Socioeconomic Determinants of Total Children in a Family

Kashaun Eghdam, Ankhee Paul, Timothy Regis, Chen Shupeng

10/18/2020

Abstract

Current social trends indicate the undertaking of family planning among the Canadian population. In this paper, we use the Canadian General Social Survey to conduct a regression analysis to determine the impact of social-demographic variables such as age, income level, and education level of respondents on the total number of children in their family. We discovered that age is a strong predictor of total children in a family and we found a positive correlation for the income levels and negative correlation for education levels regarding total children. Our findings are quite relevant as they can help to highlight the importance of family planning among Canadians.

Introduction

The Canadian General Social Survey is a program conducted by Statistics Canada that aims to gather data through a series of annually independent, cross-sectional surveys, on social trends, exploring specific themes in depth. The objective is to monitor changes in the living conditions of Canadian citizens and provide information on key social policy issues. Initially, using the Random Digit Dialling method to gather data from Canadian citizens aged 15 years and over, living in private households in each of the provinces, the mode of collection has now changed to Computer Assisted Telephone Interviewing and internet questionnaires.

The GSS focuses on several specific themes such as caregiving and care receiving and families, to name a few. It is quite common for individuals to determine their family size based on their earning capacity and age. Higher levels of maturity, as well as education, inclines towards practical family planning, especially on the part of the present generation. In this paper, we aim to explore the independent variables such as education, income, and age and their relationship with regards to the total children in a family. We are using a negative binomial distribution in order to create an appropriate regression model with our cleaned data, obtained from the GSS dataset. This model will aid the understanding of the current social trends involving family planning.

Since there were a lot of NA values in the dataset, we had to clean up the data set and filter out the variables we required for our analysis, as described in the model development section of the paper. We chose to use a generalized linear model of the negative binomial family for the regression analysis due to the characteristics and distribution of our data. After validating our model and running the analysis, we discovered the estimates, standard deviations, z-value, and p-value of each of our respective independent variables, namely age, and respondents' income and education levels. The results section of the paper displays these values we obtained. Our model suggests varying degrees of relationship between the total number of children in a family and the predictor variables. We acknowledge the limitations of this model in the weaknesses section such as the biasedness in the data due to a single collection method. Despite the age group and data bias limitations that need to be dealt with, the results of this analysis should provide a reference to the relationship between the total number of children in families and the respondents' age, income level, and education level.

In conclusion, we see that age is a very strong predictor of the total number of children in a family and we find trends related to this in the respondents' education and income levels. It appears that with an increase in income levels, there is an increase in the number of children a family has, whereas we see greater numbers

of children for respondents with lower education levels. In order to come to more conclusive answers on what factors lead to the highest number of children in a family, future work must be carried out by selecting more variables and alternative regression models to apply to the data. Still, our work will provide important information regarding specifically, the effects of our chosen predictors on total children.

Data Discussion

The data we used in this paper were collected by the Canadian General Social Survey Program back in 2017. The 2017 GSS is a sample survey with a cross-sectional design. The target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. This survey included more than 400 variables ranging from educational background to demographic characteristics.

Below we have displayed a preview of this GSS Data.

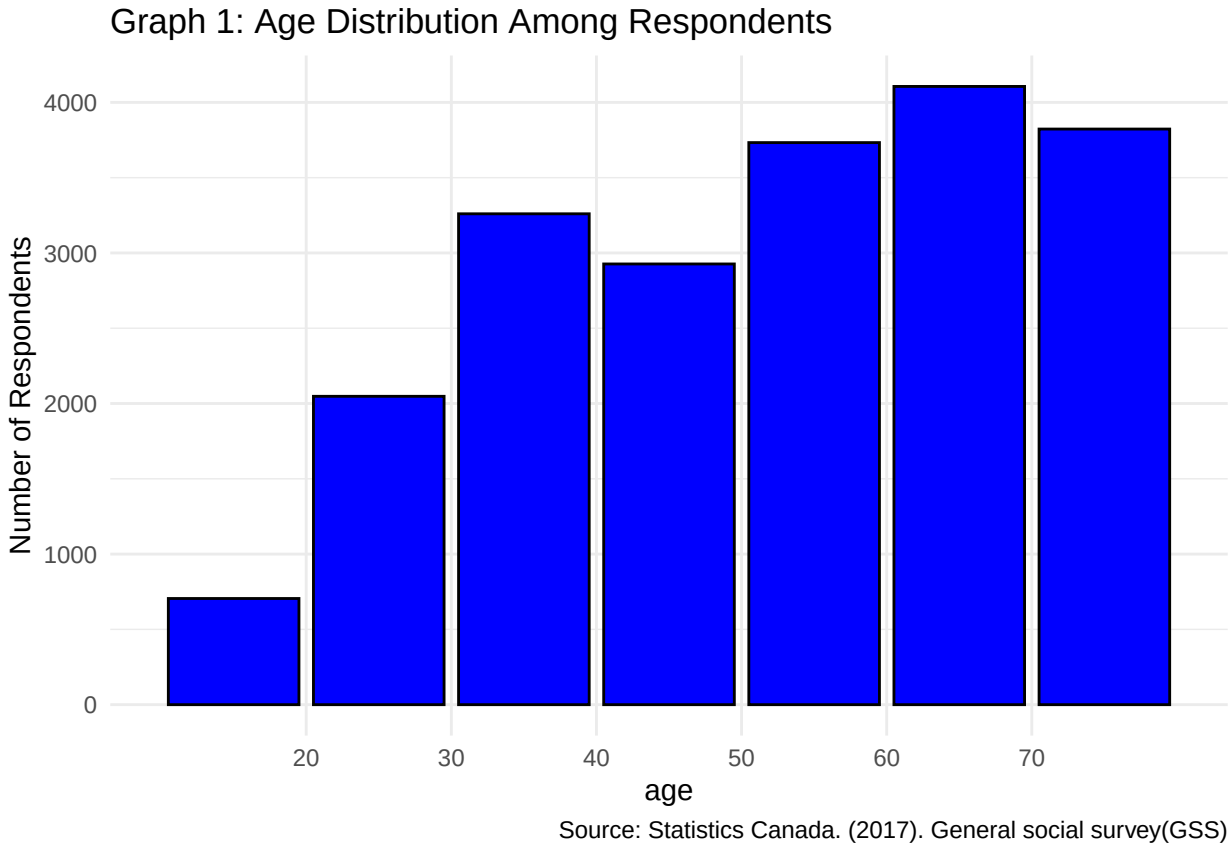
```
## # A tibble: 6 x 81
##   caseid   age age_first_child age_youngest_ch~ total_children age_start_relat~
##    <dbl> <dbl>           <dbl>            <dbl>          <dbl>            <dbl>
## 1      1  52.7              27               NA              1             NA
## 2      2  51.1              33               NA              5             NA
## 3      3  63.6              40               NA              5             NA
## 4      4  80                56               NA              1             NA
## 5      5  28                NA               NA              0             25.3
## 6      6  63                37               NA              2             NA
## # ... with 75 more variables: age_at_first_marriage <dbl>,
## #   age_at_first_birth <dbl>, distance_between_houses <dbl>,
## #   age_youngest_child_returned_work <dbl>, feelings_life <dbl>, sex <chr>,
## #   place_birth_canada <chr>, place_birth_father <chr>,
## #   place_birth_mother <chr>, place_birth_macro_region <chr>,
## #   place_birth_province <chr>, year_arrived_canada <chr>, province <chr>,
## #   region <chr>, pop_center <chr>, marital_status <chr>, aboriginal <chr>,
## #   vis_minority <chr>, age_immigration <chr>, landed_immigrant <chr>,
## #   citizenship_status <chr>, education <chr>, own_rent <chr>,
## #   living_arrangement <chr>, hh_type <chr>, hh_size <dbl>,
## #   partner_birth_country <chr>, partner_birth_province <chr>,
## #   partner_vis_minority <chr>, partner_sex <chr>, partner_education <chr>,
## #   average_hours_worked <chr>, worked_last_week <chr>,
## #   partner_main_activity <chr>, self_rated_health <chr>,
## #   self_rated_mental_health <chr>, religion_has_affiliation <chr>,
## #   regilion_importance <chr>, language_home <chr>, language_knowledge <chr>,
## #   income_family <chr>, income_respondent <chr>, occupation <chr>,
## #   childcare_regular <chr>, childcare_type <chr>,
## #   childcare_monthly_cost <chr>, ever_fathered_child <chr>,
## #   ever_given_birth <chr>, number_of_current_union <chr>,
## #   lives_with_partner <chr>, children_in_household <chr>,
## #   number_total_children_intention <dbl>, has_grandchildren <chr>,
## #   grandparents_still_living <chr>, ever_married <chr>,
## #   current_marriage_is_first <chr>, number_marriages <dbl>,
## #   religion_participation <chr>, partner_location_residence <chr>,
## #   full_part_time_work <chr>, time_off_work_birth <chr>,
## #   reason_no_time_off_birth <chr>, returned_same_job <chr>,
## #   satisfied_time_children <chr>, provide_or_receive_fin_supp <chr>,
## #   fin_supp_child_supp <dbl>, fin_supp_child_exp <dbl>, fin_supp_lump <dbl>,
## #   fin_supp_other <dbl>, fin_supp_agreement <chr>,
## #   future_children_intention <chr>, is_male <dbl>, main_activity <lgl>,
```

```
## #   age_diff <chr>, number_total_children_known <dbl>
```

To carry out sampling, each of the ten provinces was divided into strata (i.e., geographic areas). Many of the Census Metropolitan Areas (CMAs) were each considered separate strata. Compared with the previous survey, the advantages of the 2017 GSS survey are many survey specific socio-demographic questions were replaced by Statistics Canada's harmonized content questions (i.e., standardized modules for household survey variables, such as marital status, education, and labour force) and the 2017 GSS on Families uses the redesigned GSS frame, which integrates data from sources of telephone numbers (landline and cellular) available to Statistics Canada and the Address Register.

In addition, the non-response problem was well handled. The non-response adjustment was done in three stages. In the first stage, adjustments were made for complete non-response. This was done within each stratum. In the second stage, adjustments were made for non-response with auxiliary information from sources available to Statistics Canada. These households had some auxiliary information which was used to model propensity to respond. In the third stage, adjustments were made for partial non-response. These households had some auxiliary information which was used to model propensity to respond.

However, most of the variables needed to be understood in the survey dictionary, so with the help of Professor Rohan Alexander, we reduced the number of variables we were interested in down to 81. In this paper, we will focus on the three variables: age, income, and education and we want to determine the relationship between these three variables and total children in a family.

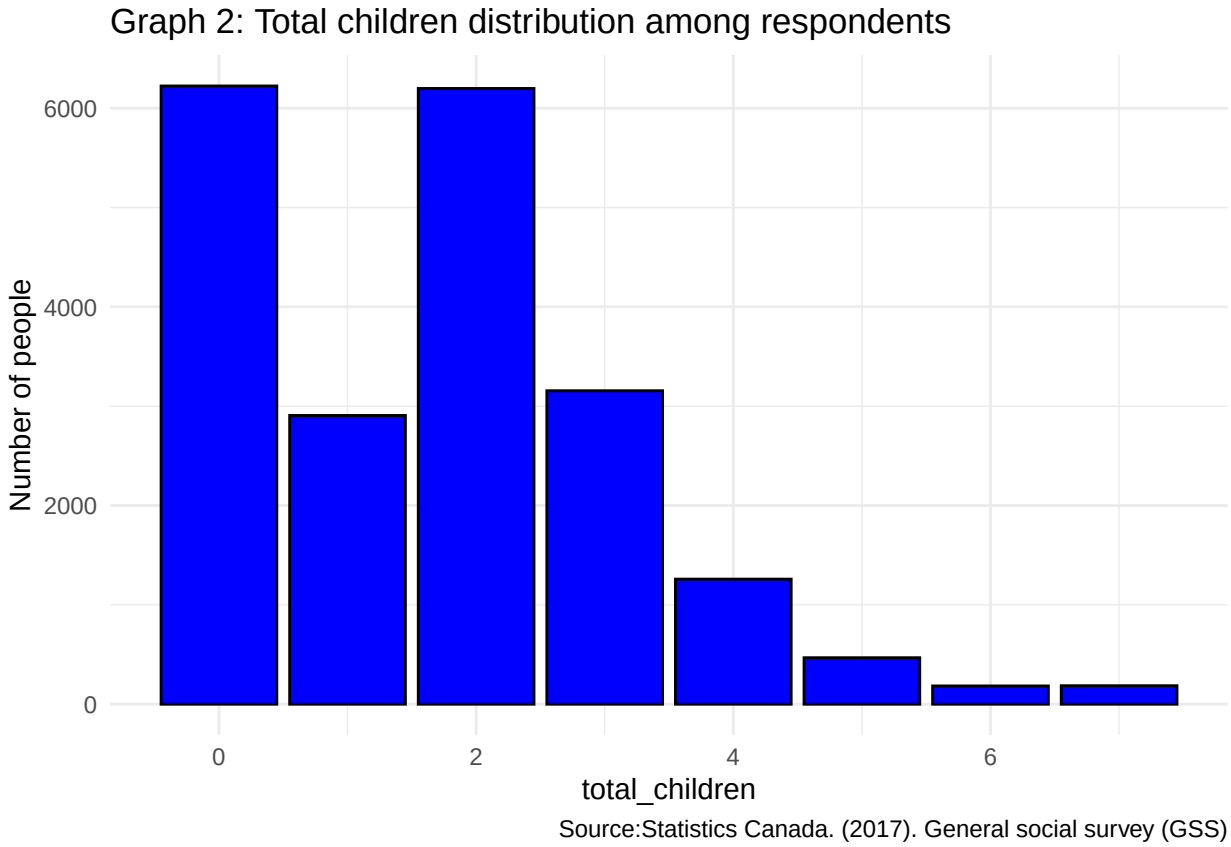


As shown in graph 1, the largest proportion of respondents were over the age of 50, with the highest number of respondents between the ages of 60 to 70. In contrast, we see the smallest number of respondents below the age of 20. Thus, the proportion of older respondents greatly outweighs the proportion of younger respondents to the survey.

Older respondents have often reached the final number of children they will have, whereas the younger proportion will likely parent more children in the future. This may cause some disruptions to our model's

strength as younger respondents with similar degrees and income levels to older respondents will likely have different numbers of children. As a result of this, we may see a slight underestimation of the strengths of these predictors.

```
## Warning: Removed 19 rows containing non-finite values (stat_count).
```



From graph 2, we can clearly see that most respondents have zero or two children with proportions totaling over 30%. As expected, the lowest proportion of respondents are seen with 4 or more children. The large number of respondents with zero children is surprising given the age proportions we saw earlier, however, this should not disturb our model's ability.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 6 x 3
##   income_respondent     number percentage
##   <chr>                  <int>      <dbl>
## 1 $100,000 to $ 124,999    846     0.0411
## 2 $125,000 and more        885     0.0430
## 3 $25,000 to $49,999      6173     0.300
## 4 $50,000 to $74,999      3896     0.189
## 5 $75,000 to $99,999      2030     0.0985
## 6 Less than $25,000       6772     0.329
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 8 x 3
##   education                                                  number percentage
##   <chr>                                                       <int>      <dbl>
## 1 Bachelor's degree (e.g. B.A., B.Sc., LL.B.)                  3753     0.182
## 2 College, CEGEP or other non-university certificate or di...  4566     0.222
```

```
## 3 High school diploma or a high school equivalency certificate   4848     0.235
## 4 Less than high school diploma or its equivalent                3036     0.147
## 5 Trade certificate or diploma                                   1483     0.0720
## 6 University certificate or diploma below the bachelor's level   732     0.0355
## 7 University certificate, diploma or degree above the bach...    1843     0.0895
## 8 <NA>                                                            341     0.0166
```

As shown in table 3 and table 4, nearly a third of respondents earn less than $25,000 a year, with the next highest proportion seen earning between $25,000 and $49,999, while only 4 percent earn more than $125,000 a year. As for educational background, more than 50% of respondents had a college diploma or above, and only about 15% had a lower educational level than high school. We may end up running into problems with our model's statistical significance for the categories with the lowest proportions of respondents as there will be less data for our model to get a clear understanding of. This will likely result in these income and education options being less significant than their alternatives in determining the expected number of children.

Model Development

Due to this being a real survey, we can find many NA values in the responses given that can prevent us from properly carrying out a regression analysis. To correct this we first selected only the variables of interest; the total number of children, age, education, and respondent income, from our original data, and made a new dataset with these. We then filtered the responses to these questions so that they would contain only complete cases and remove any NA responses. Doing this ensures that the length of our regression model's observations would match that of our cleaned dataset.

```
## # A tibble: 6 x 5
##   caseid total_children   age income_respondent education
##    <dbl>          <dbl> <dbl> <chr>             <chr>
## 1      1              1  52.7 $25,000 to $49,9~ High school diploma or a high s~
## 2      2              5  51.1 Less than $25,000 Trade certificate or diploma
## 3      3              5  63.6 $25,000 to $49,9~ Bachelor's degree (e.g. B.A., B~
## 4      4              1  80   $50,000 to $74,9~ High school diploma or a high s~
## 5      5              0  28   Less than $25,000 College, CEGEP or other non-uni~
## 6      6              2  63   Less than $25,000 High school diploma or a high s~
```

```
## [1] 1.676151
```

```
## [1] 2.212757
```

The regression model we have chosen to use in this study is a generalized linear regression model of the negative binomial family. A generalized linear model is a generalization of the ordinary linear regression which removes the assumptions made in the linear model. This model then allows us to link our response variable through a specified probability distribution, we refer to this distribution as the family of our regression model. Thus, unlike in linear regression, we do not need to have a normally distributed response variable.

We first made this choice as the independent variable we are studying, total children, is count-based and not continuous which meant that ordinary least squares regression would not work as well. Many of our variables also violate the strong requirements for a linear regression model like homoscedasticity and normality of errors. Since generalized linear regression models do not feature these same requirements, this made it the ideal model option for our data. When it comes to determining a distribution family, the skewness of our data suggests that a Poisson distribution might be suitable since it accounts for this. However, if we analyze the mean and variance of total children: Mean = 1.68 Variance = 2.21, we notice that the variance is greater, meaning that our data are overdispersed. This violates the Poisson model's equal dispersion assumption and will negatively affect the effectiveness of our model. As this equal dispersion assumption is often unrealistic for real work data, we must use a probability distribution similar to Poisson in shape, but without the harsh criterion. Thus, to account for this, we have decided to use a negative binomial probability distribution in our model. A negative binomial distribution shares almost the exact same shape as a Poisson distribution

but it does not make the same equal dispersion assumption. As it works well specifically with over-dispersed and skewed count data, which is exactly what we have, we have decided to use it throughout our study.

Alternatively, at this step, we could have chosen a zero-inflated Poisson regression model as we notice the number of respondents with zero total children is extremely large, at over 30% of our dataset. We decided against this, however, as we also see an almost equally large number of respondents with a total of two children as well as the strong overdispersion in our data which the zero-inflated model could not account for. As mentioned earlier, the explanatory variables we have chosen for our study are the respondent's; age, education, and income bracket, as they are all extremely influential in determining one's social/economic status which is known to play a key role in the number of children in a family. Since age has already been recorded as a numerical value, we did not have to alter it to fit it into our model. However, for both education and respondent income, they are set up as categorical strings in our data, thus to allow them to function properly in our generalized linear model we first needed to apply the as.factor() function from the base R package to convert them into a factor form that we can work with.

Model Validation

```
## Joining, by = c("caseid", "total_children", "age", "income_respondent", "education")
```

```
## [1] 16244
```

In order to validate our model, as seen in the appendix, we initially divided the GSS dataset into a model validation dataset and a model development dataset. We decided that the model validation dataset should be roughly 20% of the cleaned GSS dataset at 4000 observations, leaving us with a total of 16244 respondents in our training set. Through running calculations in R, we found the MSPE to be 1.22, when compared to the MSres from our model, at 1.23, we find that the difference is smaller than 0.1. An MSPE value that is close to the MSRes value based on the regression fit to the model-building data set, indicates a high predictive ability of our model. Thus, this incredibly small difference helps to prove the validity of our model selection being a negative binomial regression.

Next, using a 10-fold cross validation analysis, we find that our cross validation estimate of prediction error is equal to 1.8. This means that, if we split our data up into 10 subsamples, cross validate using 1 of these subsamples as the testing dataset, and then perform this same process for each subsample, we would see an average prediction error of 1.8. As a result of this relatively high value, our model's predictive ability will not be extremely strong as the error will likely cause us to under or overestimate the number of children we expect. This is likely due to the number of predictor variables we have chosen to include. These values suggest that there is still a lot of variance in the total number of children a family has, given their responses to these questions. By using more predictor variables we would be able to gain a more specific understanding of exactly what personal factors will result in the highest number of children in a family. Options we have considered include; the age of the respondent at the birth of their first child, whether or not the respondent has a religious affiliation, and whether or not the respondent is currently married. The addition of these variables would require different regression models depending on dispersion in the data and the distribution of the response variable. While further research must be done to reach better predictive abilities, the estimates and data we have collected will still be able to tell us the relative significance of the response options and how they affect the total number of children we expect to see. This will also allow us to see trends in the total number of children a respondent has based on what level of education they have achieved and what level of income they earn.

Results

```
## MODEL INFO:
## Observations: 16244
## Dependent Variable: total_children
## Type: Generalized linear model
```

```
##   Family: Negative Binomial(23.2708)
##   Link function: log
##
## Standard errors: MLE
## ------------------------------------------------------------------------------------
##                                                Est.      S.E.      z val.          p
## ------------------------------------------ ---------- --------- ----------- ---------
## (Intercept)                                -0.78486   0.03966   -19.79065   0.00000
## age                                         0.02384   0.00040    59.69975   0.00000
## as.factor(education)College,                0.12286   0.02091     5.87491   0.00000
## CEGEP or other non-university
## certificate or di...
## as.factor(education)High                    0.12309   0.02090     5.89078   0.00000
## school diploma or a high
## school equivalency
## certificate
## as.factor(education)Less                    0.16732   0.02323     7.20129   0.00000
## than high school diploma or
## its equivalent
## as.factor(education)Trade                   0.18530   0.02745     6.75104   0.00000
## certificate or diploma
## as.factor(education)University              0.04356   0.03651     1.19315   0.23281
## certificate or diploma below
## the bachelor's level
## as.factor(education)University             -0.05658   0.02739    -2.06579   0.03885
## certificate, diploma or degree
## above the bach...
## as.factor(income_respondent)$125,000        0.02490   0.04248     0.58622   0.55773
## and more
## as.factor(income_respondent)$25,000        -0.15414   0.03368    -4.57686   0.00000
## to $49,999
## as.factor(income_respondent)$50,000        -0.08610   0.03443    -2.50064   0.01240
## to $74,999
## as.factor(income_respondent)$75,000        -0.00408   0.03697    -0.11025   0.91221
## to $99,999
## as.factor(income_respondent)Less           -0.19229   0.03398    -5.65850   0.00000
## than $25,000
## ------------------------------------------------------------------------------------
```

In this section, we will display and talk about the core results of our model. To begin with, in Table 1, we display the output from the estimate of our model using summary statistics via the MASS(reference call) package. To interpret both the impact and the strength of our predictor variables, we will pay attention to three key columns of our regression summary; Estimate, Z-value, and P-value. The coefficient estimate value will tell us the log change in the total number of children that we expect a respondent to have given change in the predictor's value. Next, the Z-values will tell us whether or not we can reject the null hypothesis, that our coefficient estimate is truly zero. Here, to conform with a 95% confidence interval, we will be looking for absolute values greater than 1.96. Lastly, a predictor's p-value works with the z value to confirm our rejection of the null hypothesis. Keeping in line with the same 95% confidence interval we require values less than 0.05. Independent variables that satisfy both of the Z and p-value conditions we have described above will, as a result, reject the null hypothesis and prove to be significant predictors of the total number of children people have. We will discuss which specific predictors these are and their overall significance in the discussion section below.

From this output, we can determine the formula for our final model: log(Y) = -0.78486 + 0.02384X + 0.12286S1 + 0.12309S2 + 0.16732S3 + 0.18530S4 + 0.04356S5 - 0.05658S6 + 0.02490T1 - 0.15414T2 -

0.08610T3 - 0.00408T4 - 0.19229T5 Where: Y represents our response variable, total children, X represents the predictor variable age Si is an indicator variable representing the education level of the respondent, in our equation a the corresponding variable is set to 1 if the respondent chooses the category and zero otherwise. S1 corresponds to a College, CEGEP, or other non-university diploma, S2 to a High School Diploma or High School equivalency, S3 to a level below the high school equivalent, S4 to a Trade certificate or diploma, S5 to a University certificate or diploma below the bachelor's level, and S6 signifies a University certificate, diploma, or degree above the bachelor's level. If all of these options are zero, we take this to mean that the respondent is part of the group we condition on, those who have earned a Bachelor's degree. Ti represents the income level of the respondent and it operates in nearly the exact same way as Si. T1 means an income level above $125,000, T2 is the income bracket $25,000-$49,999, T3 is the bracket $50,000-$74,999, T4 is the bracket $75,000-$99,999, and T5 represents incomes below $25,000. As with Si, when all options equal zero this means that the respondent is part of the income bracket we condition on, $100,000-$124,999.

```
# Plot Coefficient Estimates and Standard Deviations
plot_summs(negbinmodel, scale= FALSE)
```

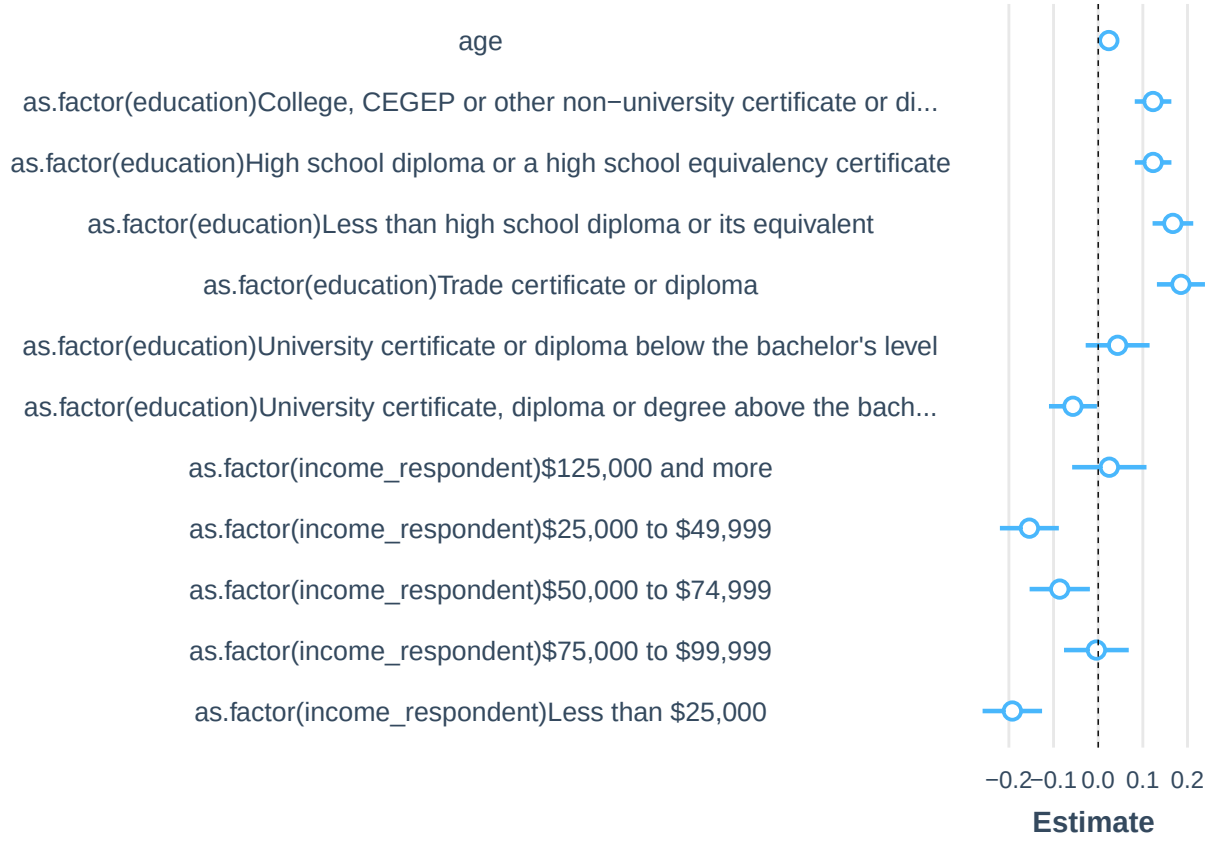


Figure 1: Figure 6

Next, In Figure 6, we can see the estimated coefficient values and their respective standard deviations. As described above, estimates are concerning the logarithmic change we expect to see in total children given a change in the predictor variables. The standard deviation tracks the variability between these estimates, and thus shows how precise they are. For age, we observe that the standard deviation is extremely small and thus the estimate of the coefficient is likely very precise. Next for the level of education completed, we can see for the majority of categories, the standard deviation was relatively low, leading us to believe that the estimated coefficients for these categories are accurate. However, categories; "University certificate or diploma below the bachelor's level", "University certificate or diploma above the bachelor's level" and "Trade certificate or diploma" had higher standard deviations than the other levels of education, thus we cannot

trust their estimated coefficients as surely. Finally, for levels of annual income, we see that the standard deviation is relatively high in comparison to education. The average standard deviation for income categories was about 0.04 whereas education was only about 0.03, thus the estimates of the coefficients for variables in this category can be seen as less precise.

Discussion

Based on the results obtained from our model, the impact and strength of the predictor variables can be determined by the three values of each predictor, namely, the estimate, the Z-value, and the P-value. The estimates of the coefficient values are essentially the slopes of the regression equation, informing us of a logarithmic change in the total number of children that respondents intend to have, given a unit change in the predictor's value. Considering a null hypothesis of zero, that is, there is no relationship between the predictors, age, education, and income and the response variable, we also look at the Z-values and P-values obtained for each predictor. The Z-value indicates the degree of uncertainty that surrounds the estimate of the co-efficient whereas the P-value also works with the Z-value and essentially allows us to verify the truth of the null hypothesis. In our analysis, we conform to a 95% confidence level thereby, desiring Z-values more than 1.96 and P-values less than 0.05 in order to reject the null hypothesis.

Age:

The estimated coefficient for the predictor "age" is 0.02. It is a positive value indicating a change of 0.02 in the logarithmic value of the total children in a family caused by a unit change in the age of the respondent. A huge z-value of 67.5 and an extremely low p-value of less than 2eˆ-16, allows us to reject the null hypothesis. It can be safely assumed that with an increase in the age of the respondent, the total number of children increases. Having children is an extremely long process but even considering parenthood at a young age, families can still choose to have more children with maturity and age. Hence, the reality, very practically, conforms to the results derived from this model.

Education:

Using the education level of a Bachelor's degree as conditional, we can think about the other values as the difference in the number of total children we would expect to see from a respondent with a different level of education. Analysing the results, it is very apparent that all the levels of education, when compared to the Bachelor's degree, have positive estimates indicating a higher number of children in the family. However, when respondents achieve a degree above the Bachelor's, the estimated value is -0.06. Thus, the general trend as observed from an educational degree at the Bachelor level or above is that a higher level of education likely provides enough knowledge, maturity, and understanding for the respondent to undertake proper family planning to have the correct amount of children according to their capacity. Obtaining a trade school certificate or less than a high school diploma resulted in the highest estimates at 0.19 and 0.17 respectively. Hence, it seems that a lower level of education deprives the respondents of accurate information or knowledge regarding family planning, and in most cases, leads to the highest number of children in these families compared to the families of respondents with a higher degree.

The z-values and the p-values for all education levels except a University certificate or diploma below the bachelor's level, allow us to reject the null hypothesis, thus confirming that their estimates are significant. However, a University certificate or diploma below the bachelor's level has a very high p-value of 0.23 ($>$ 0.05), disabling us from rejecting the null hypothesis. As mentioned earlier, this was to be expected since the proportion of respondents that chose this option was relatively low.

Income:

Conditioning on the $100,000-$124,999 income bracket, we can interpret the estimates from the different income categories as differences in the total number of children expected in the family of a respondent with an income between $100,000 and $124,999 per year. We can observe that for income levels below $75,000, the estimates are negative, at -0.09, and keeps decreasing as income decreases, with the smallest estimate at -0.19 for income below $25,000. This realistically indicates that the lower the income, less would be the number of children in a respondent's family as they would likely not have sufficient funding to maintain or provide

for these children. Moreover, the highest estimate of 0.02 for income above $125,000 suggests that a higher income allows the respondent to comfortably provide for more children. Hence, we can see a trend in the level of income for respondents where higher levels lead to more children in a family.

Examination of the z-values and p-values for the income levels reveal that for income less than $25,000 and the brackets of $25,000-$49,999 and $50,000-$74,999, the p-values are low enough to reject the null hypothesis and confirm their significance to our model. However, for income level above $75,000, the p-values are huge and thus, we do not consider these specific income categories significant. As was the same case with the less significant education categories, the low proportion of respondents we saw in these income brackets is the likely explanation for their insignificance in our model.

To conclude, our study discovered that there was a positive relationship between total children and a respondent's age and income levels, and a negative relationship to the respondent's education levels. These results are not too surprising when we think about the real world effects of living in different socioeconomic conditions. However, our regression model works to strengthen the understanding of this relationship and with future work, can be adapted to provide conclusive answers on what factors lead to the most children in a family.

Weaknesses

The most prominent limitation of our study is the fact that we are dealing with a survey where we must rely on total honesty from respondents. It would be extremely easy for respondents to falsify their answers, thereby making our conclusions less dependable. Since there is no way we can confirm the truthfulness of answers, we must assume that we have been provided with completely honest responses. As mentioned earlier, we can also expect some bias in our model given the distribution of the ages of respondents. To protect against this, we could have chosen a method to take a sample of our dataset that would provide a better representation of all age groups equally, however, this would have greatly reduced our model's predictive ability as we would lose a significant number of observations. The biasedness of data collection is mainly caused by a single data collection method as interviews are held over the phone. The biggest disadvantage of this method is that it ignores many people who either do not have mobile phones or the time to take part in a lengthy study. As the data shows, there is a very low proportion of younger respondents likely due to the fact that most often, younger people have less spare time in their days to complete a survey compared to the older population. Thus we would be less likely to see a large proportion from younger age groups in our data which can negatively impact our model's success. Another limitation of our model comes from the number of variables we have chosen for our study. By selecting so few questions out of a very long survey, we still end up seeing a large variability in the total number of children a respondent has given their responses. If we were to add additional questions into our model we would be able to gain a greater predictive ability and a higher significance of our estimates. Further work can be carried out wherein more predictor variables are selected for our model. This would likely result in a much stronger predictive ability, thereby allowing us to gain an even better understanding of the determinants of the number of children in a family.

Appendix

Code for this study can be found at:

References

Ben Bolker and David Robinson (2020). broom.mixed: Tidying Methods for Mixed Models. R package version 0.2.6. http://github.com/bbolker/broom.mixed

Government of Canada, S. C. (2017, February 27). The General Social Survey: An Overview. Government of Canada, Statistics Canada. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org

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.

Lionel Henry, Hadley Wickham and Winston Chang (2020). ggstance: Horizontal 'ggplot2' Components. R package version 0.3.4.

Long JA (2020). *jtools: Analysis and Presentation of Social Scientific Data.* R package version 2.1.0, <URL: https://cran.r-project.org/package=jtools>. Negative Binomial Regression | STATA Annotated Output:UCLA, Institute for Digital Research and Education, Statistical Consulting. https://stats.idre.ucla.edu/stata/output/negative-binomial-regression/

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rohan Alexander (2020).GSS_cleaning

Sachin. (2020, May 10). Generalized Linear Models. https://towardsdatascience.com/generalized-linear-models-9ec4dfe3dc3f

Sachin. (2020, May 8). An Illustrated Guide to the Poisson Regression Model. https://towardsdatascience.com/an-illustrated-guide-to-the-poisson-regression-model-50cccba15958

Sachin. (2020, September 13). Negative Binomial Regression: A Step by Step Guide. Medium. https://towardsdatascience.com/negative-binomial-regression-f99031bb25b4

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

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

- `Wickham et al., (2019). Welcome to the tidyverse. Journal of Open`

Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686