first couple responses along the variables of interest.

Table 1

Case ID	Sex	Have Children	Age	Family Income	Education Level	Urban / Rural
1	Female	1	52.7	25k to 50k	HS or less	Urban
3	Female	1	63.6	75k to 100k	University Undergraduate	Rural
4	Female	1	80.0	100k to 125k	HS or less	Urban
6	Female	1	63.0	50k to 75k	HS or less	Urban
7	Female	1	58.8	25k or less	HS or less	Urban
8	Female	1	80.0	25k or less	HS or less	Urban

The data for the GSS was collected from Feb. 2nd, 2017 to Nov. 30th, 2017. The target population of this survey was all non-institutionalized people living in Canada's 10 provinces and who are 15 years or older (GSS 2017). This notably immediately excludes those who have been imprisoned or are in hospitals or nursing homes. It also excludes those who live in the Canadian territories. Both are notable exclusions as they may have very different experiences as compared to the average respondent. Specifically many of Canada's First Nation and Inuit reserves are in the territories.

The Survey is further reduced by it's chosen frame. It uses a combination of telephone numbers (landline and cellular) with Stats Canada's Address Register (GSS 2017). Which removes the possibility of those who do not own phones, or do not have a stable address from those that could be sampled. This is also important as it removes individuals who may be experiencing acute poverty as well as those who have recently moved. It is notable even those with stable housing may not be included if they do not have a phone number.

This frame is then sampled via cross sectional design, where geographical regions were split into strata. Each strata was then randomly sampled without replacement. Households were then called when they then had to confirm if an individual above the age of 15 lived there. Once that was confirmed a random member of the household over 15 was selected and interviewed over the phone (GSS 2017). This requirement for voluntary agreement to partake in a long form telephone interview does provide some sampling bias.

This will affect all of our data in the same manner. While we will be working with a subset of the data, only looking at female respondents, these women living in Canada will have had to have an address, phone number, live in the provinces, and be willing to voluntarily take a phone survey.

Something of note in the data collection is that income was collected by connecting households to their household tax returns. This means that much of the ambiguity often present in self reported income survey responses was removed from the data. This is hugely useful for better studying low and high income individuals as they have commonly been known to over or under report their earnings.

For further information the User Guide and Raw Codebook of the GSS has been included in the "Data Information" file in this projects related Github.

Based on the number of respondents able to be used in this analysis, 10,916, and the overall female population of Canada 17.8 million, the analysis should have a general confidence level of 95% (Survey Systems 2020). The average likelihood or probability that a woman would have children in our sample was 0.73. Next we will begin to model this probability to see how it is affected by our independent variables.

Model

The model that we are using is logistic regression. We are interested in how variables such as age, family income, education level, and the type of population center(of the area that she lives in) affect a woman's decision to have children. The response variable that we are interested in is whether a woman has children, which is a categorical variable with two levels: 0 represents not having children, and 1 means having children. The first explanatory variable is age. This is the variable that we want to be specific because the specific age instead of age group provides a lot more information than age groups. For example, it is very unlikely that a woman has children at age 22 but not very unlikely for 23 or 24. Which makes sense because at 22, most people are still studying, and it is possible to get married after graduation and to have children after then. The ladder three are categorical variables because we are interested in their effects on the childbirth decision as a group. We want it to be easy to get the rough probability that a woman has children, given some of the most basic information.

Logistic regression allows us the simplicity of analyzing the categorical response variable while incorporating both numerical and categorical explanatory variables. Due to the large sample size and the nature of GSS, the sample represents the population well for a rough estimate. If a more precise estimate is necessary, we can simply post-stratify on top of the logistic regression model results.

Logistic regression estimates $\beta_0 ... \beta_k$ in the following equation:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

where p is the probability of event A that we are interested in, $\beta_0$is the intercept, $x_1 ... x_K$ are our variables of interest and $\beta_1 ... \beta_k$ are parameters for each of these variables. Based on the result, we are able to estimate p for a particular case given all the variables.

In our case, we want to estimate $\beta_{age}, \beta_{inc}, \beta_{edu}, \beta_{pop}$ in:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_{age} x_{age} + \beta_{inc} x_{inc} + \beta_{edu} x_{edu} + \beta_{pop} x_{pop}$$

where p is the probability that a woman has children, β_0 is the intercept, β_{age} is the parameter for age, β_{inc} is the parameter for family income, β_{edu} is the parameter for education level and β_{pop} is the parameter for the type of population center.

We use `glm()` from `stats` package in `R` to fir the model to our data. We use `as.factor()` to incorporate dummy variables for all the categorical variables: family income, education level and the type of population center. For each categorical variable with n levels, we need n-1 dummy variables to fully study its influence on our response variable.

The dummy variables setting ups are stated in table 3, 4 and 5 in the appendix.

Results

Table 2 summaries our model results:

Table 2: Summary of Losgistic Estimates

variable	estimate	pvalue
intercept	-1.513991	< 2e-16
age	0.060790	< 2e-16
income greater than $125,000	0.161424	0.0810
income between $25,000 to $49,999	-0.680099	8.24e-13
income between $50,000 to $74,999	-0.449230	3.49e-06
income between $75,000 to $99,999	-0.197017	0.0491
income less than $25,000	-0.872898	< 2e-16
high school or less education	-0.102521	0.0982
University Graduate	-0.779989	< 2e-16
University Undergraduate	-0.468140	9.25e-13
Poplation centered at PEI	0.276375	0.0409
Rural areas and small population centres(non CMA/CA)	0.526007	2.86e-14

In an equation, this means

$$log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0 + \hat{\beta}_{age}x_{age} + \hat{\beta}_{inc_1}x_{inc_1} + \hat{\beta}_{inc_2}x_{inc_2} + \hat{\beta}_{inc_3}x_{inc_3}$$

$$+\hat{\beta}_{inc_4}x_{inc_4} + \hat{\beta}_{inc_5}x_{inc_5} + \hat{\beta}_{HS}x_{HS} + \hat{\beta}_{graduate}x_{graduate} + \hat{\beta}_{under}x_{under}$$

$$+\hat{\beta}_{PEI}x_{PEI} + \hat{\beta}_{rural}x_{rural}$$

Notice that we have incorporated our categorical variables using dummy variables. "inc1", "inc2", "inc3", "inc4", "inc5" represent the six income categories; "HS", "graduate" and "under" describe the four education levels; "PEI" and "rural" represent the three population center.

The model predicts the following result for the probability a woman has children p, rounded to three decimal places. Since log is an one to one function with p, we say the change on $log(\frac{\hat{p}}{1-\hat{p}})$ is isomorphic to any change on p, the probalbility that a woman in Canada has children.

$$log(\frac{\hat{p}}{1-\hat{p}}) = -1.51 + 0.061x_{age} + 0.161x_{inc_1} - 0.68x_{inc_2} - 0.449x_{inc_3}$$

$$-0.197x_{inc_4} - 0.873x_{inc_5} - 0.103x_{HS} - 0.78x_{graduate} - 0.468x_{under}$$

$$+0.276x_{PEI} + 0.526x_{rural}$$

The interpretation of the dummy variables' prediction results is comparing it to a certain level that is not in the equation above. A $\hat{\beta}_{HS} = -0.103$ does not mean the coefficient for a high school level education is -0.103. It represents the difference of influence (on childbirth decision) between a woman who has a college level of education and a high school level of education. -0.103 indicates that a woman who has a high school degree or below is less likely to have children than a woman who has a college degree if it is influential at all depending on the p-value.

Using $\alpha = 0.05, H_0 : \hat{\beta} = 0, \hat{\beta} \neq 0$, the p values indicate weak evidence that having a family income higher than 125,000 dollars/year affects a woman's probability of having children. At the same time, it is evident that other income levels do influence the probability. A p-value of $0.0982 > 0.05$ indicates weak evidence to reject H_0; therefore, we cannot say having high school or fewer education influences a woman's probability of having children.

There is evidence that both having a family income between 75,000 to 99,999 dollars per year and living in an area with a population center at PEI have influences on p because their p values are smaller but close to 0.05. There is strong evidence that the rest of the variables influences p.

All else the same, an older woman is more likely to have children than a young woman. A woman's family income affects her decision to have children when the income is below 125,000 dollars per year. Having a university graduate or undergraduate degree reduces a woman's probability of having children comparing to having a college degree while having high school or less education does not affect that decision. A woman living in PEI or rural areas and smaller population centers are more likely to have children than a woman living in urban areas and larger population centers.

Discussion

In conclusion modeling the GSS data shows that there are significant effects on a woman's choice to have children by age, income, education, and the type of population center they live in. This data is important as it both helps understand the decision making process percpective mothers must face and it can help identify possible barriers to having children.

Firstly, older women are more likely to have made the decision to have children. This could both be from the fact that older women have had more opportunities to have children and/or be based on the views older generations have on child bearing. This data shows evidence of the common tendency found in the Canadian Demographic make up known as the "Reverse Pyramid." That there are continually fewer young children being born then their were previously.

Secondly, unlike the common myth that the poor are the ones having large numbers of children the data shows that those making under $100,000 a year as a household were far less likely to have children then those who did make $100,000 annually. This does make some intuitive sense as havign a child is a large commitment in resources the family must provide. This also could be a barrier to childbirth for women who are interested in having children.

This is further supported by the finding that when a woman's household begin making $100,000 or more the effect income had on her decision to have children essentially dissappeared. This suggests that once the prospect of having a child is no longer going to strain a families income it becomes a matter of choice based upon a woman's own preferences rather than external forces.

Thirdly, when looking at education level we see that the more highly educated a woman becomes the less likely they are to have children. This is an oft mentionned phenomina that could be linked to the amount of time and resources it takes to pursue higher education. Or it could be linked to other hypothesis on how damaging takign a break in your career to have children can be to loing term career success. This does work off the assumption that those seeking higher education are highly career focused however.

Fourthly, it is interesting to note that women living in PEI or in Rural areas in other provinces were significantly more likely to have children then those living in dense urban centers. This could represent a difference in cultural norms and gender roles in rural communities. Or it could represent the fact that rural communities are often safer places to raise children, due to their lower crime rates and increased space. It is difficult to parce out the reason for this effect without further study.

It is however important to re-examine possible sampling errors that exist with in the data set. In order to do this let's review the ages of respondents, their incomes, and their education levels, seeing if these distributions match teh expected population distributions.

Figure 1 displays the general breakdown of respondents ages. This is followed by a table taken from Anne Milan's report on the age of the Female population in Canada (Milan 2016). As can be seen the sampling method caused an over representation of older individuals and an under representation of the young. This could have to do with the requirement of a stable address, or because of the voluntary nature of the survey.

Figure 1

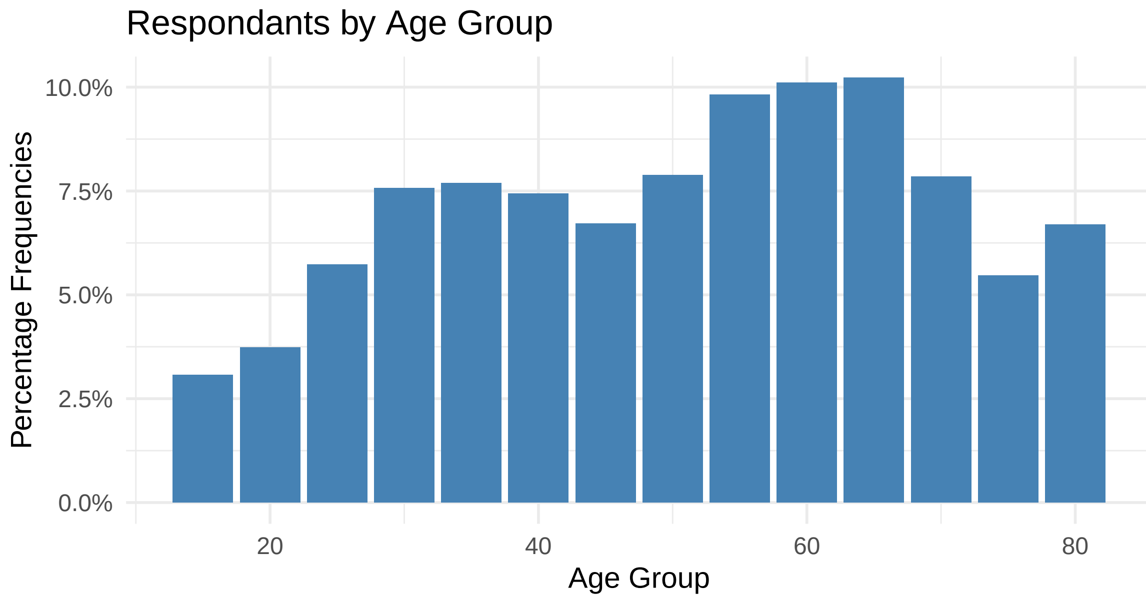

Respondants by Age Group
Percentage Frequencies
10.0%
7.5%
5.0%
2.5%
0.0%
20
40
60
80
Age Group

Figure 1.B Canadian Female Population by Age 2014

Population by age group and sex, Canada, 2014

Age group	Females			Males	
	thousands	percentage	as a percentage of the age group	thousands	percentage
4 and under	938.6	5.2	48.8	986.3	5.6
5 to 9	935.3	5.2	48.8	982.9	5.6
10 to 14	906.7	5.1	48.6	958.8	5.4
15 to 19	1,039.1	5.8	48.6	1,099.6	6.2
20 to 24	1,209.0	6.7	48.9	1,263.1	7.2
25 to 29	1,212.3	6.8	49.7	1,225.4	7.0
30 to 34	1,242.6	6.9	50.1	1,237.3	7.0
35 to 39	1,187.4	6.6	50.2	1,179.8	6.7
40 to 44	1,179.1	6.6	50.0	1,179.5	6.7
45 to 49	1,241.8	6.9	49.8	1,250.3	7.1
50 to 54	1,381.4	7.7	49.8	1,393.2	7.9
55 to 59	1,281.1	7.2	50.1	1,276.2	7.2
60 to 64	1,096.9	6.1	50.6	1,071.3	6.1
65 to 69	937.8	5.2	51.2	893.1	5.1
70 to 74	691.4	3.9	52.6	623.2	3.5
75 to 79	528.6	3.0	54.4	443.6	2.5
80 to 84	422.3	2.4	57.3	314.3	1.8
85 to 89	292.5	1.6	62.9	172.8	1.0
90 and over	191.2	1.1	72.0	74.4	0.4
Total population	**17,915.4**	**100.0**	**50.4**	**17,625.0**	**100.0**

Note: Adjusted for net census undercoverage. Preliminary estimates.

Source: Statistics Canada, Cansim table 051-0001, 2014.

(Milan 2016)

Figure 2 displays the general breakdown of respondents by Family/Household income. When comparing this to a table provided by the 2010 National Household Survey, despite the difference in bin definition, we see that there are highly skewed results towards those with higher incomes (Stats Canada 2018). Both data sets used non-institutionalized Canadians aged 15 and older, who lived together in a household. Again this could be both from the fact that addresses were required meaning those living more transiently were not tracked properly, and individuals had to be willing to provide a large amount of their time voluntarily. Something that could be more difficult for someone making a lower income.

Figure 2

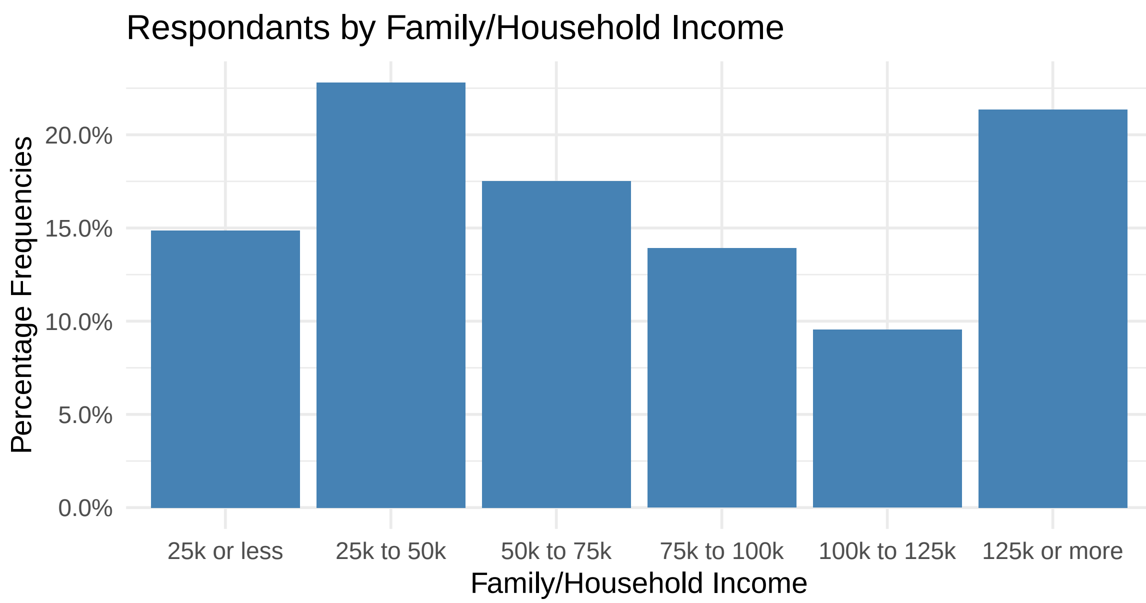
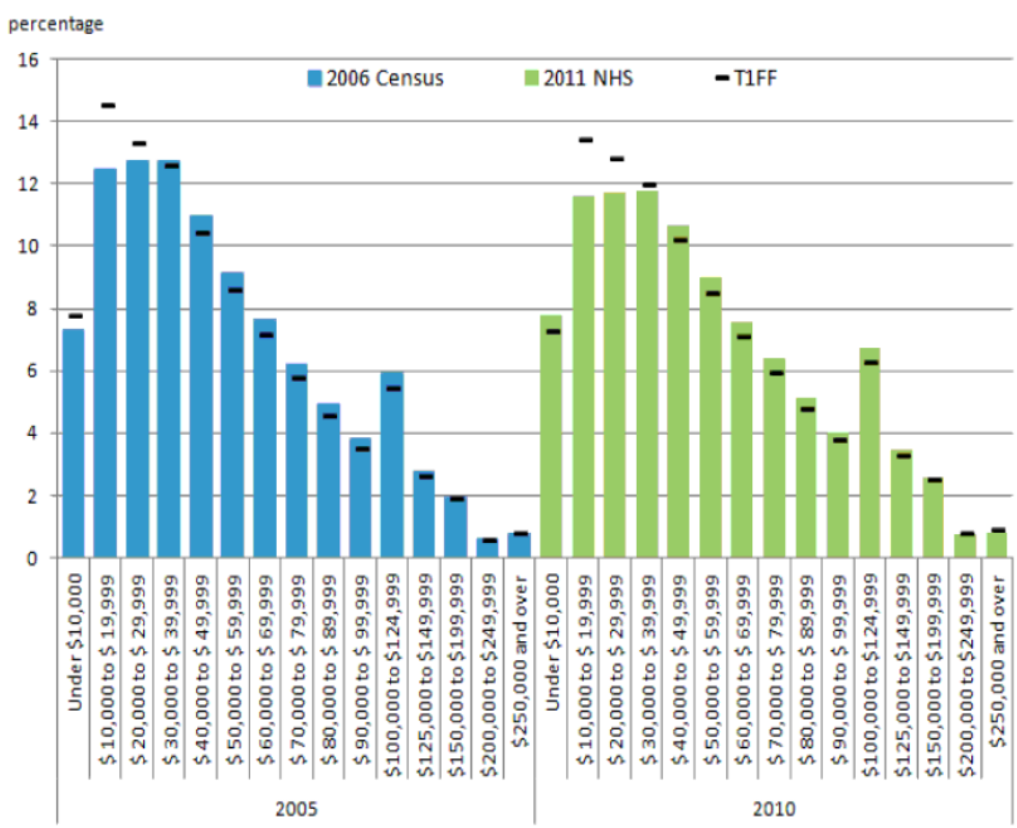


Figure 2.B Canadian Family Income Levels 2010



(Stats Canada 2018)

What these two modes of analysis show is that the frame and methodology of the data collection had a significant effect on those that were surveyed.

However, not all demographic variables were skewed like this. Looking at Figure 3 one can see the breakdown by Education Level. When comparing this table to a distribution from the 2016 Census we see that the sample population is similarly distributed to the Canadian population (Stats Canada 2019). While the bins are defined differently, combining the census bins together as well as combining our split of Undergraduate versus Graduate University degrees all percentages are near population levels. There only being a small 10% difference from one to the other.

Figure 3

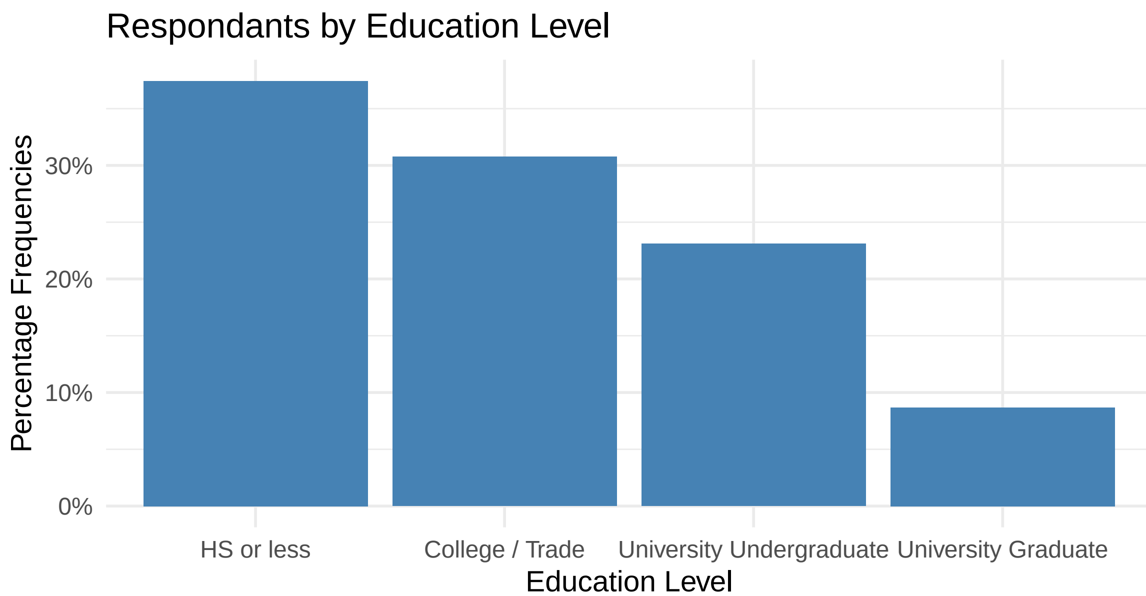


Figure 3.B Canadian Education Levels by Percentage 2016

Geographic name	Total - Highest certificate, diploma or degree [1]	No certificate, diploma or degree	Secondary (high) school diploma or equivalency certificate [2]	Apprenticeship or trades certificate or diploma [3] [4]	College, CEGEP or other non-university certificate or diploma	University certificate or diploma below bachelor level	University certificate, diploma or degree at bachelor level or above [5]
Canada	100.0	11.5	23.7	10.8	22.4	3.1	28.5

(Stats Canada 2019)

Weaknesses and Next Steps

Using the GSS dataset from 2017 may lead to weaknesses in our dataset. Because we use an exclusive dataset to the year 2017, our data is limited and may be exposed to biases from that year. This may skew our results in a direction that might not be indicative of the real results. A solution to this problem is to survey participants over multiple years; this allows our data to contain more information and be less biased.

Another potential weakness of the dataset is that it only contains information on Canadian participants. This can be viewed as a weakness because it prevents our results from generalizable to a different country. If we examined another country that is not in North America, our results might differ drastically. A potential solution to this weakness is to gather information in different countries around the world. That way, we have information on other countries, and our results can be more generalizable.

Since we are modeling what factors influence a woman's choice to have children, it would be ideal for collecting longitudinal data. Because we are only collecting information at a certain point in time, our analysis may be missing out on certain influential factors. Collecting data from someone who is 80 years old will be drastically different from collecting data from someone in their 20s. The older our respondents are, the more likely they are to have already had a child at some point in their life. A solution to this problem could be to collect longitudinal data to compare all respondents from the same perspective. Another solution, in this case, could be to conduct an in-depth interview rather than a survey. An in-depth interview ensures that we receive more detailed responses. Asking questions about when they decided to have children and what their life situation was like can provide much more meaningful information and more meaningful results.

The dataset we used in the analysis has an average that is much older than the average Canadian population and an average income relatively high. To address this issue, the next step in this analysis can be to use post-stratification. This allows our sample to reflect the population, making our results more meaningful and more accurate.

With these next steps this data can help to identify patterns in women's decisions to give birth and possible barriers women making that decision may have to face.

Appendix

Dummy variable setups for income, education, and population center

Below are the dummy variable setups for income, education, and population center

Table 3

Table 3: Dummy Variable Coding Set Up for Income Levels

income.level	inc1	inc2	inc3	inc4	inc5	inc6
greater than $125,000	1	0	0	0	0	0
between $25,000 to $49,999	0	1	0	0	0	0
between $50,000 to $74,999	0	0	1	0	0	0
between $75,000 to $99,999	0	0	0	1	0	0
less than $25,000	0	0	0	0	1	0
between $100,000 tO $124,999	0	0	0	0	0	0

Table 4

Table 4: Dummy Variable Coding Set Up for Education Levels

education.level	HS	graduate	under	college
high school or less education	1	0	0	0
University Graduate	0	1	0	0
University Undergraduate	0	0	1	0
college/trade	0	0	0	0

Table 5

Table 5: Dummy Variable Coding Set Up for Population Center

population.center	PEI	rural	urban
Poplation centered at PEI	1	0	0
Rural areas and small population centres(non CMA/CA)	0	1	0
Larger urban population centres (CMA/CA)	0	0	0

Git Repository

The link to the Git Repository is: https://github.com/STA304-PS3/PS3

Code

Below you will also find the R code used in this analysis:

Preamble — setting up libraries and setting up the logistic regression

```
knitr::opts_chunk$set(echo = FALSE)
library(dplyr)
library(knitr)
library(ggplot2)
library(tidyverse)
library(scales)
library(kableExtra)
```

```
#logistic regression
my_logit_dd <- glm(
  child ~ age + as.factor(income_family) + as.factor(education_level)
  + as.factor(pop_center),
  data = my_data, family = "binomial"
)
```

Code for tables and figures used in analysis

Table 1

```
my_data <- read.csv("gss-prepared-for-analysis.csv") #read in the data
display_data <- my_data %>% #Select variables of interest
  select(
    caseid,
    sex,
    child,
    age,
    income_family,
    education_level,
    pop_center
  )
```

```
#Shorten the response descriptions
display_data[display_data=="Larger urban population centres (CMA/CA)"] <- "Urban"
display_data[display_data==
             "Rural areas and small population centres (non CMA/CA)"] <- "Rural"
display_data[display_data=="$25,000 to $49,999"] <- "25k to 50k"
display_data[display_data=="Less than $25,000"] <- "25k or less"
display_data[display_data=="$50,000 to $74,999"] <- "50k to 75k"
display_data[display_data=="$75,000 to $99,999"] <- "75k to 100k"
display_data[display_data=="$100,000 to $ 124,999"] <- "100k to 125k"
display_data[display_data=="$125,000 and more"] <- "125k or more"
# Table 1 of the first couple responses
Table1 <- head(display_data) %>%
  kable(align = "c",
        format = "simple",
        col.names = c("Case ID", "Sex", "Have Children", "Age",
                      "Family Income", "Education Level", "Urban / Rural"),
        ) %>%
  kable_styling(font_size = 5)

Table1
```

Table 2

```
#creating the summary table
my_log_res <- data.frame(
  variable = c("intercept","age", "income greater than $125,000", "income
             between $25,000 to $49,999", "income between $50,000 to $74,999",
             "income between $75,000 to $99,999", "income less than $25,000",
             "high school or less education", "University Graduate",
             "University Undergraduate","Poplation centered at PEI", "Rural
             areas and small population centres(non CMA/CA)"),#list of all
  #variable names including levels
  estimate = c(-1.513991, 0.060790, 0.161424, -0.680099, -0.449230, -0.197017,
             -0.872898, -0.102521, -0.779989, -0.468140, 0.276375, 0.526007),
  # list of estimate results
  pvalue = c("< 2e-16", "< 2e-16", "0.0810", "8.24e-13", "3.49e-06", "0.0491",
           "< 2e-16", "0.0982", "< 2e-16", "9.25e-13", "0.0409", "2.86e-14" )
  # list of p values
)
kable(my_log_res, caption = "Summary of Losgistic Estimates",
      label = "Data Source: GSS2017")
```

Table 3

```
dummy_income <- data.frame(
  "income level" = c("greater than $125,000", "between $25,000 to $49,999",
                     "between $50,000 to $74,999", "between $75,000 to $99,999",
                     "less than $25,000", "between $100,000 t0 $124,999"),
  inc1 = c(1,0,0,0,0,0),
  inc2 = c(0,1,0,0,0,0),
```

```
  inc3 = c(0,0,1,0,0,0),
  inc4 = c(0,0,0,1,0,0),
  inc5 = c(0,0,0,0,1,0),
  inc6 = c(0,0,0,0,0,0)
)
kable(dummy_income, caption = "Dummy Variable Coding Set Up for Income Levels")
```

Table 4

```
dummy_edu <- data.frame(
  "education level" = c("high school or less education", "University Graduate",
                        "University Undergraduate", "college/trade"),
  HS = c(1,0,0,0),
  graduate = c(0,1,0,0),
  under = c(0,0,1,0),
  college = c(0,0,0,0)
)
kable(dummy_edu, caption = "Dummy Variable Coding Set Up for Education Levels")
```

Table 5

```
dummy_pop <- data.frame(
  "population center" = c("Poplation centered at PEI", "Rural areas and small
                          population centres(non CMA/CA)", "Larger urban
                          population centres (CMA/CA)"),
  PEI = c(1,0,0),
  rural = c(0,1,0),
  urban = c(0,0,0)
)
kable(dummy_pop, caption = "Dummy Variable Coding Set Up for Population Center")
```

Figure 1

```
Figure1 <- my_data %>% #Create an overall view of Age of respondents
  ggplot(aes(x = age_group)) +
  geom_bar(aes(y = ..prop..), stat = "count", fill = "steelblue") +
  scale_y_continuous(labels=scales::percent) +
  labs(title = "Respondants by Age Group", y = "Percentage Frequencies",
       x = "Age Group")+
  theme_minimal()  #Adjust how it is displayed
```

Figure 2

```
display_data2 <- my_data %>% #select smaller subset fo data
  select(
```

```
    age_group,
    income_family
  )
#Shorten the responses
display_data2[display_data2=="$25,000 to $49,999"] <- "25k to 50k"
display_data2[display_data2=="Less than $25,000"] <- "25k or less"
display_data2[display_data2=="$50,000 to $74,999"] <- "50k to 75k"
display_data2[display_data2=="$75,000 to $99,999"] <- "75k to 100k"
display_data2[display_data2=="$100,000 to $ 124,999"] <- "100k to 125k"
display_data2[display_data2=="$125,000 and more"] <- "125k or more"
Figure2 <- display_data2 %>%
  #Create an overall view of family income for the respondents
    mutate(income_family = fct_relevel(income_family,
            "25k or less", "25k to 50k", "50k to 75k",
            "75k to 100k", "100k to 125k", "125k or more")) %>%
  ggplot(aes(x = income_family)) +
  geom_bar(aes(y = (..count..)/sum(..count..)), stat = "count",
           fill = "steelblue") +
  scale_y_continuous(labels=scales::percent) + #Make it a percentage based scale
  labs(title = "Respondants by Family/Household Income",
       y = "Percentage Frequencies", x = "Family/Household Income")+
  theme_minimal()  #Adjust how it is displayed
```

Figure 3

```
Figure3 <- my_data %>% #Create an overall view of education level of respondents
  mutate(education_level = fct_relevel(education_level,
            "HS or less", "College / Trade", "University Undergraduate",
            "University Graduate")) %>%
  ggplot(aes(x = education_level)) +
  geom_bar(aes(y = (..count..)/sum(..count..)), stat = "count",
           fill = "steelblue") +
  scale_y_continuous(labels=scales::percent) +
  labs(title = "Respondants by Education Level", y = "Percentage Frequencies",
       x = "Education Level")+
  theme_minimal()  #Adjust how it is displayed
```

References

- Monica Alexander (2019). “Analyzing Name Changes after Marriage Using a Non-Representative Survey.” Monica Alexander, 7 Sept. 2019, www.monicaalexander.com/posts/2019-08-07-mrp/.
- Rohan Alexander and Sam Caetano (2020). “GSS_Cleaning” Retreived from https://q.utoronto.ca/courses/184062
- 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. URL https://rmarkdown.rstudio.com.
- 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
- GSS. (2017). General Social Survey – Family (GSS). Retrieved from https://www.statcan.gc.ca/eng/survey/household/4501
- Lauren Kennedy, and Jonah Gabry (2020). “MRP with Rstanarm.” Rstanarm, mc-stan.org/rstanarm/articles/mrp.html

- Anne Milan (2016). “Female Population.” Government of Canada, Statistics Canada, www150.statcan.gc.ca/n1/pub/89-503-x/2015001/article/14152-eng.htm.
- Statistics Canada (2018). “Income Reference Guide, National Household Survey, 2011.” Government of Canada, Statistics Canada, www12.statcan.gc.ca/nhs-enm/2011/ref/guides/99-014-x/99-014-x2011006-eng.cfm.
- Statistics Canada (2019). “2016 Census.” Highest Level of Educational Attainment (General) by Selected Age Groups 25 to 64, Both Sexes, % Distribution 2016, Canada, Provinces and Territories, 2016 Census – 25% Sample Data, www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/edu-sco/Table.cfm?Lang=E&T=11&Geo=00&SP=1&view=2&age=2&sex=1
- Survey Systems (2020). “Sample Size Calculator.” Sample Size Calculator - Confidence Level, Confidence Interval, Sample Size, Population Size, Relevant Population - Creative Research Systems, 2020, www.surveysystem.com/sscalc.htm.
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.
- Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963
- 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
- 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.
- 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
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- 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