Sta304 - Assignment 2

Jessica Glustien, Nick Callow, Min Zhang, Dongqi Bi

Oct 19 2020

Abstract

This paper examines different possible factors that can affect the age a woman gives birth to her first child at. Using data collected in the 2017 GSS, a multiple regression model was created to examine the relationship between women's marital status, education level, total number of children, and the age that the first child was born at. This model showed a negative relationship between total number of children and age at first birth, and a positive relationship between the level of education achieved and the age at first birth. A strong correlation was not found between marital status and age at first birth.

Introduction

In this assignemnt, our team will use data from the 2017 General Social Survey on Family (GSS) package monitored by Statistics Canada to estimate a woman's age at her first birth. The 2017 GSS is a sample survey conducted from February 2 to November 30 2018 via telephone. The target population includes all Canadians who are 15 or over living in the 10 provinces.

When our team members are doing research about factors affecting a woman's age at her first birth, we found that "women with college degrees have children an average of seven years later than those without."(Bui & Miller, 2020) This interests us a lot as in general, if a woman is pursuing a professional education like a law degree, she needs to put much more effort and time to complete it, which will delay her age at her first birth. This also potentially affects their marriage age, that is, if a woman gets married at a relatively old age, she may have her first child at a relatively old age too. However, if a woman decides to have more children, she may start to have her first child at a relatively young age.

In consequence, we decide that we will set variables Ever_married, Education and Total_children as predictor variables and Age_at_first_birth will be response variable from gss.csv dataset. Based on our assumptions, we also think that the statistical model we will create will be linear regression model.

Data

The data set selected for analysis is the 2017 General Social Survey (GSS). Data collection took place between February and November of 2017, and the goal of the survey is to learn more about Canadian family structures (Statistics Canada, 2017). In this section, we provide an analysis of the survey, including its key features, methdology, and merits.

Target Population and Sampling Frame

To learn more about Canadian family structures, the GSS focuses on a specific section of Canadians. This group, called the *target population* is "all non-institutionalized persons 15 years of age or older, living in the ten provinces of Canada" (Statistics Canada, 2017, Data Sources and Methdology Section). Of note is that

residents of the territories are excluded from the GSS. This decision by Statistics Canada is explored further in the *Weaknesses* section below.

To survey individuals in the target population, we need to know who they are and how to contact them. This list of people is called the *sampling frame*, and Statistics Canada recruits individuals from this list (according to the sampling approach detailed below). The sampling frame of the GSS is built in the following way. First, a list of landline and cellular phone numbers are obtained from the Census (Statistics Canada, 2017). These phone numbers are then grouped into households if they share the same address (Statistics Canada, 2017). Ideally, everyone in the target population is included in the sampling frame (Wu & Thompson, 2020). However, when this is the not the case, it introduces a bias in the survey known as *coverage error*. The sampling frame of the GSS does not capture everyone in the target population. Specifically, individuals without telephone service, or a known phone number, will be excluded (Statistics Canada, 2017). This coverage error implies that the GSS may not be entirely representative of the Canadian population, and therefore, our estimates should be approached cautiously.

Sampling Approach

After creating the sampling frame, Statistics Canada needs to decide how to pick individuals from this list, since it would be too costly to survey them all. The method of selecting Canadians from the frame is called the *sampling approach*. The GSS employs a *stratified random sampling approach*. First, the target population is divided into smaller groups, called strata, where individuals share a common characteristics or property (Caetano, 2020b). In this survey, Canadians are grouped by geographic location for a total of twenty seven strata (Statistics Canada, 2017). Second, a percentage of households in each stratum are randomly selected (Statistics Canada, 2017). Third, one eligible member of each chosen household is randomly selected for participation (Statistics Canada, 2017). These last two steps employ *simple random sampling*, meaning survey participants are picked randomly, and each member of the sampling frame has an equal chance of selection (Caetano, 2020b).

There are several benefits to this method. First, the GSS uses a probability sampling approach (stratified random sampling). The odds of any one individual being included in the survey are known, and can be weighted to reflect the characteristics of the populations (Caetano, 2020a). Therefore, we know that our data reflects well Canadian family structures. Second, this approach can be replicated, allowing for verification and comparison (Caetano, 2020a). However, there are some limitations. First, probability sampling, including stratified random sampling, is very costly. The decline in GSS sample sizes since 1999 may reflect this fact (Statistics Canada, 2019). Second, random selection implies that participants may not want to participate in the survey (Caetano, 2020a). Therefore, the non-response rate may be high, and this may introduce bias into the survey (see Sampled Population and Missing data below).

Sampled Population and Missing Data

Figure 1: Distribution of Ages

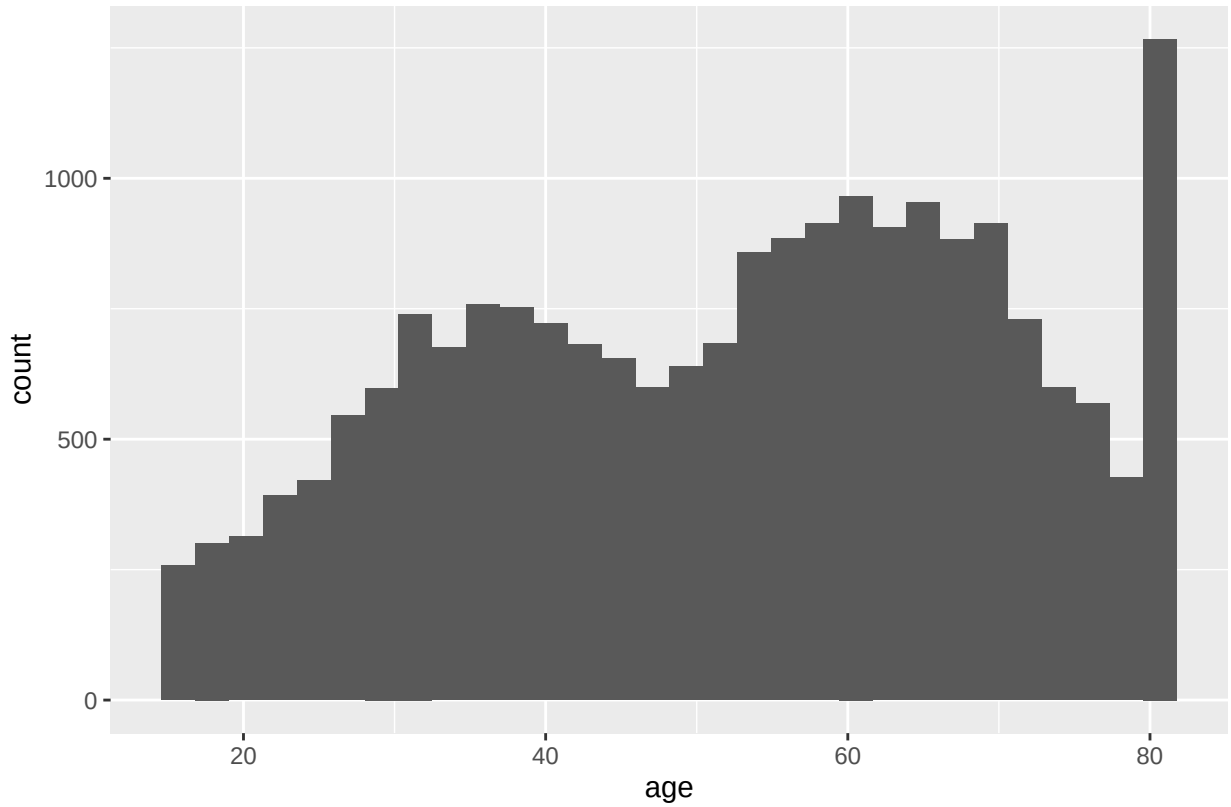


Figure 2: Distribution of Marital Status

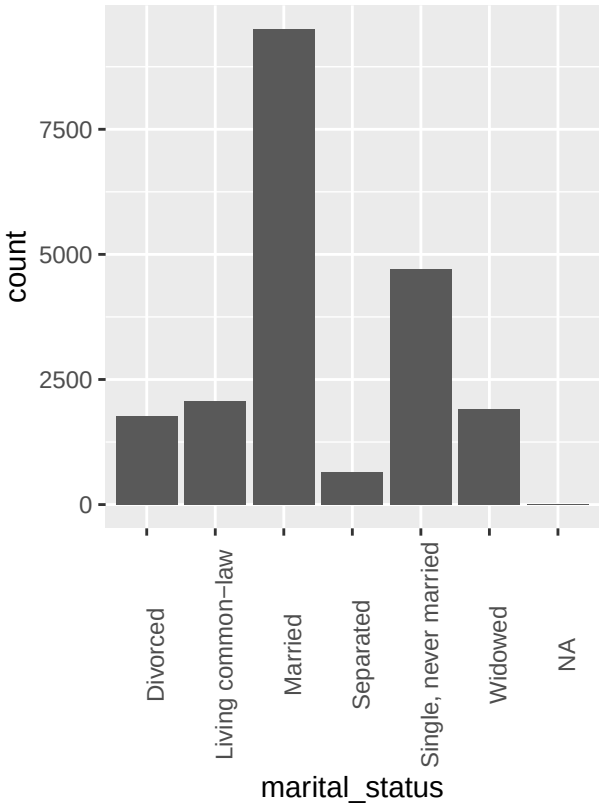
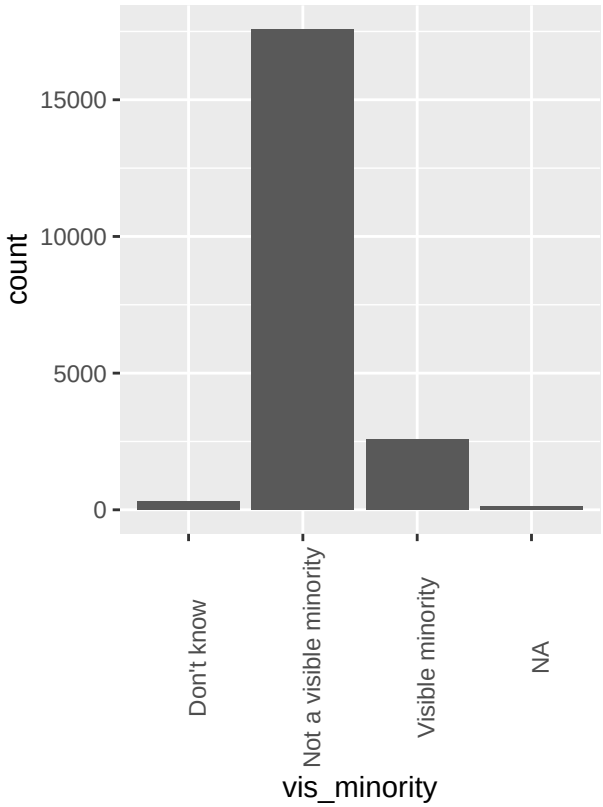


Figure 3: Distribution of Visible Minority Status



Not everyone selected to participate in the survey will do so. The set of individuals who take part are called the *sampled population* (Wu & Thompson, 2020). In the GSS, a total of $n = 20,602$ individuals completed some or all of the survey questions. This yielded a response rate of 52.6%. The non-response rate appears fairly high, raising concerns about whether some demographics were consistently left out the data. The above figures represent distributions of several key demographic variables (age, marital status, and visible minority status). We compared these distributions with results from the 2016 Census.

Age

The distribution of ages in the GSS differs somewhat from the Canadian population, as represented in the Census. There are too many older individuals in the survey, particularly those eighty-years and older (Statistics Canada, 2019). Likewise, Census data portrays a *unimodal* distribution of ages, skewed slightly toward the older side of the spectrum (Statistics Canada, 2019). In contrast, the GSS appears *bi-modal*, with spikes in the mid-thirties and early sixties.

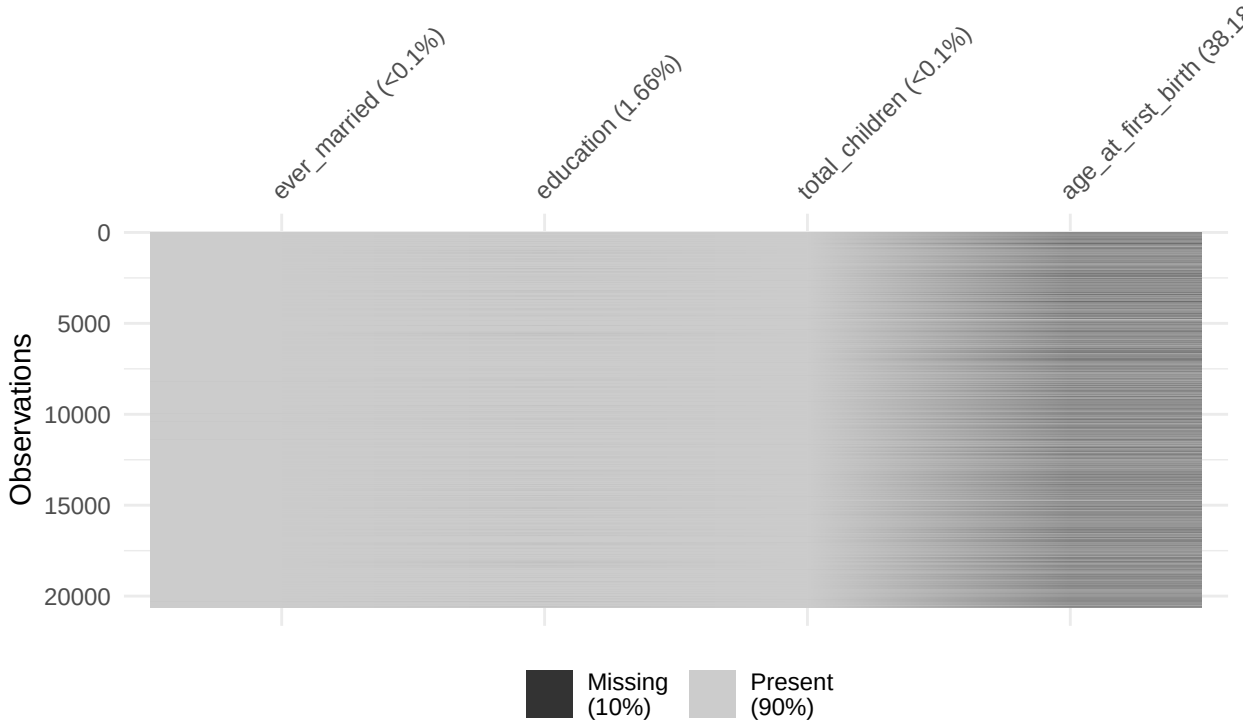
Marital Status

The sampling of marital status remains fairly consistent between the GSS and Census 2016. Specifically, 43% and 44% of respondents reported not being in a couple (divorced, separated, single, widowed) in the Census and GSS, respectively (Statistics Canada, 2019). This result indicates the GSS is representative of the Canadian population on this indicator, which is of particular importance to our analysis.

Visible Minority Status

The GSS has undersampled visible minorities. According to Census data in 2016, approximately 25% percent of respondents identified as a visible minority (Statistics Canada, 2019). Comparatively, only approximately 13% of GSS respondents identified as a visible minority.

Figure 4: Missing Data in Variables of Interest



Missing data is a concern for any survey. Individuals who agree to participate may not complete all the questions. When this happens, holes appear in the data, and this makes analysis more difficult. As seen in Figure 1, all of our predictor variables have limited missing data. In contrast, our response variable, *age_at_first_birth*, has a non-response rate of 38.18%. While this may seem large and concerning, this rate is over inflated. Many of the participants with missing data in *age_at_first_birth*, have no children. Therefore, this question was not applicable to them. In supplementary analysis (see Appendix), we find that of the $n = 20,602$ respondents, only $1,286$ individuals, or 6.2%, who had children did not respond to *age_at_first_birth*. Therefore, we consider the variable still worthy and capable of analysis.

```
## # A tibble: 4 x 3
##   `outcome_missing == 1` `outcome_valid == 1`     n
##   <lgl>                  <lgl>                <int>
## 1 FALSE                  FALSE                13074
## 2 FALSE                  NA                    6224
## 3 NA                     FALSE                 1286
## 4 NA                     NA                      18
```

Model

A multiple regression model will be used to predict the value of age_at_first_birth, by using auxilliary variables income_family, ever_married, education, and total children. This model was chosen as our goal is

to predict a numeric value by examining the relationship between multiple other predictors. ever_married is a boolean value and both ever_married and income_family are categorical values, so the regression model will use dummy variables to simulate them.

Since ever_married is boolean, it will only require one dummy variable to represent it. education is split into seven categories, and so will require six dummy variables. These will be less_than_highschool, high_school, trade, college, uni_bachelor, uni_above_bachelor. If all siz of these variables are turned off, that will indicate that the responder has an education level of uni_bachelors_degree.

The final model will take the following form:
$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \epsilon_i$

Where:
$x_1 = ever\_marriedYes$
$x_2 = college$
$x_3 = highschool$
$x_4 = less\_than\_highschool$
$x_5 = trade$
$x_6 = uni\_bachelor$
$x_7 = uni\_above\_bachelor$
$x_8 = total\_children$

Our calculated model will be an estimate of that exact model, and can be represented by:

$\hat{E}(Y_i|X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6}, X_{i7}, X_{i8}) = \hat{\beta_0} + \hat{\beta_1} X_{i1} + \hat{\beta_2} X_{i2} + \hat{\beta_3} X_{i3} + \hat{\beta_4} X_{i4} + \hat{\beta_5} X_{i5} + \hat{\beta_6} X_{i6} + \hat{\beta_7} X_{i7} + \hat{\beta_8} X_{i8}$

Y represents the output variable age_at_first_birth, which are using our model to present. The X values represent all of the auxilliary variables: ever_married, total_children, and education, which we are using as the basis of our prediction. The β's are the values our model will calculate, and represent the weight of each factor on the outcome. Larger negative or positive β values will represent that the variable has a large effect on the outcome, while smaller values will indicate it does not have much of an effect. Positive β values indicate that the predictor variable tends to increase the age at first birth, while negative values would indicate that the predicator variable tends to decrease it.

When using R to create this model, we had to choose between using the lm and the surveylm functions. The main difference between these two functions is that surveylm allows for the use of a population correction, and can take into account the sizes of stata for the stratified sampling technique. While both of those features increase accuracy, neither the population size or the individual strata size are known and so would have to be approximated. We decided it would be better to use the lm function, instead of introducing the possible error caused by estimating these values for the surveylm function.

Results

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Figure 1

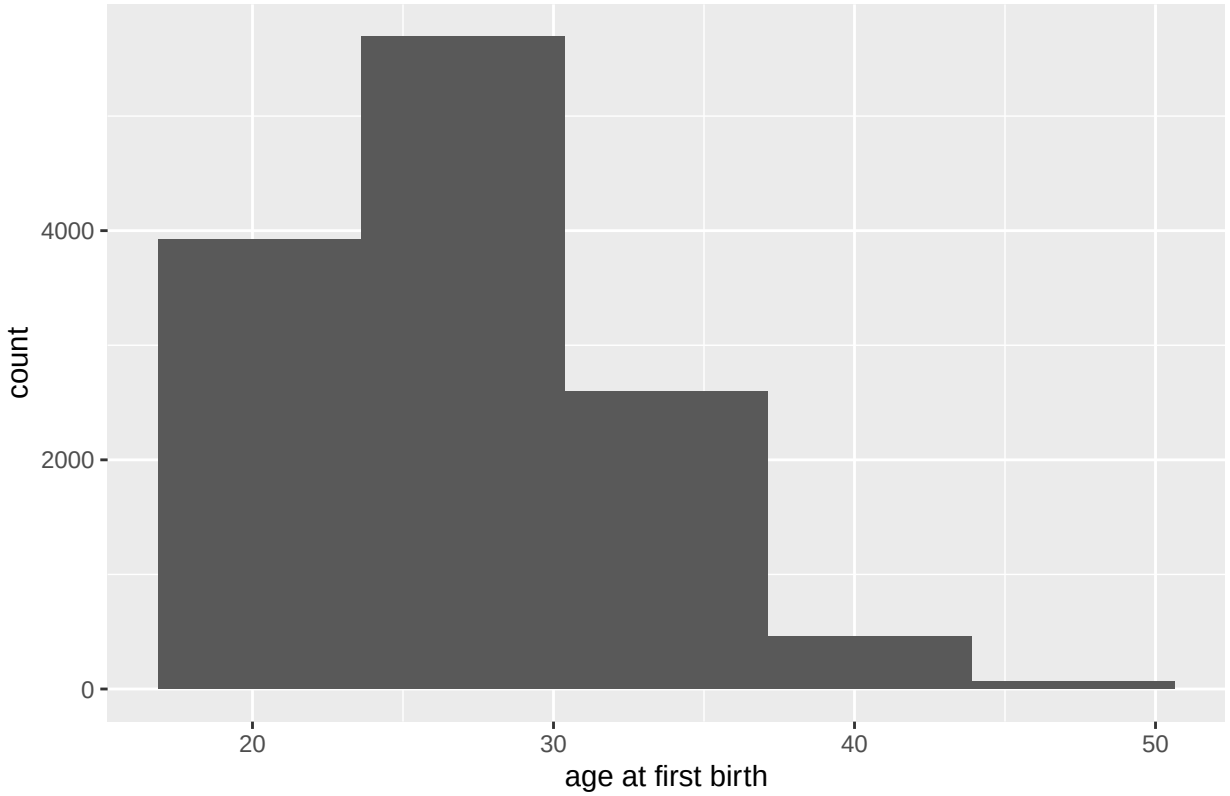
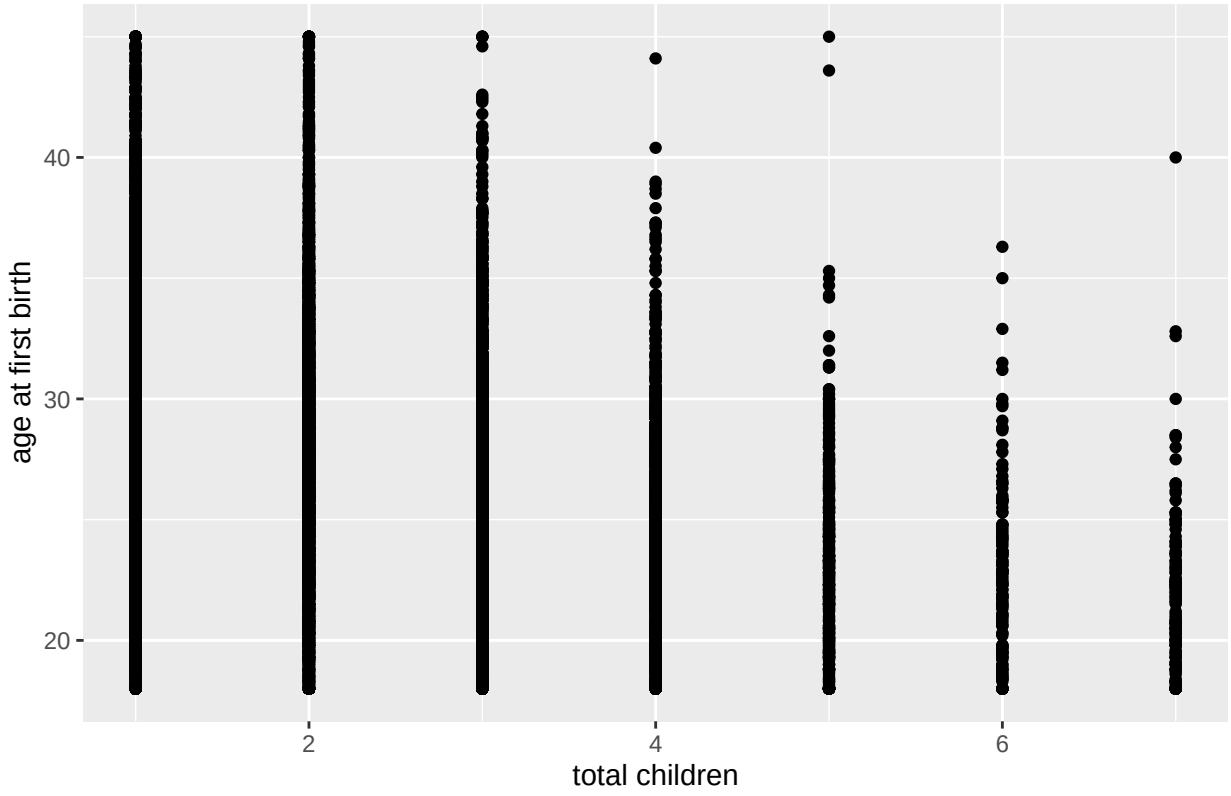


Figure 2



```
##
## Call:
## lm(formula = age_at_first_birth ~ ever_married + education +
##     total_children, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9825  -3.3168  -0.3985   2.9018  23.0780
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        32.16108    0.17365 185.205  < 2e-16 ***
## ever_marriedYes                     0.32160    0.13802   2.330   0.0198 *
## educationCollege                   -2.40253    0.13585 -17.686  < 2e-16 ***
## educationHigh School               -3.60030    0.13572 -26.527  < 2e-16 ***
## educationLess than high school     -4.48961    0.15124 -29.686  < 2e-16 ***
## educationTrade certificate         -2.52099    0.18493 -13.632  < 2e-16 ***
## educationUniversity above bachelor  0.89187    0.17723   5.032 4.92e-07 ***
## educationUniversity below bachelor -1.70838    0.23994  -7.120 1.14e-12 ***
## total_children                     -1.39208    0.03816 -36.484  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.787 on 12520 degrees of freedom
##   (204 observations deleted due to missingness)
## Multiple R-squared:  0.2182, Adjusted R-squared:  0.2177
## F-statistic: 436.7 on 8 and 12520 DF,  p-value: < 2.2e-16
```

After we cleaned the gss data using the gss_cleaning program provided by Rohan Alexander and Sam Caetano, we selected the variables we are interested in and removed the "NA" and "Don't know" values in the data.

In Figure1, we created a geom_histogram which shows the distribution of the counts of age at first birth for our sample population. Since our response variable in this study is the age at first birth, we think it's valuable to show its distribution.

In Figure2, we created a scatter plot of total children vs age at first birth. We are trying to study if there's a linear relationship between the two variables or any observable trends.

The model we chose to help analyzing our study is general linear regression model because we have multiple predictor variables and our response variable is numerical. We also assume our parameters have fixed values. The response variable is the age at first birth. We have one numerical predictor variable which is total children and we have two categorical predictor variables which are ever married and education. We created a summary of the statistics of our model to show the coefficients, intercept value and p-values of our model for further analysis. The two categorical predictor variables are split into multiple dummy variables. The reference group for variable "ever_married" is "No" and the reference group for variable "education" is "bachelor". We also renamed the categories of education into shorter names.

Discussion

The estimated regression equation we have is:

$$\hat{E}(Y_i) = 32.16108 + 0.32160 ever\_married - 2.40253 college - 3.60030 highschool - 4.48961 less\_than\_highschool - 4.48961 trade$$

From the results table, We can see that the p-value for ever_married is very big which means that it cannot support the hypothesis that ever_married and age_at first_birth are correlated. The other variables all have

really small p-values which shows that they have correlations with age_at_first_birth.

To interpret this model well, the example will be estimating a woman's age at her first birth given that she gets married and graduated with a master degree and has 2 kids. So the estimated age will be:

$$\hat{p} = 32.16108 + 0.32160 * 1 - 0.89187 * 1 - 1.39208 * 2 = 28.8$$

In this example, we can see that the estimated age for this woman's first birth is likely to be 28.8 years old.

Weaknesses

This data and study is not without its limitations. While much can be done to correct for missing data, sampling bias, and other problems, some issues are beyond our control or unfeasible given the timeline of this project. There are two particular weaknesses we want to highlight and discuss. First, our multivariate linear regression results are calculated on the assumption of census data. That is, adjustments (finite population correction) were not made to account for taking only a sample of the target population. To make these corrections, we require data on the size of each stratum. However, this information is not published by Statistics Canada, and we lack a reliable means of estimating these values. This implies that some of our results, including the slope and intercept estimates, standard errors, and significance values, may be slighly off.

Second, as discussed briefly earlier, Statistics Canada elects to exclude residents of the territories from the sample. Census data from 2016 indicates that in Nunanvut and the Northwest Territories, Indigenous peoples make up the majority of the population, at 85% and 50%, respectively (Statistics Canada, 2020). Therefore, this group is not particularly well represented in the GSS, and subsequently our analysis. In general, the GSS excludes residents of the territories from its target populations. Two recent exceptions exist to this convention, the 2009 and 2014 cycles on victimization (Statistics Canada, 2019). Future research may examine whether there are significant differences between residents of the territories and provinces on the GSS. This work could settle whether the GSS target populations are representative of the Canadian population.

Next Steps

A possible future step in terms of statistical analysis would be to test for colinearity between the predictor variabels used in our regression model. This would help to verify the accuracy of our model and that correlation exists between our predictor variables.

Going forward we also would be interested in expanding the area of our investigation and adding in additional predictor variables that we believe are relevant to the outcome. One area that we did not have a chance to explore in this report would be the effect of income on the age women have their first child at. Our analysis did not find a strong correlation between marital status and age at first birth, but potentially if we were to examine the interaction between marital status and household income, this would have a stronger relationship with age at first birth.

References

Bui, Q., & Miller, C. (2020). The Age That Women Have Babies: How a Gap Divides America (Published 2018). Nytimes.com. Retrieved 19 October 2020, from https://www.nytimes.com/interactive/2018/08/04/upshot/up-birth-age-gap.html.

Caetano, S. (2020a, September 21). Probability versus Non-Probability Sampling [PowerPoint Slides]. Quercus. https://q.utoronto.ca/courses/184060/files/8975319?module_item_id=1816946

Caetano, S. (2020b, September 28). Sampling Techniques [PowerPoint Slides]. Quercus. https://q.utoronto.ca/courses/184060/files/9058287?module_item_id=1828122

General social survey on Family(cycle31),2017

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 2.4. URL https://rmarkdown.rstudio.com.

Rohan Alexander and Sam Caetano. gss_cleaning.2020

Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://github.com/sfirke/janitor

Statistics Canada (2019, February 2019). *General Social Survey: An Overview, 2019*. Government of Canada. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm

Statistics Canada. (2020). Aboriginal Peoples Highlight Tables, 2016 Census. Retrieved October 19, 2020, from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/abo-aut/Table.cfm?Lang=Eng

Statistics Canada (2017, December 19). *General Social Survey - Family (GSS)*. Government of Canada. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816

Statistics Canada. (2019, April 3). Historical Age Pyramid. Retrieved October 18, 2020, from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/pyramid/pyramid.cfm?type=1

Statistics Canada. (2019, June 18). Census Profile, 2016 Census. Retrieved October 18, 2020, from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E

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

Wu, C., & Thompson, M. E. (2020). *Sampling Theory and Practice.* Cham: Springer International Publishing.