Exploration of the Relationship among Age, Family Members and Income

Yue Chen Shen 1005350790, ANNI LIN 1004141751, Andong Cai 1005035908, Yizhou Hu 1005342808

October 19, 2020

Abstract

In this report, the data based on the general social survey on family in 2017. The Canadian general social surveys are provided by Statistics Canada and all the surveys are being conducted through telephones. The content of the data is about the characteristics and conditions of Canadian families. The report will be using classification variables such as age, number of children and income to create analysis and find relationships between these.

Introduction

The goal of this analysis is to explore the relationship between income versus number of children and age versus number of children. Throughout this analysis, we are interested in finding out if it is the case that higher income families have more children. Also about how one's age is associated with the number of children and see if the number of children increases as one becomes older. In order to demonstrate and explore a type of relationship between the income and the number of children, a histogram or bar graph will be constructed to model the relationship using easy to read bars, the horizontal axis will be splitted into several income levels while the vertical axis will be number of children, hence creating a straight forward view of the volume and total numbers. In terms of different ages and the corresponding number of children, a scatterplot will be applied to see the distribution and how each age corresponds to the number of children. Logistic regression models will also be applied to predict the growth of age and the change in number of children. The above variables are definitely closely related with each other, by utilizing the following data and model, we will be able to obtain a more detailed and specific relationship and trend between the variables.

Data

I downloded raw data, dictionary and labels from the website CHASS http://www.chass.utoronto.ca/.

Then I used the given code to clean the data and named the data gss.

For Figure 1 and Table 1.We want to analyse the relationship between age and total_children. Therefore, We used plot function to make a scatter plot and draw a linear line through it by abline function. We wonder whether people have more children when they are older or it is not related at all. Age is a numeric variable and total_children is a categorical variable. Age and total_children both have 20602 observations in it. Variable age represents a person's age. Variable total_children represents total number of children that person has. We clearly see the linear relationship of those two variables, but since total_children is a categorical variable, it cannot be performed well by linear regression model. In order to check the specific statistics of our plot, we use summary function to get the details of our linear regression model.

For Figure 2 and Table 2. We also chose age_at_first_birth and total_children to see if the people who have the first child early tend to have more children. As the same as above, we use the same method to get a scatter plot and a linear line. We also use summary to get information from the plot. Variable total_children

has been introduced before, so I am going to introduce age_at_first_birth. Variable age_at_first_birth is a numerical variable, representing the person's age when he/she has the first child. Vector age_at_first_birth has 20602 observations with some missing values in it, which makes sense as people may not have child yet. We struggled choosing between age_at_first_birth and age_first_child, since they are similar and related. In the end, we decided to use age_at_first_birth as we thought people may have a child with older age because they are old. We cannot determine if people have the first child early by the variable age_first_child.

For Figure 3, the variable income_family is a categorical variable which contains different levels of income in a family. In order to know whether family who has a higher income tends to have more children or vice versa. I extracted income_family, total_children from gss data to form a new data named gss_income_children_1. What is more, I eliminate the missing data in total_children for getting a cleaner data. Then I combined income_family and total_children to create a new variable named income_children in the gss_income_children_1 data. The variable income_children got grouped by its income level and total number of children. Variable income_children is a categoricla variable and I count the number of observations for each group then make a histogram. Since income_children got grouped by 0 to 7, 8 separated total_children and 6 different income levels, the variable has 48 observations. I can know people's tendency about how many children they want in each income range by this variable. The disadvantage of it is that it makes x-axis of the histogram too messy. I was considering separating the histogram into 6 different histograms by income levels. However, I think it is clearer to see the differences among each level if I do not separate the histogram. Besides, the age of people is a potential influential variable, for example, we can have two persons who are 20 years old and 70 years old. Both of them may be in the same income range. But 70 years old person has more chances to have more children.

In Figure 4, I make a new data called gss_income_children_2, which is almost the same as gss_income_children_1. Except it was classified by total_children first, then by income_family. And I named the new combined variable as children_income. I made a histogram by variabl children_income as a predictor variable and counts of it as a response variable. By this variable, I am able to say that which income range of family tends to have 0 or 1 or ... or 7 children.

In Figure 5, I selected variable average_hours_worked and income_family. The variable average_hours_worked is a categorical variable which contains different worked hours. We wonder if a family's income is positively related to working hours. Therefore, I group_by variables income_family and average_hours_worked to form a new data called gss_hours_income_1. Then, I combined those two variables to form a new variable income_hours, which is sorted by income_family firstly and average_hours_worked secondly. I counted the number of observations for each group. Afterwards, the variable income_hours has 24 observations. After I ploted the histogram, I can clearly see which hours worked is the most common in each income range.

For Figure 6, I selected the same variable as above. And the new data gss_hours_income_2 is almost the same as gss_hours_income_1. The difference is that I make the variable hours_income sorted by worked hours first, then the income level. For this data, we can clearly see what is the usual income of people for certain working hours.

In the Table 3, I chosse has_grandchildren, age, total_children and age_first_child variables from gss data to make a new dataset called gss_grandchildren. I change "Yes" to 1 and "No" to 0 for the variable has_grandchildren to apply for logistic regression model. The variable has_grandchildren is a categorical variable with value 1 or 0. Variable age_first_child is a continuous variable represents the age of the first child. We want to use age, total_children and age_first_child as predictor variables to predict the probability of has_grandchildren. People tends to have grandchildren after 50 years old. However, we have many people who are under 50, which may lead to misinterpretation.

Model

In order to find relationships between age, number of children, and income levels, our group decided to use three different models using r code.

Firstly, we use histograms to find the relationship between income and the number of children, and the relationship between income and number of worked hours. Figure 1 is a density histogram that shows the number of children, ranging from 1 to 7, of six different income brackets, while figure 2 is a density histogram that shows the level of incomes, given a specific number of children. Figure 3 and 4 are also density histograms that reveal relationships between worked hours and income. Our group chose to use histogram to analyze the data because histograms are an approximate representation of the distribution of our data. This can help our group to identify whether the data is normally distributed, left-skewed, or right-skewed.

Our group believes that there is a positive relationship between age and total children, and there is a negative relationship between total children and age at first birth. Therefore, we decide to use a linear regression model to determine the relationship between variables of interest. The explanatory variable in the first linear regression model is age, and the response variable is total children. In the second linear regression model, the explanatory variable is age at first birth, and the response variable is total children. Since $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is the model we would like to estimate, we expect β_1 to be a positive number for the linear regression model of age and total children, and we expect β_1 to be a negative number for the linear regression model of total children and age at first birth. In the case of age and total children, y would be the total number of children, x value is the age of individual, β_0 is the estimated intercept parameter, and β_1 is the estimated slope of regression line. In the model of age at first birth and total children, y would be the total number of children, x would be the at at first birth, β_0 is the estimated intercept parameter, and β_1 is the estimated slope of regression line.

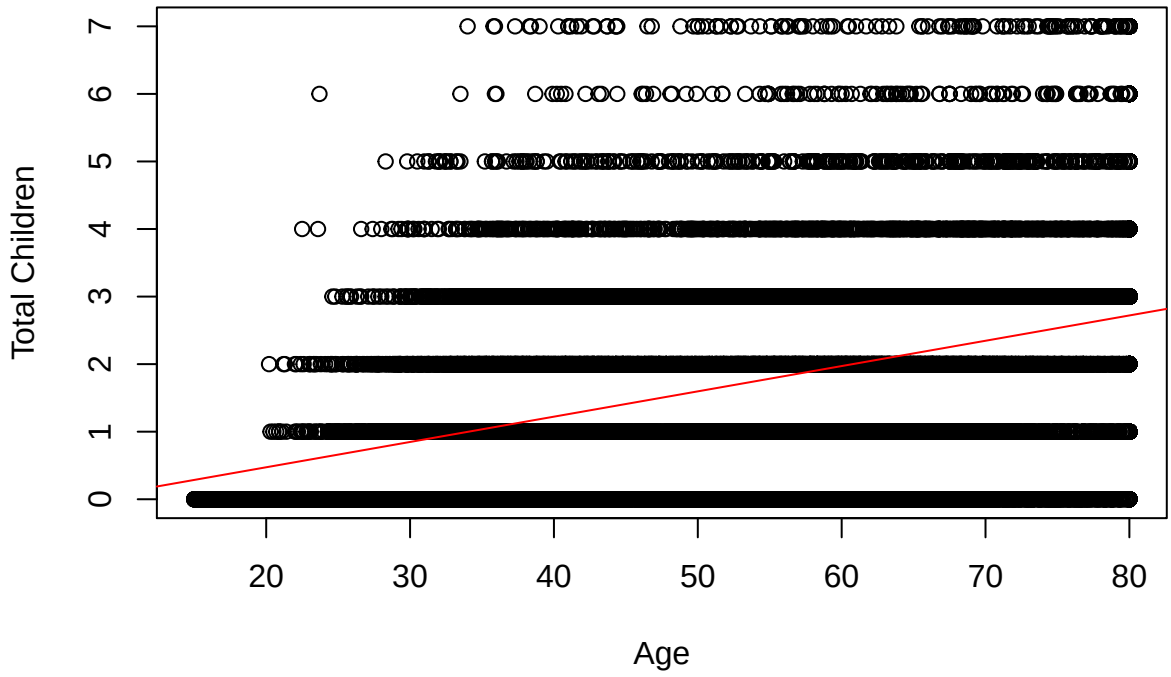
Lastly, we use a logistic model to determine the probability of a person having grandchildren. We are interested in how age, total children, adn age of first children affect the probability whether a person has grandchildren. In order to find the relationship, we are going to run a logistic regression analysis on these variables, and we are expecting the following regression equation: $log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{totalchildren} + \beta_3 x_{agefirstchildren}$. In this model, p is the probability of an individual having grandchildren, $log(\frac{p}{1-p})$ is the log probability of an individual having grandchildren. β_0 is the intercept parameter of the logistic model. β_1 is the expected change in log odds of having a grandchildren for every additional unit increase in age, β_2 is the expected change in log odds of having a grandchildren for every additional unit increase intotal children, and β_3 is the expected change in log odds of having a grandchildren for every additional unit increase in age of first children. One problem we faced during constructing the model is that our data for whether a person has grandchildren is recorded with "yes" or "no", instead of 0 or 1. However, we did a transformation of our, and replaced "yes" or "no" with 1 or 0.

Results

Here you will include all results. This includes descriptive statistics, graphs, figures, tables, and model results. Please ensure that everything is well formatted and in a report style. You must also provide an explanation of the results in this section. You can overflow to an Appendix if needed.

Please ensure that everything is well labelled. So if you have multiple histograms and plots, calling them Figure 1, 2, 3, etc. and referencing them as Figure 1, Figure 2, etc. in your report will be expected. The reader should not get lost in a sea of information. Make sure to have the results be clean, well formatted and di-

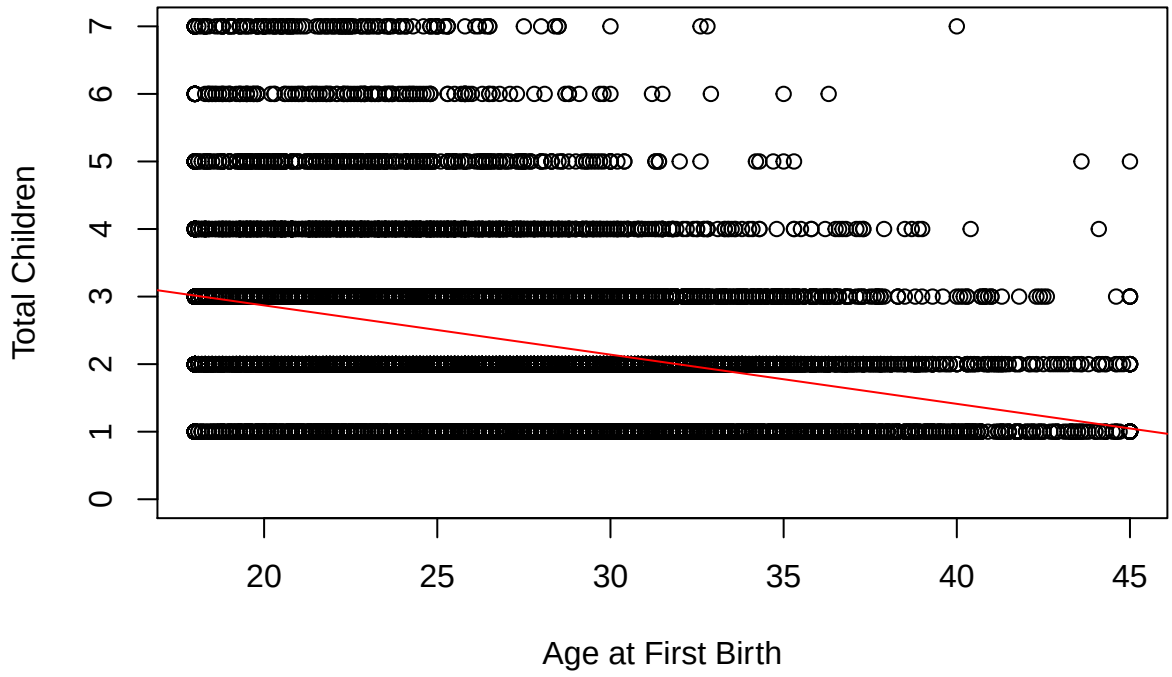
Figure 1. Scatterplot of Age and Total Children



gestible.

```
##
## Call:
## lm(formula = total_children ~ age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7205 -0.8271 -0.1662  0.7814  6.0025
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.2760540  0.0288599  -9.565   <2e-16 ***
## age          0.0374574  0.0005235  71.547   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.333 on 20581 degrees of freedom
##   (19 observations deleted due to missingness)
## Multiple R-squared:  0.1992, Adjusted R-squared:  0.1991
## F-statistic:  5119 on 1 and 20581 DF,  p-value: < 2.2e-16
```

Figure 2. Scatterplot of Total Children and Age at First Birth




```
##
## Call:
## lm(formula = total_children ~ age_at_first_birth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0151 -0.7018 -0.1844  0.4950  5.5880
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.326775   0.048825   88.62   <2e-16 ***
## age_at_first_birth -0.072870   0.001782  -40.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.089 on 12734 degrees of freedom
##   (7866 observations deleted due to missingness)
## Multiple R-squared:  0.1161, Adjusted R-squared:  0.116
## F-statistic:  1673 on 1 and 12734 DF,  p-value: < 2.2e-16
```

Figure 3. Income with Number of Children

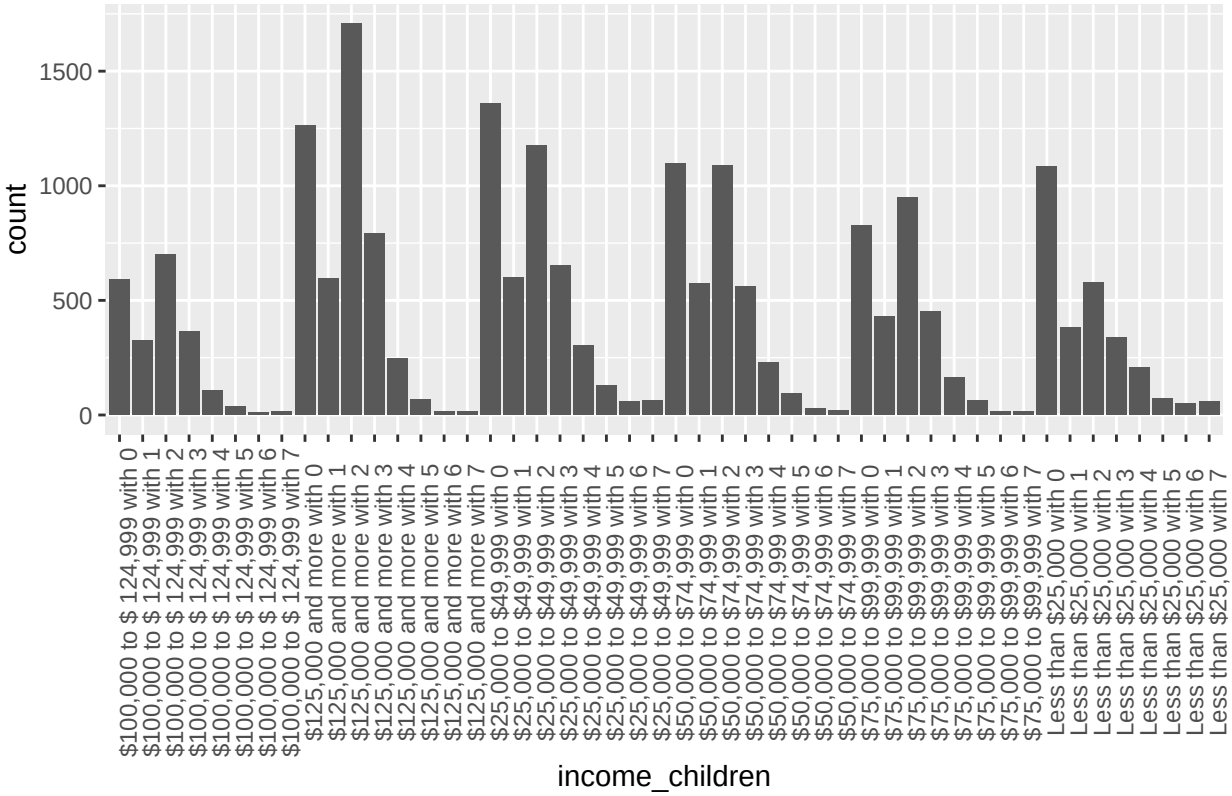


Figure 4. Number of Children with Income

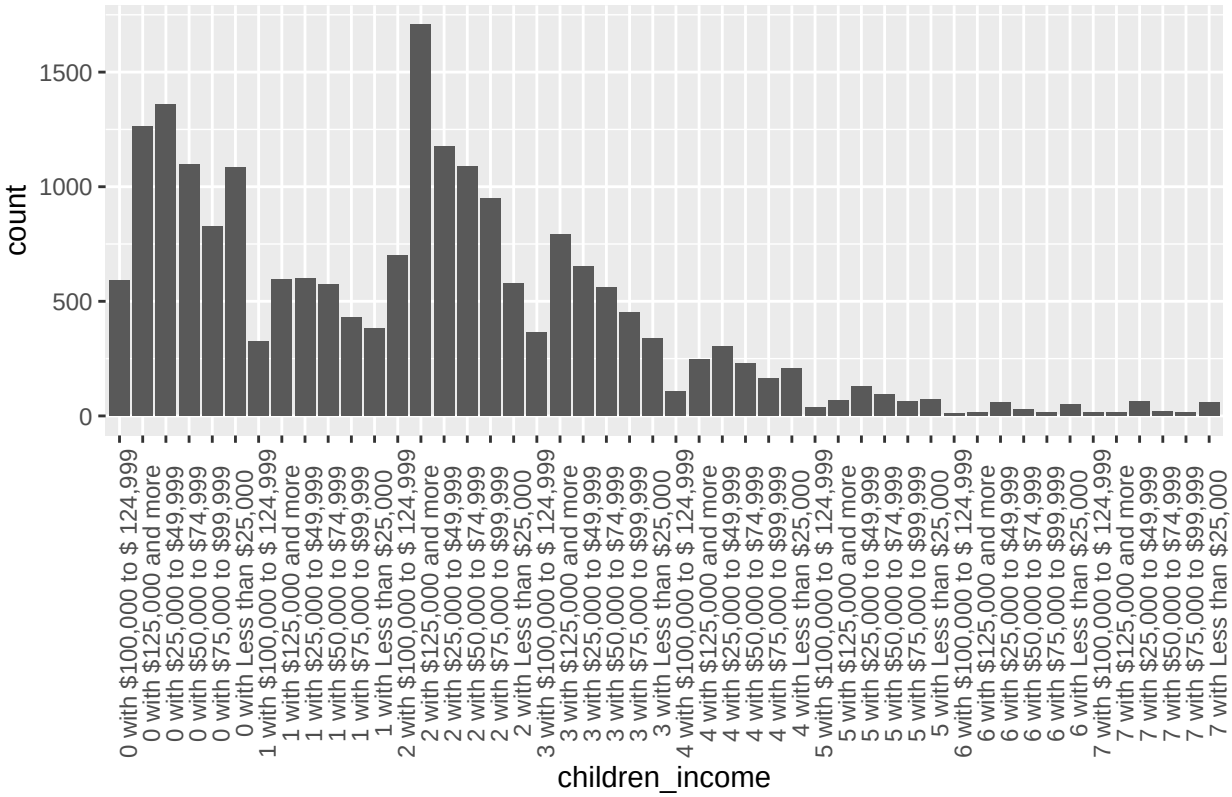


Figure 5.Income with Worked Hours

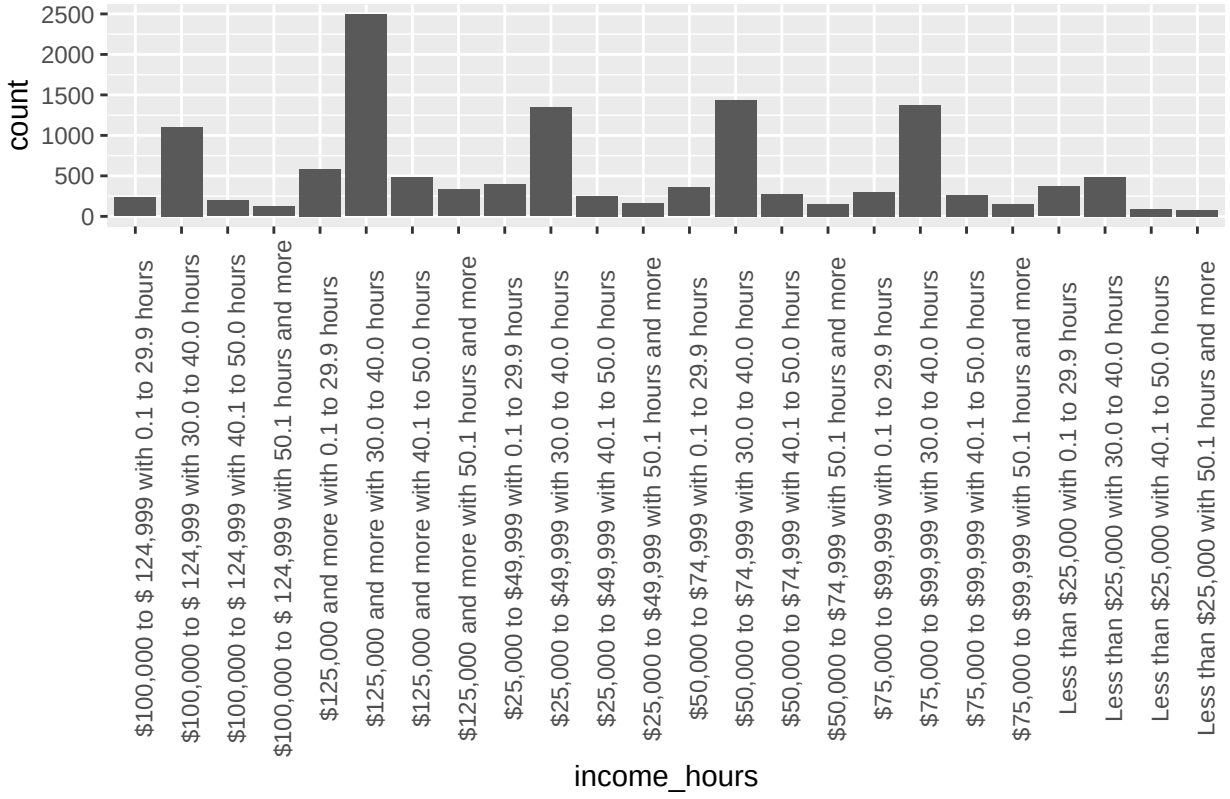
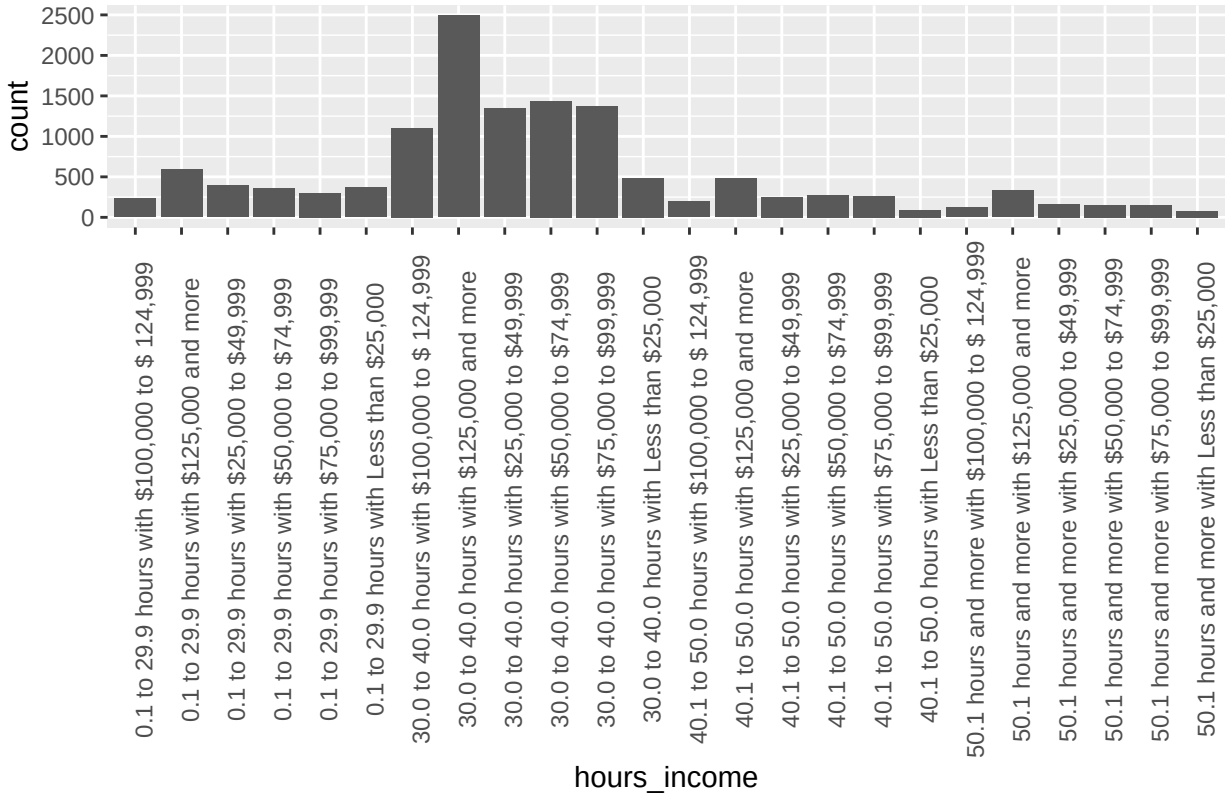


Figure 6. Worked Hours with Income



```
##
## Call:
## svyglm(formula = has_grandchildren ~ age + total_children + age_first_child,
##     design = grandchildren.design, family = "binomial")
##
## Survey design:
## svydesign(id = ~1, data = gss_grandchildren, fpc = fpc.srs)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -6.461177   0.134539 -48.025  < 2e-16 ***
## age             -0.021943   0.003383  -6.486 9.09e-11 ***
## total_children   0.359246   0.017269  20.802  < 2e-16 ***
## age_first_child  0.212570   0.003895  54.579  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1.598742)
##
## Number of Fisher Scoring iterations: 6
```

Discussion

The scatter plot from figure 1 discusses the relationship between the age of the mother and the total number of children that she has. From 0 to 4 children, we can see 5 very solid lines, that is because there are too many data points fitted so they scatter into a solid line, which again means that the number of individuals that have 5 and more children are a lot less in population when compared to the rest. Based on the plot, a linear slope can be fitted and is upward sloping meaning that older people will have more children, as a person grows older and enters a different stage of life, they will gradually have more children. Following the linear regression model from table 1, on average, when a person's age increases by one, the number of children that person has will increase by 0.037.

The other scatter plot from figure 2 is describing the age of first child and the total number of children, agains, we still see that the bottom four lines are more solid than the top three lines, this is because most families will have a max capacity of having 4 children, which causes the bottom four lines to be more dense than the other three. From this plot, we can conclude that most people in Canada give birth at a young age, which is mostly between 20 and 30 years old. If a person gives birth at relatively old age, such as after 40 years old, there exists a high likelihood that the total number of children will not be more than 2. As people age, their physical condition will not be able to give them as many births as younger individuals, which is why we will see a downward sloping trend between these two variables.

The associated table 2 also supports the above conclusion, one year increase in the age of giving first birth, the total number of children will decrease by 0.07.

From figure 3, it is evident that higher income is not directly proportional to more children, the lowest income family tends to have the most count of 5 children and more, whereas high income to middle income families usually have either no children or two children. Most Candian families are in the range of 125000 annual income and more, since that category has the highest count and highest volume of the bars. No matter the income, we can observe that 0 children or 2 children always have the highest bar, indicating that Canadian families tend to favor having 2 children or not having children at all.

From figure 4, the horizontal axis is categorized by the number of children, whereas figure 1 is categorized by different income. The overall shape of the figure is like a wave, there are two peaks at 0 children and 2 children, which supports the previous conclusion from figure 1, that most Canadian families like to have 2 children or no children. Near the end of the wave, there are very few counts of 5 children and more, which

means that no matter how high the family income is, 5 children seems too much for a family to be taken care of. However, 5 children and more have more counts in low income families, which can be explained by the fact that these families need to have more kids as a way of supporting the family's livelihood.

The histogram from figure 5 shows that 30-40 hours of working happens most frequently and has the highest count. Which means that the majority of the Canadian families are working on average 30-40 hours with minimum earnings starting at 75000 annually. That is a good sign of indicating that the overall Canadian economy and GDP is doing well since the majority of the family is able to earn 75000 and are considered in the middle class. Moreover, it is directly proportional most of the times that when you work more hours, you will get paid more. There are differences in wages which is why there are different income levels, the minimum wage in Canada is also an important factor which guarantees a certain income for the hour worked, which is why we are able to see the 5 outstanding bars.

Figure 6's graph has a normal distribution curve, the highest bar is at 125000, which means that when working 30-40 hours per week, lots of people are able to earn an annual salary of 125000 and more. The end of the curve shows that high earners do not necessarily work longer. Majority of the people in Canada work on average only 30-40 hours, which is less than 8 hours per day, this also shows that Canada has a relatively expensive labor cost.

Table 3 is about a logistic linear regression model, by using age, total children and age having the first child, we are able to estimate the probability of having a grandchild. For example, by sub in an individual that has one child, age 23, had the first child at age of 22, by taking the log and calculating the p value, the probability of a young individual having a grandchild will be close to zero.

Weaknesses

One of the major weaknesses in the data is that income value is not accurate enough, through the analysis, the average income is to be around 125000 annually, which is considered to be a relatively high income in Canada. The weakness and flaw is that income should be changed to disposable income or after tax income, because different income levels will correspond to different income tax. If the income variable can be changed to after tax income, it will be more precise and more valuable in providing information on a family's income and salary, the income will not be as overwhelming as right now.

Next Steps

One improvement that could be made next time is to avoid the use of scatter plot, since we are dealing with a huge amount of data, using scatterplot will likely just generate a solid line instead of distributed dots, which is not straight forward enough when viewing the plot.

References

Rupnarain, Kimberly. "Why Do the Poor Have Large Families?" World Vision Canada, Organization, 13 July 2020, www.worldvision.ca/stories/why-do-the-poor-have-large-families.

Urbina, Jilber. "Create new variable by combining 2 Categorical variables in R", 29 Oct. 2014, https://stackoverflow.com/questions/26641602/create-new-variable-by-combining-2-categorical-variables-in-r.

smillig, "How do I change a value coded as 'Yes' to a value of 1 in R?", 20 Aug. 2012, https://stackoverflow.com/questions/12033960/how-do-i-change-a-value-coded-as-yes-to-a-value-of-1-in-r.

Appendix

GitHub Link: https://github.com/rexyizhouhu/STA304_Problem_Set_2.git