Title of Your Report

Wing Yi Ma, Qiuyun Han, Tong Wu, Minghui Yu

October 19th 2020

Title of your Report

Names of Authors

Wing Yi Ma, Qiuyun Han, Tong Wu, Minhui Yu

Date

October 19th 2020

Abstract

The falling population is a common phenomenon in the developed countries, and it brings a series of social problems on aspects such as labour force and aging of population. We extract data from the 2017 General Social Survey on Family, a sample survey with cross-sectional design which collects a large amount of data from each selected respondent. To investigate the factors that affect the decreasing population, we construct a multivariate regression model on the respondent's total number of children. And we have evidence to conclude that respondent's age and household size are positively correlated to total number of children, respondents with high education level and respondents with single status tend to have fewer number of children.

Introduction

The goal of this report is to find how the factors will relate to the total number of children in Canadian households. We first get the GSS survey data of Canadian Family in 2017, then we made a clean of the data and the raw data after cleaning is called "gss". We have made this clean because we want to remove the insignificant, invalid observations or variables that are not useful to our report, also we rename every variable in the new data so we can visualize the dataset. In the report, we made analysis on some variables that we believe important to see how they correlated with the total number of children in families, such as the age of respondent, the household size of the respondent, and we also investigate their marriage status and educational level. We have used a multivariate regression method to measure the relations, in our model, total children number is the response variable and the age, household size, marriage status and educational level are our expansionary variables.

Data

```
## Rows: 20,602
## Columns: 81
## $ caseid                        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,...
## $ age                           <dbl> 52.7, 51.1, 63.6, 80.0, 28.0, 63.0...
## $ age_first_child               <dbl> 27, 33, 40, 56, NA, 37, 40, 59, NA...
## $ age_youngest_child_under_6    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ total_children                <dbl> 1, 5, 5, 1, 0, 2, 2, 7, 0, 1, 0, 0...
## $ age_start_relationship           <dbl> NA, NA, NA, NA, 25.3, NA, NA, NA, ...
## $ age_at_first_marriage            <dbl> NA, NA, NA, NA, NA, NA, NA, 22.1, ...
## $ age_at_first_birth               <dbl> 25.9, NA, 23.2, 27.3, NA, 25.8, 18...
## $ distance_between_houses          <dbl> 30, NA, NA, NA, NA, NA, NA, NA, NA...
## $ age_youngest_child_returned_work <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ feelings_life                    <dbl> 8, 10, 8, 10, 8, 9, 4, 10, 8, 5, 1...
## $ sex                              <chr> "Female", "Male", "Female", "Femal...
## $ place_birth_canada               <chr> "Born in Canada", "Born in Canada"...
## $ place_birth_father               <chr> "Born in Canada", "Born in Canada"...
## $ place_birth_mother               <chr> "Born in Canada", "Born in Canada"...
## $ place_birth_macro_region         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ place_birth_province             <chr> "Quebec", "Ontario", "Ontario", "A...
## $ year_arrived_canada              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ province                         <chr> "Quebec", "Manitoba", "Ontario", "...
## $ region                           <chr> "Quebec", "Prairie region", "Ontar...
## $ pop_center                       <chr> "Larger urban population centres (...
## $ marital_status                   <chr> "Single, never married", "Married"...
## $ aboriginal                       <chr> "No", "No", "No", "No", "No", "No"...
## $ vis_minority                     <chr> "Not a visible minority", "Not a v...
## $ age_immigration                  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ landed_immigrant                 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ citizenship_status               <chr> "By birth", "By birth", "By birth"...
## $ education                        <chr> "High school diploma or a high sch...
## $ own_rent                         <chr> "Owned by you or a member of this ...
## $ living_arrangement               <chr> "Alone", "Spouse only", "Spouse on...
## $ hh_type                          <chr> "Low-rise apartment (less than 5 s...
## $ hh_size                          <dbl> 1, 2, 2, 2, 2, 2, 1, 1, 1, 6, 5, 1...
## $ partner_birth_country            <chr> "Canada", "Canada", "Canada", "Can...
## $ partner_birth_province           <chr> "Quebec", "Manitoba", "Ontario", "...
## $ partner_vis_minority             <chr> "Not a visible minority", "Not a v...
## $ partner_sex                      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ partner_education                <chr> "Trade certificate or diploma", "B...
## $ average_hours_worked             <chr> "30.0 to 40.0 hours", "50.1 hours ...
## $ worked_last_week                 <chr> "Yes", "Yes", "No", "No", "No", "N...
## $ partner_main_activity            <chr> "Working at a paid job or business...
## $ self_rated_health                <chr> "Excellent", "Good", "Very good", ...
## $ self_rated_mental_health         <chr> "Excellent", "Good", "Good", "Very...
## $ religion_has_affiliation         <chr> "Has religious affiliation", "Don'...
## $ regilion_importance              <chr> "Somewhat important", "Don't know"...
## $ language_home                    <chr> "French", "English", "French", "En...
## $ language_knowledge               <chr> "French only", "English only", "Bo...
## $ income_family                    <chr> "$25,000 to $49,999", "$75,000 to ...
## $ income_respondent                <chr> "$25,000 to $49,999", "Less than $...
## $ occupation                       <chr> "Sales and service occupations", "...
## $ childcare_regular                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ childcare_type                   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ childcare_monthly_cost           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ ever_fathered_child              <chr> NA, "Yes", NA, NA, "No", NA, NA, N...
## $ ever_given_birth                 <chr> "Yes", NA, "Yes", "Yes", NA, "Yes"...
## $ number_of_current_union          <chr> NA, NA, NA, NA, "Second union", NA...
## $ lives_with_partner               <chr> "No", "No", "No", "No", "Yes", "No...
## $ children_in_household            <chr> "No child", "No child", "No child"...
## $ number_total_children_intention  <dbl> NA, NA, NA, NA, 2, NA, NA, NA, NA,...
## $ has_grandchildren                <chr> "No", "Yes", "Yes", "No", "No", "Y...
## $ grandparents_still_living        <chr> "No", "No", "No", "No", "Yes", "No...
## $ ever_married                     <chr> "No", "Yes", "Yes", "Yes", "No", "...
## $ current_marriage_is_first        <chr> NA, "Yes", "Yes", "Yes", NA, "Yes"...
## $ number_marriages                 <dbl> 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0...
## $ religion_participation           <chr> "Once or twice a year", "Don't kno...
## $ partner_location_residence       <chr> "In the same province", NA, NA, NA...
## $ full_part_time_work              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ time_off_work_birth              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ reason_no_time_off_birth         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ returned_same_job                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ satisfied_time_children          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ provide_or_receive_fin_supp      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_child_supp              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_child_exp               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_lump                    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_other                   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ fin_supp_agreement               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ future_children_intention        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ is_male                          <dbl> 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0...
## $ main_activity                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ age_diff                         <chr> NA, "Respondent is 4 years older",...
## $ number_total_children_known      <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1...
```

The data set used in this report is the 2017 General Social Survey (GSS) on Family. The data was collected through telephone calls from February 2nd to November 30th 2017, with telephone numbers provided by Statistics Canada’s Address Register, which is the sampling frame of this survey.The population of the survey is the entire Canadian population. Its target population was the non-institutionalized persons who were 15 years of age and older and lived in the 10 provinces of Canada, excluding residents of the Yukon, Northwest Territories, and Nunavut, and the full-time residents of institutions. The frame population was the owners of the telephone numbers that were called, and the sampled population was the ones who answered the survey question through the calls. The sampling method used was stratified random sampling. The 10 provinces were each divided into strata according to whether they are Census Metropolitan Areas (CMAs) or not. Some of the CMAs were grouped together by extra rules and the non-CMAs were grouped to form 10 more strata, resulting in a total of 27 strata. Such sampling method is not the most ideal because the division of strata for the smaller CMAs and the non-CMAs is not as systematic as the others, when some unrelated areas were grouped together. As well, the method for data collection allows for inaccurate data if respondents choose to not provide true information through phone calls.

In the data set, there are in total 461 variables and 81 of them were selected for our study after cleaning, and 20602 observations were collected. This data set was chosen for our study because it is large and the most recent for our topic. A major drawback of this data set is that not all observations were useful when the survey provides irrelevant options to questions, such as “do not know” and “valid skip”. Due to that, we have to clean the data to exclude useless observations, which would then decrease the size of our sample and lead to less accurate results.

Code and data support this report is available at https://github.com/QiuyunHan/PS2.

Model

We used a multivariate linear regression model to measure the correlation between the total children of respondents and their age, size of household, marital status and degree of education. We use this model because we believe these variables would impact the number of children, and there is a linear relation between the variables. A multivariate linear regression model can predict the linear relation of more than one independent variable and the dependent variable. In our model, the total number of children is considered the response(dependent) variable, which is numeric. Other variables are used as independent variables, more specifically, the age and size of household are numerical variables. Marriage status and educational level are dummy variables, which means they are the categorical variables in the regression model. There are reasons why we choose these variables. Firstly, we choose age rather than age group because age is a continuous numerical variable, it is more accurate than the age group. Same reason for choosing the household size. Besides, we believe the degree of educational level and his marital status would affect his decision on being a parent, thus it would impact the total number of children in the household. We used R to run this model as learnt in the lectures. Since this model was built based on a survey data and they have used stratified sampling during the survey, we used survey package and the stratified method to build the mode. However, this survey has used different province as the stratum, we considered every province have the same weight in our model, and we were build this without the finite population correction. For the alternative model, we believe logistic regression model is also appropriate, but we need to turn the response variable as binary, for example if the respondent has more than 5 children or not, and we can use the same expasionary variable to predict the result.

By using the linear model, we have predicted a formula.

(1) Formula of Our Model:

total_children = β_0 + $\beta_1$age + $\beta_2$hh_size + β_3College, CEGEP or other non-university certificate or di… + β_4High school diploma or a high school equivalency certificate + β_5Less than high school diploma or its equivalent + β_6Trade certificate or diploma + β_7University certificate or diploma below the bachelor's level + β_8University certificate, diploma or degree above the bach… + β_9Living common-law + $\beta_1$0Married + $\beta_1$1Separated + $\beta_1$2Single, never married + $\beta_1$3Widowed

(2) Interpretation of the Model:

Total_children is our response variable, age,hh_size,education and marital status are expansionary variables. + β_0 : This is the intercept for our model + β_1 : This is the parameter for variable age + β_2 : This is the parameter for variable household size + β_3 to β_8: These are the parameter for variable education in different degrees. + β_9 to $\beta_1$3: These are the parameter for variable marital_status represents different status of marriage of repondents.

Results

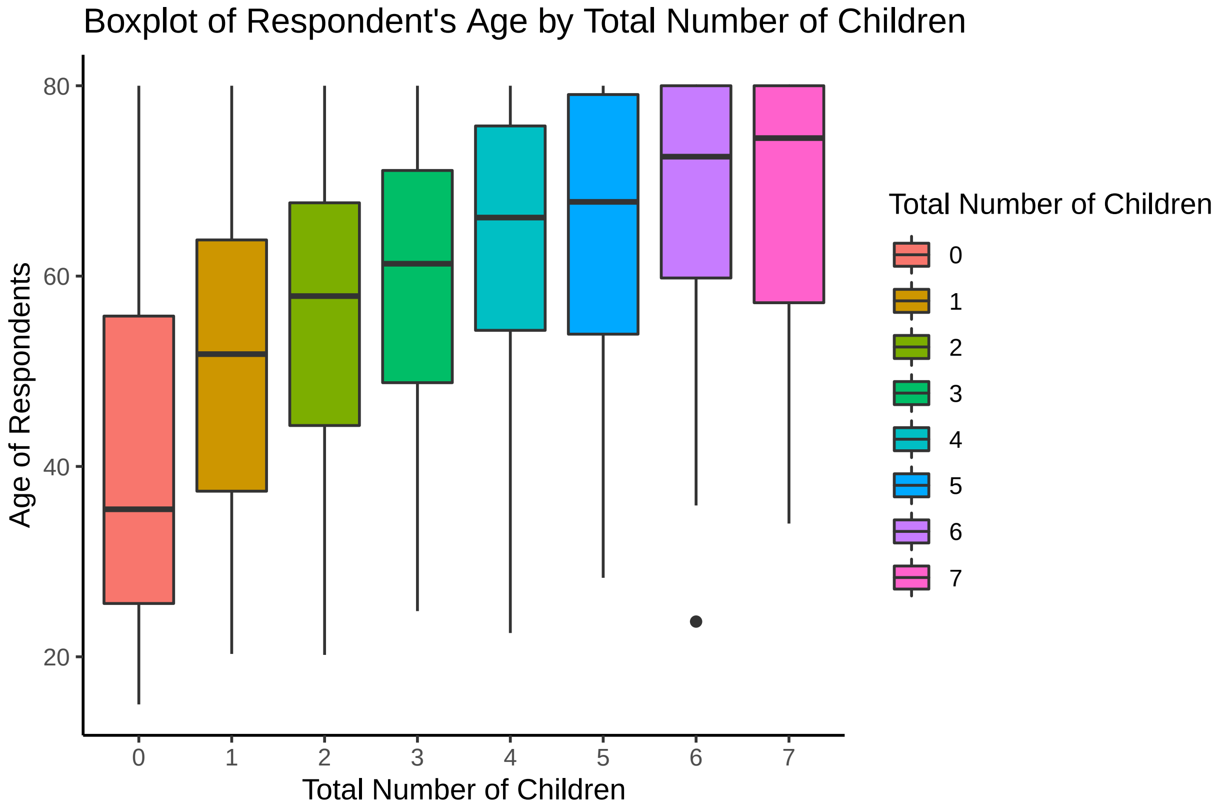

Figure 1 Boxplot of age of respondents grouped by number of total children

From the box plot of age of respondents by each number of children(Figure 1), we can observe the distribution of age in each number of children. We can see that respondents without kid has lowest age median, which is around 35 years old. And the respondents who has seven or more children are around 75 years old. There is one outlier, a respondent who has 6 kids is approximately 30 years old, while most resondents' age has the median of 73 years old with 6 children.

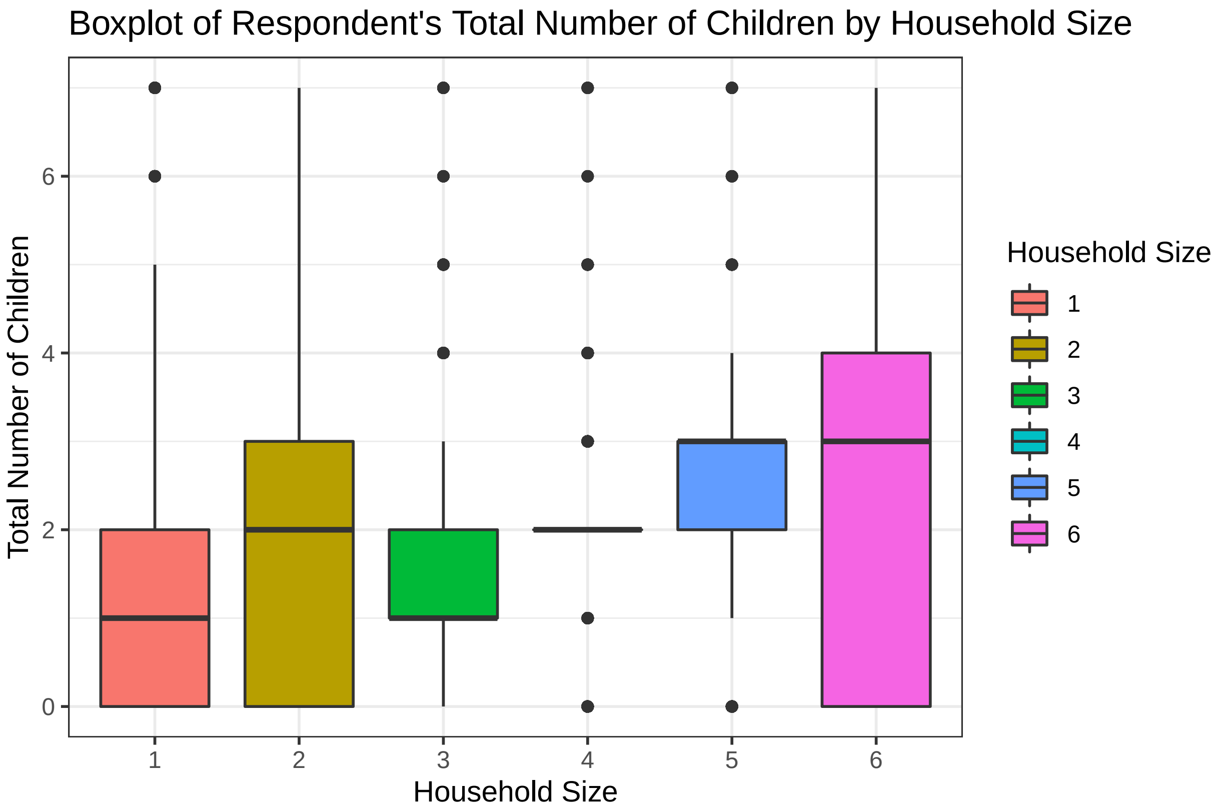

Figure 2 Boxplot on respondent's total number of children grouped by household size

Above is the boxplot of total number of children versus household size(Figure 2). We can see from the boxplot that small household size generally have smaller amount of children, but it does not vary much between groups. Household size of 5 and 6 have the same mean ammount of children around 3, and household size of 1 and 3 have the same amount of children around 1.

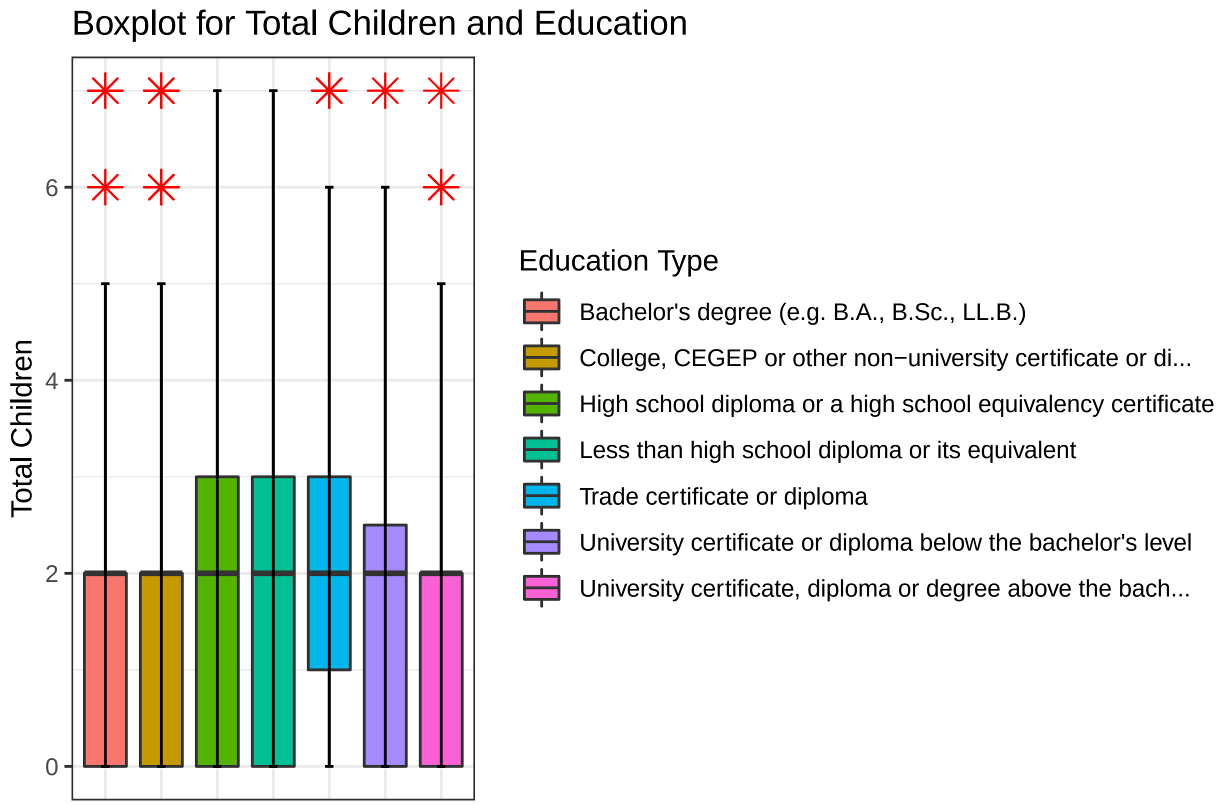

Figure 3 Boxplot of total children

Figure 3 shows that for each type of education the lower quartile are all 0 except “Trade certificate or diploma”, but the upper quartile are different which is 2 or 3. The median is 2. The lower bound are all 0 , and the upper bound are 5, 6, or 7. There are outliers in the first three types and the last two types and all the outliers are at the larger value parts. For a sample with a standard normal distribution, only a few points are outliers. The more outliers, the heavier the tail and the smaller the degree of freedom (i.e. the number of free changes). However, skewness indicates the degree of deviation. Since the outliers are concentrated on the side with larger values, the distribution is right-skewed. We can get a simple conclusion that the higher the education, the less the number of children, however, most family will choose two children.

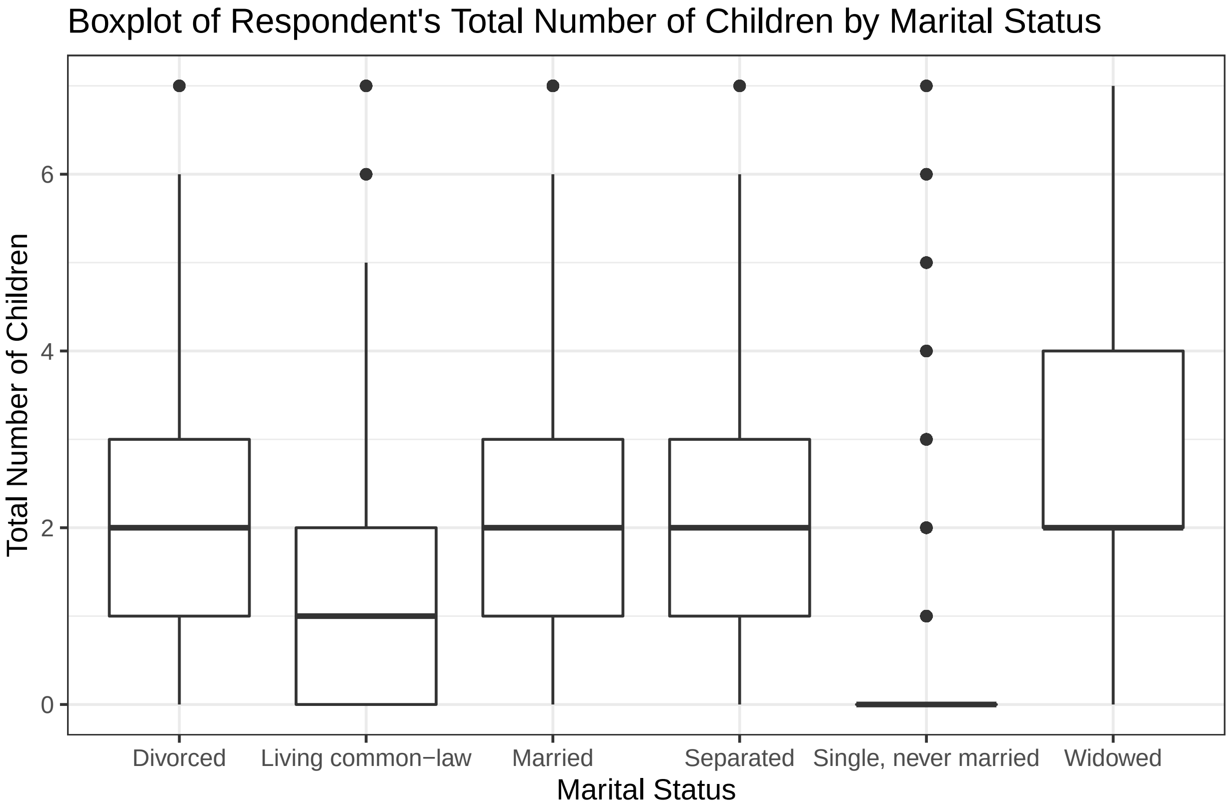

Figure 4 Boxplot of respondent's total number of children grouped by marital status

From the box plot of marital_status verses total_children(Figure 4), we can see that the medians of most marital status lie around 2 other than "living common-law" and "single,never married". Also, most "widowed" seem to have more children than other status while most "single, never married" have the least. Most of the respondents who are "Single, never married" have 0 children but there are still some of them with different number of children.

Table 1: Summary of Regression Model on Total Children

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.123	0.057	-19.593	0.000
age	0.038	0.001	60.045	0.000
hh_size	0.434	0.009	49.160	0.000
educationCollege, CEGEP or other non-university certificate or di…	0.168	0.022	7.493	0.000
educationHigh school diploma or a high school equivalency certificate	0.227	0.023	9.810	0.000
educationLess than high school diploma or its equivalent	0.386	0.031	12.523	0.000
educationTrade certificate or diploma	0.323	0.034	9.387	0.000
educationUniversity certificate or diploma below the bachelor's level	0.130	0.046	2.851	0.004
educationUniversity certificate, diploma or degree above the bach…	-0.063	0.027	-2.292	0.022
marital_statusLiving common-law	-0.456	0.043	-10.706	0.000

	Estimate	Std. Error	t value	Pr(>|t|)
marital_statusMarried	-0.280	0.036	-7.843	0.000
marital_statusSeparated	0.259	0.062	4.196	0.000
marital_statusSingle, never married	-1.064	0.038	-27.979	0.000
marital_statusWidowed	0.174	0.048	3.619	0.000

R Squared Computation

```
## [1] 0.3957997
```

Above is the summary table of the multivariate regression model on respondent's total number of children, with four regressor: respondent's age, household size, education level and marital status. And the computed R squared value is 0.396.

Discussion

(1) Formula of Our Model:

total_children = β_0 + $\beta_1$age + $\beta_2$hh_size + β_3College, CEGEP or other non-university certificate or di… + β_4High school diploma or a high school equivalency certificate + β_5Less than high school diploma or its equivalent + β_6Trade certificate or diploma + β_7University certificate or diploma below the bachelor's level + β_8University certificate, diploma or degree above the bach… + β_9Living common-law + $\beta_1$0Married + $\beta_1$1Separated + $\beta_1$2Single, never married + $\beta_1$3Widowed

- In this formula, age and hh_size are numerical variables and others are dummy variables. And the estimate value, the standard error, the t-value and the p-value of each variable are shown as the table above.

- For example, Living common-law = 1, if the marital status of the person being interviewed is Living common-law; Living common-law = 0, otherwise.

(2) Interpretation:

- β_0 = -1.123: This is the intercept for our model which represent that when age = 0, hh_size = 0, education = "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" and marital_status = "Divorced", total_children = -1.123. However there is no real meaning here, because there will be no -1.123 children.
- β_1 = 0.038: If age of the person being interviewed increases by one unit, then total_children increases by 0.038 units.
- β_2 = 0.434: If hh_size of the person being interviewed increases by one unit, then total_children increases by 0.434 units.
- β_3 ~ β_8:
- These are all dummy variables for education.
- The baseline is "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)".
- Since β_3 = 0.168, then the meaning of β_3 is if the education type change from "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" to "College, CEGEP or other non-university certificate or di… ", total_children will increase by 0.168 with other conditions remain unchanged. The explanation of β_4 ~ β_8 is the same.
- Since β_4 - β_3 = 0.227 - 0.168 = 0.059, we can also say that if the education type change from "College, CEGEP or other non-university certificate or di… " to "High school diploma or a high school equivalency certificate", total_children will increase by 0.059 with other conditions remain unchanged. Same for β_3 ~ β_8
- β_9 ~ $\beta_1$3: *These are all dummy variables for marital status.
- The baseline is "Divorced".

- Since β_9 = -0.456, then the meaning of β_9 is if the marital status change from "Divorced" to "Living common-law", total_children will decrease by 0.456 with other conditions remain unchanged. The explanation of $\beta_1$0 ~ $\beta_1$3 is the same.
- Since $\beta_1$0 - β_9 = -0.280 - (-0.456) = 0.176, we can also say that if the marital status change from "Living common-law" to "Married", total_children will increase by 0.176 with other conditions remain unchanged. Same for other combinations for β_9 ~ $\beta_1$3.

(3) Evaluation for Our Model:

- R^2 = 0.396: We can say that 39.6% variation in total_children can be explained by our model.
- Pr(>|t|) - p-value: Since all variables' p-value is smaller than 0.05, then we can say that all variables are statistically significant.

(4) Conclusion:

Our conclusion is that the larger the age, the larger the hh_size, the lower the education level, and the better the marital status, the greater the total number of children in Canadian families. This model can be used for reference by people of different ages, for example, for young couples, they will have the expected number of children about them. In other words, through our model, people can find a reference value for the number of total children. Combined with our plots, we can see that the proportion of people with a high education level will be 50%, and single, never married will account for 23%, which will result in fewer children in Canadian families. If things go on like this, fewer children will lead to more serious social problems such as population aging and a decline in labor, so the government can encourage people to have more children.

Table 2: Table for Respondents' Education

Education	Count
Bachelor's degree (e.g. B.A., B.Sc., LL.B.)	3752
College, CEGEP or other non-university certificate or di…	4563
High school diploma or a high school equivalency certificate	4843
Less than high school diploma or its equivalent	3033
Trade certificate or diploma	1483
University certificate or diploma below the bachelor's level	731
University certificate, diploma or degree above the bach…	1835

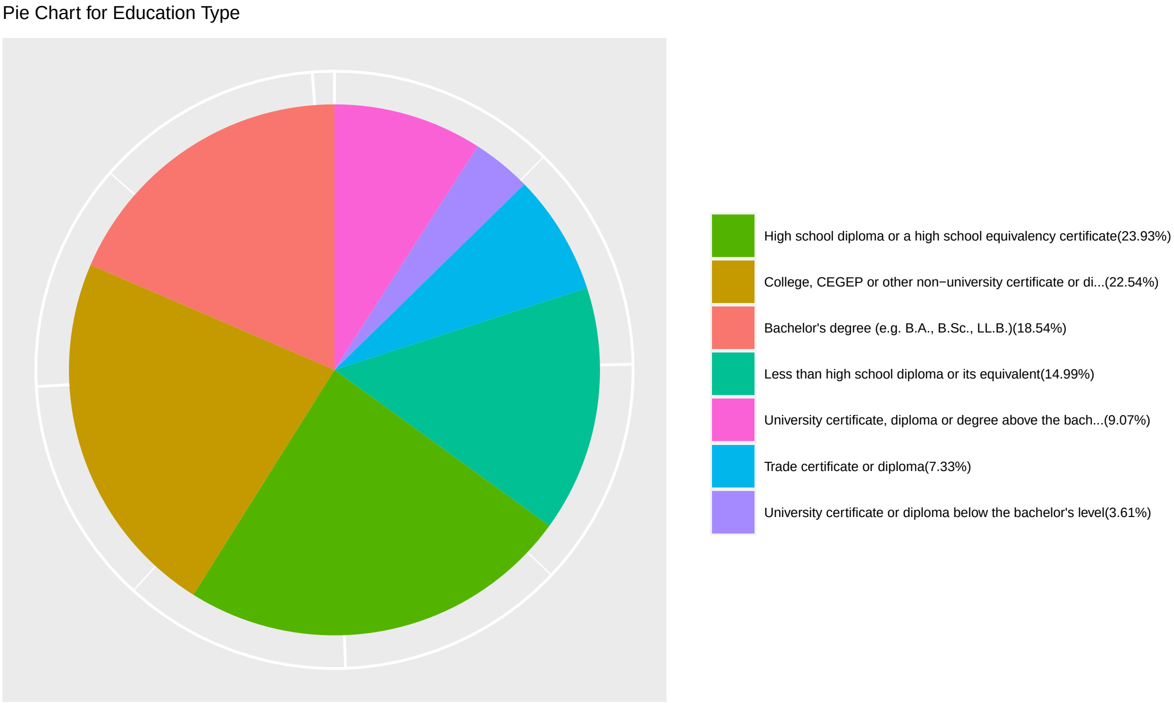

Figure 5 pie chart of respondents' education

From this pie chart we can observe that if it is based on whether or not to complete college, the proportion of people with high education and low education is basically 50%.

Table 3: Table for Respondents' Marital Status

Marital Status	Count
Divorced	1734
Living common-law	2030
Married	9332
Separated	629
Single, never married	4647
Widowed	1868

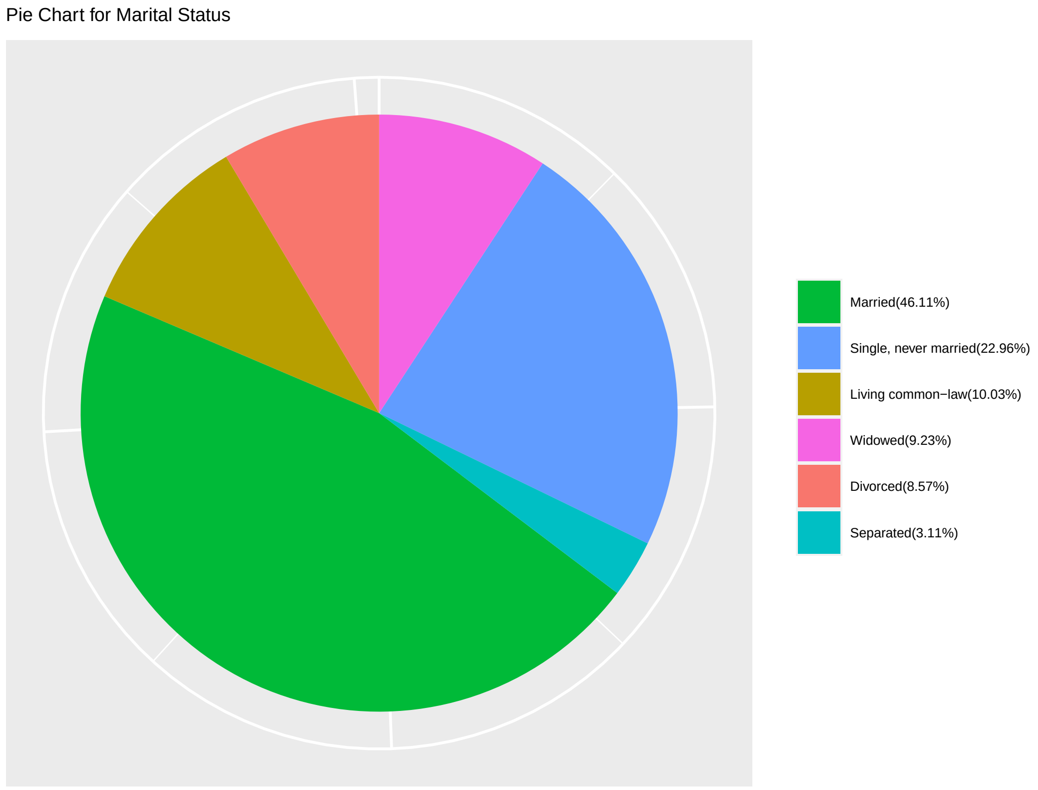

Figure 6 Pie chart of respondents' marital status

From this pie chart we can conclude that married people accounted for the largest proportion, reaching 46%, followed by single, never married people, accounting for 23%, almost half of the number of married people.

Weaknesses

Some limitations and drawbacks of the survey and the sampling method are as mentioned in the data section. As well, the sampled population includes people from different pronvinces in Canada. The various environments in Canada might also be a factor on parents' decisions with the number of children they would like to have. Therefore, our model might not have included all factors that possibly influence the total number of children in families. Another weakness of our model would be that linear regression models are sensitive to outliers, as we can see from the plots of each predictor variable versus total_children, the existence of outliers was obvious. This could have lowered the accuracy of our results. Finally, we did not use fpc in our model, which might have caused the standard errors of the estimates to be too big.

Next Steps

Our model is a stratified sampling without finite population correction. So we can add this part to strengthen our model and make the model more reasonable. Since our model is only based on 2017, if we want to obtain more information on relationship of other variables or want to see if our conclusion of the model is correct, we need follow-up surveys in 2018 and after. And the more data we have the more accurate our model is. To make our model more convincing, we can investigate more people. Expanding the data range will make the results more accurate.

References

Dataset (MLA)

Beaupre, Pascale. *General Social Survey Cycle 31: Families*, *2020*. Statistics Canada Minister of Industry [distributor]. Web. April 2020.

Software

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Packages

Dabao Zhang (2020). rsq: R-Squared and Related Measures. R package version 2.0. https://CRAN.R-project.org/package=rsq

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

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.

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

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

Websites

Yihui Xie, Christophe Dervieux. "R Markdown Cookbook." 10.1 The Function Knitr::Kable(), 21 Sept. 2020, bookdown.org/yihui/rmarkdown-cookbook/kable.html.

Telling Stories With Data, 17 May 2020, tellingstorieswithdata.com/.