Report: Predicting the Sex of an Individual Based on their Age at the Birth of their First Child in New Brunswick

Colin Babineau (1003799482)

October 19th, 2020

Code and data supporting this analysis is available at: "https://github.com/colinbabineau/nbagefirstbirth"

Abstract

This analysis looks at predicting the sex of an individual based on their age at the birth of their first child in New Brunswick (this province was chosen as it is my home province). To do this, I took data from the 2017 Canadian General Social Survey and isolated for New Brunswick. I then used a logistic regression model and graphed it, which showed that there was strong evidence that there is a clear correlation between one's sex and their age at first birth, with men being more likely to have their first child later than women. More research needs to be done to make a definitive conclusion, but I do have good reason to believe that these results are significant and are useful to understanding family dynamics associated with parental age differences.

Introduction

This analysis is focused on predicting the sex of a New Brunswick resident based on their age at the birth of their first child. Specifically, we are looking to see if the age at which someone has their first child increases, whether they are consistently more likely to be male or female. This analysis is in line with the 2017 GSS which serves to understand family dynamics in Canada, as I am analyzing whether it is more common for the father or mother to be older in general in New Brunswick families. Analysis will be carried out via logistic regression modeling and the results will be discussed. This in turn could be used to further understand family dynamics in general, as age gaps of parents may impact parenting styles and how families interact.

Data

The data used for this analysis comes from the 2017 Canadian General Social Survey, isolated to just the responses from those residing in New Brunswick (my home province). The GSS data originally had 20 602 of observations on a variety of responses related to family. Once it was cut down to New Brunswick, only 1337 remained, and then 847 responses were removed for NAs for age at first birth/is_male (R Select(), Filter(), Arrange(), Pipeline with Example, n.d.). This survey was conducted from February to November 2017 (Government of Canada, 2020).

Discussion on questionnaire:

While the GSS has a lot of useful questions, there are a lot of unnecessary questions that can be answered from other questions. For example, feelings_life and self_rate_mental_health. A more succinct survey would probably be more beneficial, as approximately the same amount of information can be obtained with

the respondent more motivated to finish the survey. However, one benefit to questions that are essentially the same, is that you can find inconsistent answers and get rid of those respondent's answers (this may be indicative that they are not answering all questions honestly). There is also a question on self-rated health, which also can lead to a lot of bias.

Fortunately, for my purposes, the questionnaire suffices for answering the question of this report.

Discussion on methodology:

The target population of the GSS was all non-institutionalized Canadians over the age of 15, living in all provinces of Canada (so the territories were excluded). They stratified the survey into 27 populations across the country (only two of these strata are used in my analysis as I am only focusing on people living in New Brunswick).

The frame population of the GSS included residents associated with the phone numbers of Canadian residents that were available to Statistics Canada. Phone numbers of Canadian residents in the population frame were also obtained through the Address Registrar.

Within each stratum, simple random sampling was used. They sampled 43 000 households to guarantee a response from at least 20 000, which they were successful in achieving (Government of Canada, 2020).

While telephone numbers for surveys allows for easy access to many Canadians, there may be some bias as to who responds. Many younger generations do not answer the phone unless it is someone that they know, as well as younger generations may be busier and not want to participate. On the other hand, it does allow for a cost-effective way to survey a large amount of people.

Variables of interest

The variables of interested are is_male (whether the respondent is male or not) and age_at_first_birth (the age at which a parent was when their first child was born). Below are tables summarizing these variables (Table 1 for is_male and Table 2 for age_at_first_birth). I could have used gender as a variable instead of is_male, but this ensured a smooth process for modeling a binary variable (some responses for gender could have been misspelled, etc.).

Summary statistics of is_male

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4375  1.0000  1.0000
```

Table 1

Summary statistics of age_at_first_birth

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   18.00   21.73   25.30   25.70   28.98   42.70     463
```

Table 2

Model

I used R to program a logistic regression model to predict the probability of an individual being male based on their age at the birth of their first child. The age_at_first_birth variable is used to predict is_male. A logistic model is being used as a binary outcome is being predicted (whether someone is male or not), so a logistic regression model is appropriate for these circumstances. Other models (like linear regression) would not be conducive for binary outcomes. Logistic regression accounts for binary variables, however, while still using a continuous predictor. Age, therefore, is more useful than age groups for this model, while also being able to get more precise predictions than a range of ages. The gender variable could have been used for this model instead of is_male, but is_male ensures a binary response variable, as it uses 0 and 1 instead of strings (which can be misspelled and end up in different categories). After taking everything into consideration for the purposes of this analysis, a logistic regression model with the selected variables made the most sense.

I used svydesign and svyglm in the survey package to account for the stratification in New Brunswick (Saint John as one stratum and the remainder of the province for the other stratum).

In mathematical terms, the logistic regression model is calculating:

log(p/p-1) = b0 + b1x

Where p is the probability that the respondent is male, x is age at first birth, b1 is the change on average in log odds that the respondent is male when age at first birth is increased by 1, and b0 is the log probability that the respondent is male when x is 0 (although for obvious reasons, x will never be 0 in this model).

Unfortunately, I could not use finite population correction because I couldn't find an estimate for the number of adults in New Brunswick with at least one child, so the model doesn't adjust the standard error to the population.

In addition, since the GSS surveyed all adults over 15, not all of them had children and therefore there were some observations that could not be used as age_at_first_birth was NA.

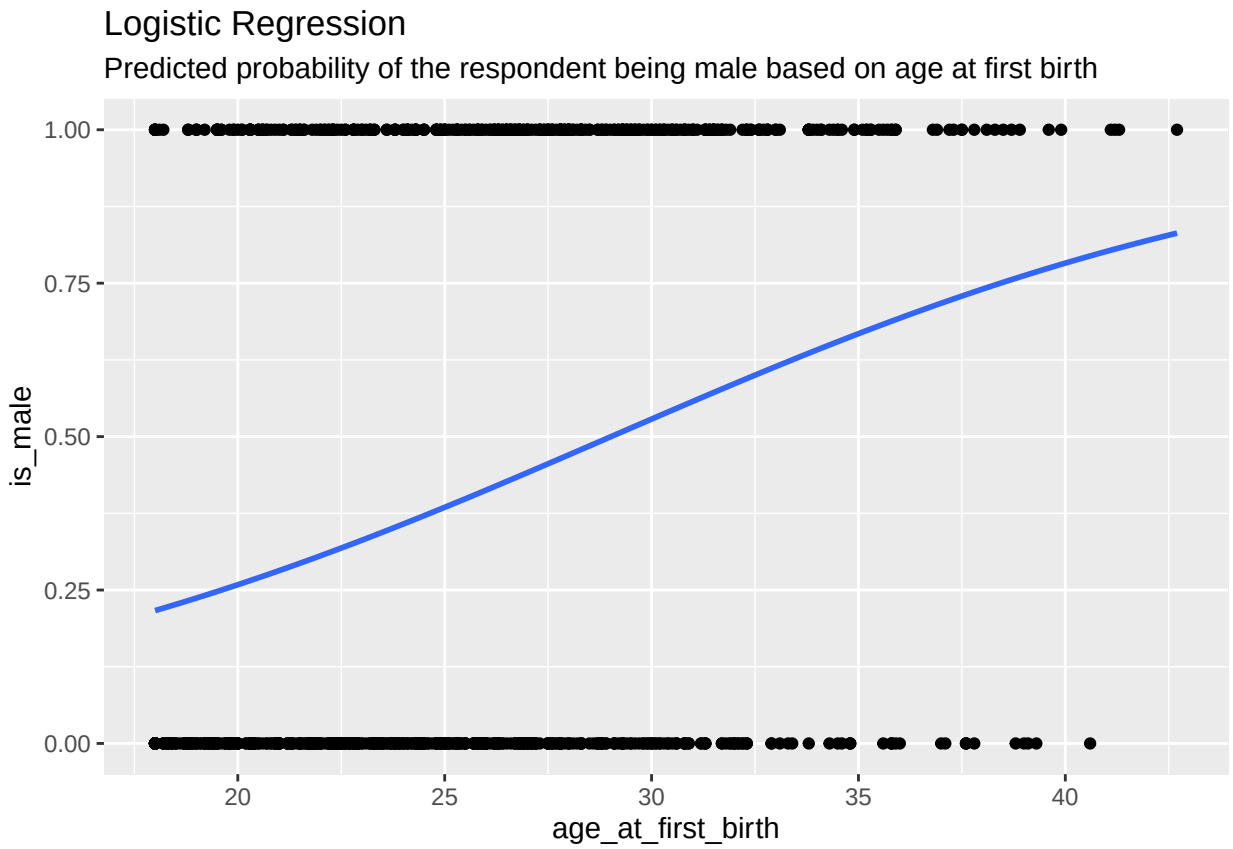
Results

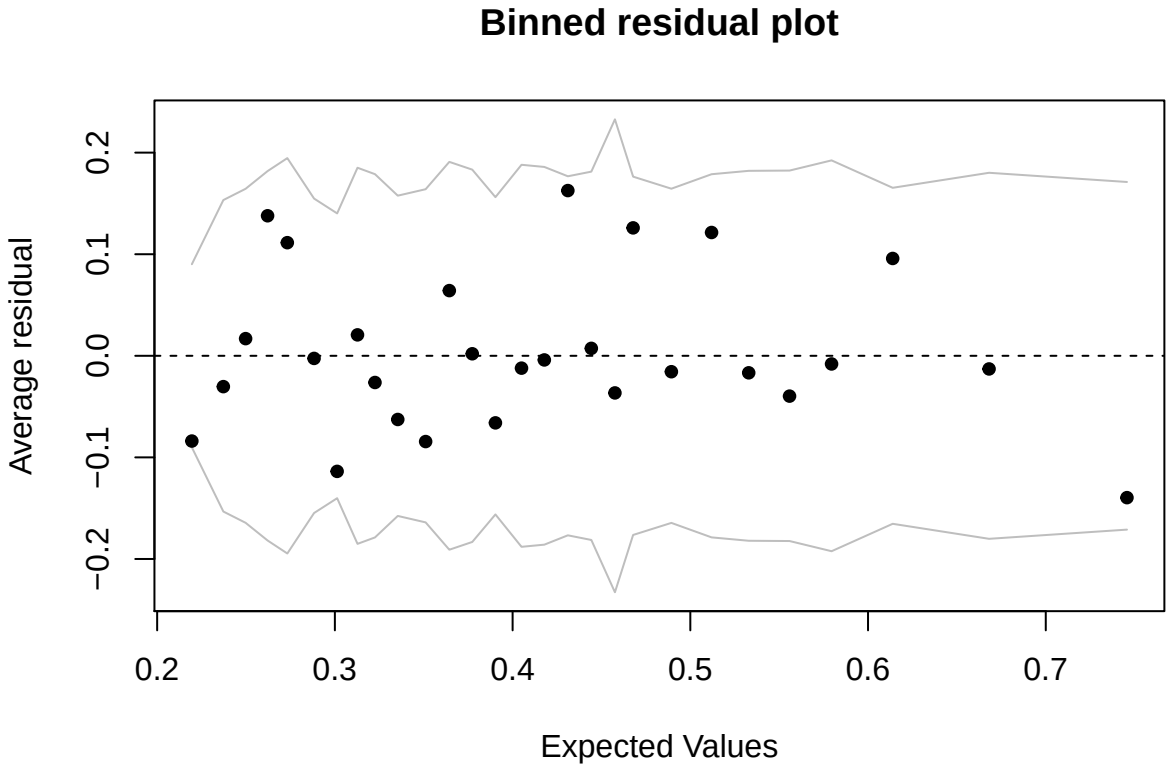
summary of logistic regression model

```
##
## Call:
## svyglm(formula = is_male ~ age_at_first_birth, design = survey.design,
##     family = "binomial")
##
## Survey design:
## svydesign(id = ~1, data = nb_data, strata = ~pop_center)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -3.38715    0.39470  -8.582  < 2e-16 ***
## age_at_first_birth  0.11672    0.01505   7.758 2.41e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1.00049)
##
## Number of Fisher Scoring iterations: 4
```

Table 3

```
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 463 rows containing non-finite values (stat_smooth).
## Warning: Removed 463 rows containing missing values (geom_point).
```





Plot 1 and Plot 2 (Respectively)

Here we see the summary of the logistic model (Table 3), and a graph which shows a graph of the logistic regression that predicts the probability that an individual's sex is male based on the age at first birth (Plot 1). We also see a binned residual plot for logistic regression diagnostics, showing where we would expect to contain 95% of the observations (Plot 2) (Modify axis, legend, and plot labels - labs., n.d.).

The summary statistics for is_male and age_at_first_birth can be viewed above in Table 1 and Table 2.

Discussion

Nothing seems violated in terms of logistic regression properties. The S shape curve is prominent in the prediction of probability. The p-values are very significant in the summary table (essentially zero), indicating that these results are most likely by chance alone, meaning we reject the null hypothesis that the probability of someone being male is 0.5 for all values of age at first birth.

The summary table for the model indicates that for each year added to age at first birth in New Brunswick, on average, the log odds increase by 0.11672.

Additionally, the binned residual plots show no obvious violations of this model. The average residuals hover around 0 with relatively constant variance. They also all fall within the grey bands which is where we would expect to contain 95% of the observations (Webb, 2017).

This indicates that in New Brunswick, as the age that someone has their first child increases, the more likely they are to be male, which is exactly the goal that we were attempting to find. We wanted to know if there was a connection between the likelihood of a parent being male based on how old they are when they have their first child, and this shows that the older a parent is when they have their first child, the more likely they are to be male. By my model, it is expected that only about 25% of new parents around the age of 20 are male, but at 35, around 70% of new parents are expected to be male.

Weaknesses

One issue with this study is that the sample size is not that large, at only 847 people. Although it is not particularly small, there are definitely many potential observations that are missing and the results could be slightly off as a result.

In addition, we do not know an estimate for all individuals in New Brunswick over 15 who have at least one child, so we were unable to get a finite population correction to get a more accurate standard error.

It should also be noted that the distribution is not even for those who are male and those who aren't, as only 44% of the New Brunswick Residents surveyed in the GSS were male, possibly skewing the results slightly and giving some bias to predicting sex (although it is still quite close to 50%).

Next Steps

For next steps, there are a variety of ways of proceeding after this report. For example, another survey could be taken in New Brunswick to get a larger sample size, as well as ensuring a better distribution of sexes by stratifying for male and female and ensuring equal numbers in each stratum. This would allow us to get a more accurate standard error and ensure less skewed data.

In addition, other surveying techniques could be implemented beyond telephone calls, such as online surveys, in-person surveys, or other methods to get a broader range of respondents.

These steps could also be taken on a larger scale outside of New Brunswick to see if this trend is applicable to other geographic locations.

References

Government of Canada, S. (2020, April 30). General Social Survey – Family (GSS). Retrieved October 11, 2020, from https://www.statcan.gc.ca/eng/survey/household/4501

Modify axis, legend, and plot labels - labs. (n.d.). Retrieved October 12, 2020, from https://ggplot2.tidyverse.org/reference/labs.html

R Select(), Filter(), Arrange(), Pipeline with Example. (n.d.). Retrieved October 11, 2020, from https://www.guru99.com/r-select-filter-arrange.html

Webb, J. (2017, September 03). Course Notes for IS 6489, Statistics and Predictive Analytics. Retrieved October 12, 2020, from https://bookdown.org/jefftemplewebb/IS-6489/logistic-regression.html