Multiple Linear Regression Model of GSS Data

Jo-Yen Lin

October 19th, 2020

Code and data supporting this analysis is available at: https://github.com/jlin213/Multiple-Linear-Regression-Of-GSS-Data

Abstract

Multiple Linear Regression Model allows statisticians to explore the relationship between two or more independent variables and one continuous variables. In this study, General Social Survey (GSS) conducted by Statistics Canada in 2017 is used to investigate the relationship between Canadians' income, age, marriage, and number of children. Through the multiple linear regression model, the study concludes that married Canadians generally have more children as they are older, regardless of income. However, income affects at which age Canadians to have more children when they are single.

Introduction

The General Social Survey (GSS) that conducts annually in 10 provinces of Canada provides insightful data to social and cultural trend of Canadians. This study is interested in how we can predict the age of Canadians through total number of children, their income, and their marital status. By examining the relationship of these variables, we can investigate the question such as "For Canadians with 2 children, what age does people achieve different income brackets?" or "Does marital status affect how many children people have at different stages of life?" Through the multiple linear regression model that will be introduced later, this study will answer all the preceding questions. Then, results, weakness, and improvements of the model will be discussed to conclude this analysis.

Data

The data was collected by Statistics Canada through General Social Survey (GSS) from February 1st to November 30, 2017 through telephone interviews and participation was voluntary. The population is targeted at all Canadians who are 15 years old or older and living in the 10 provinces. From Statistics Canada, "the GSS uses a frame that combines landline and cellular telephone numbers from the Census and various administrative sources with Statistics Canada's dwelling frame" (Government of Canada, 2020). Additionally, only one person from each household is selected. The key features of the survey include large sample size (43,000 people for year 2017), with diverse characteristics of childhood, immigration status, marriage, education background. GSS allows research in marriage and family to have extensive data. It also helps with investigating how Canadian families are facing changes at the current year compared to other years. This study chooses the 2017 data set, in particular, to have a fresh and most recent data of the social and family life of Canadians who participated in the survey.

Table 1: Descriptive Summary of Age, Number of Children, Marital Status, and Income

	Respondent Count
Less than $25,000	6724
$25,000 to $49,999	6143
$50,000 to $74,999	3877
$75,000 to $99,999	2021
$100,000 to $124,999	843
$125,000 and more	872

	Number of Children
Min.	0.00000
1st Qu.	0.00000
Median	2.00000
Mean	1.67876
3rd Qu.	3.00000
Max.	7.00000

	Age
Min.	15.00000
1st Qu.	37.30000
Median	54.20000
Mean	52.19726
3rd Qu.	66.80000
Max.	80.00000

	Respondent Count
Married	15792
Single, never married	4688

Figure 1

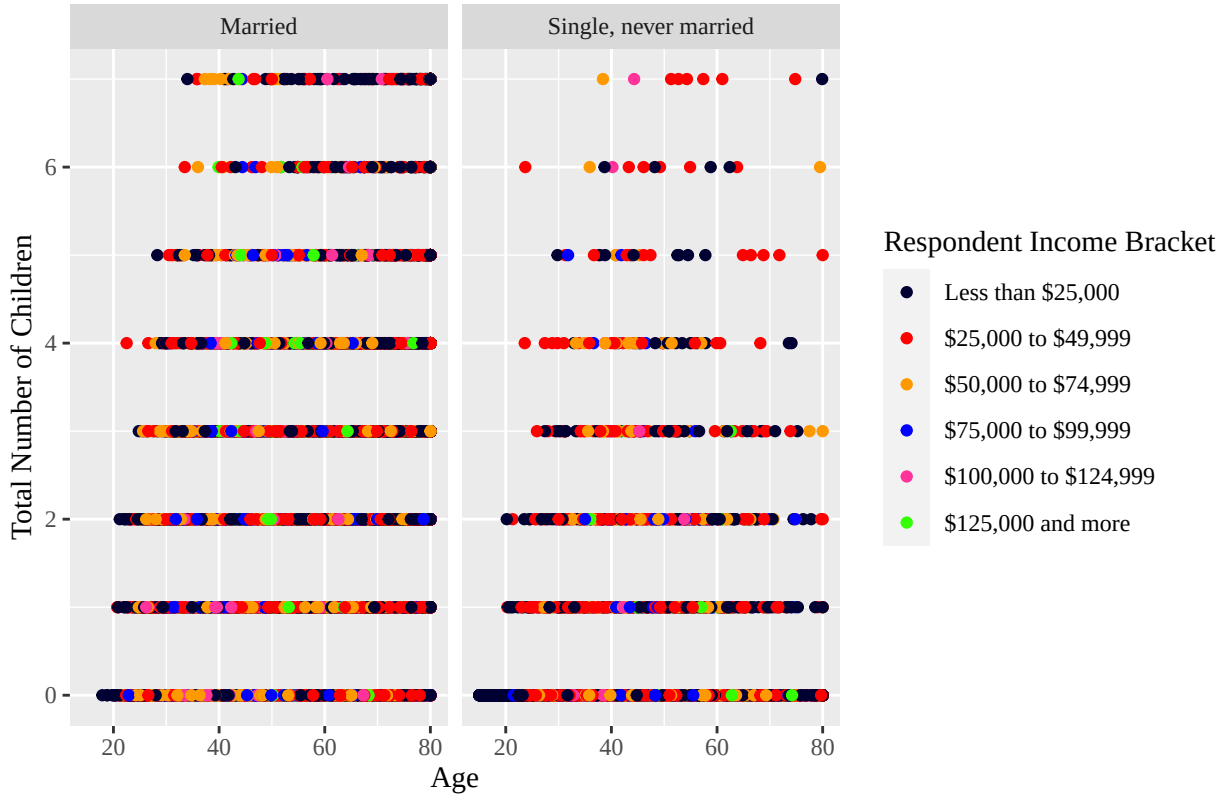
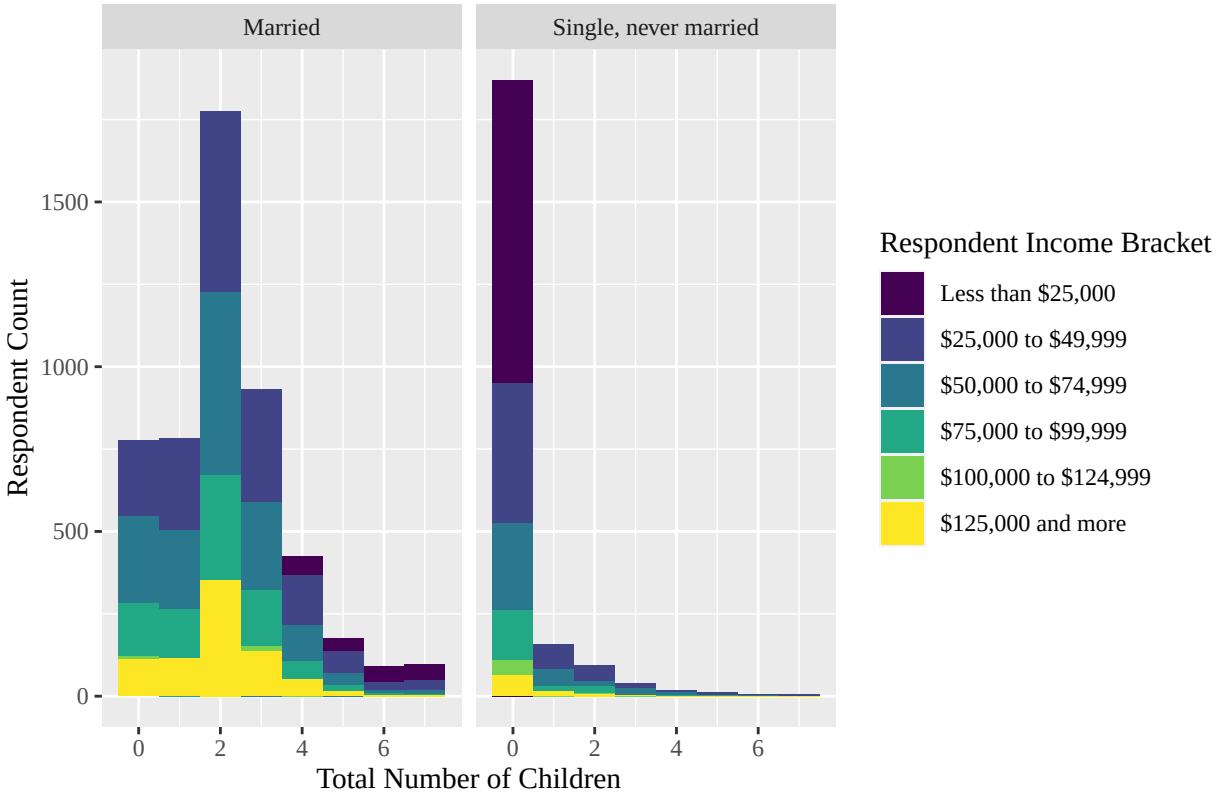


Figure 2



Using GSS raw data that was collected and cleaned, Figure 1 shows a scatter plot of the raw data, with age and total number of children as variables, with colours separating different income brackets, under different marital status. Figure 2 shows how different marital status reflects how many children each respondent have. This visualizes the distribution of overall number of children by different income brackets in the raw data.

Model

In this study, a multiple linear regression model is used to predict age of individuals based on their income, marital status, and total of children they have. A mathematical expression of the multiple regression model is:

$$y_i = \beta_0 + \beta_1 X_{i:25-49} + \beta_2 X_{inc:50-74} + \beta_3 X_{inc:75-99} + \beta_4 X_{inc100-124} + \beta_5 X_{inc:>125} + \beta_6 X_{children} + \beta_7 X_{mar:Married}$$

(1)

Table 2: Definition of Each Variables In the Model

Variable	Definition
$X_{i:25-49}$	Respondent's income is between 25,000 and 49,999
$X_{inc:50-74}$	Respondent's income is between 50,000 and 74,999
$X_{inc:75-99}$	Respondent's income is between 75,000 and 99,999
$X_{inc:100-124}$	Respondent's income is between 100,000 and 124,999
$X_{inc:>125}$	Respondent's income is more than 125,000

Variable	Definition
$X_{children}$	Total number of children
$X_{mar:Married}$	Marital status: Married, Divorced, Living common-law, Separated, Widowed

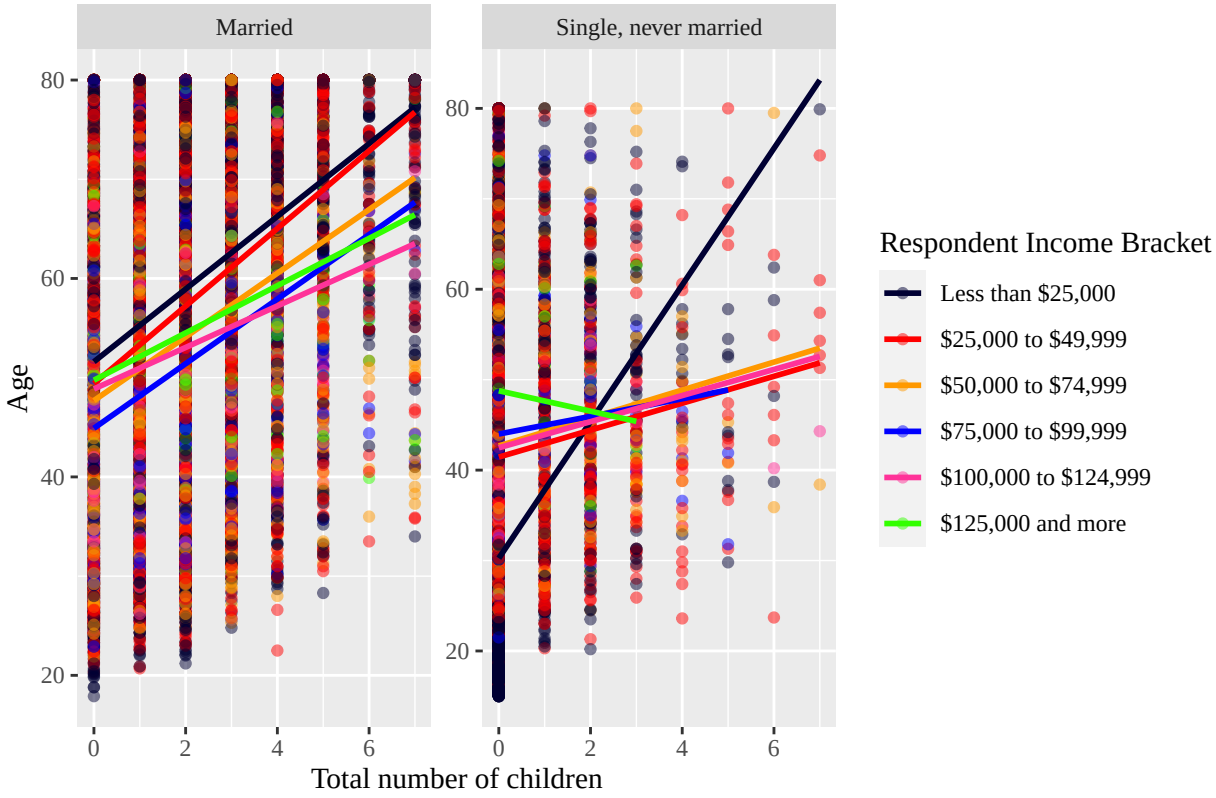
In this model and the data collected, income of the respondent is categorical and is factorized during the modeling. From Equation (1), if the individual has annual income between 25,000 and 49,999, $X_{i:25-49} = 1$, while other variables associate with income, such as $X_{i:50-74}$, will equal to 0. Similarly, marital status is also categorical variable; therefore, there is dummy variables $X_{mar:Single}$ and $X_{mar:Married}$. An important note that this model has done is modifying the marital status to be binary. That is, all statuses beside Single: never married, all fall with in Married category. This includes Married, Divorced, Living common-law, Separated, Widowed in marital status. For total number of children, it is discrete and not categorical variable. Therefore, there are no multiple dummy variables for total number of children.

Results

Table 3: Regression Coefficients of the Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	48.159740	0.2257885	213.295835	0.0000000
$X_{i:25-49}$	-1.846314	0.3882713	-4.755216	0.0000020
$X_{inc:50-74}$	1.457828	0.3546077	4.111102	0.0000395
$X_{inc:75-99}$	2.168918	0.3763339	5.763279	0.0000000
$X_{inc:100-124}$	-1.360229	0.3678510	-3.697771	0.0002181
$X_{inc:>125}$	-0.335565	0.3175266	-1.056809	0.2906111
$X_{children}$	3.667404	0.0811252	45.206726	0.0000000
$X_{mar:Married}$	-12.324942	0.2915348	-42.276055	0.0000000

Figure 3



From Table 3, the intercept is estimated to be 48.15, which means if a person has 0 child, and income of less than 25,000 and single, they are predicted to be 48 years old. β_1, which is the regression coefficient of the variable $X_{inc:>125}$, has a p-value of 0.29 that is greater than 0.05, which means that it is not significant different from other categories. However, it is important to note that the model should keep this category, as it does not make sense to throw away the "more than $125,000" income bracket.

Discussion

By referring to Figure 3, the linear predictions for participants who are married are generally similar across all income brackets. This suggests generally people who are married will have more kids when they are older regardless of income. It is also important to note that for those who are married have more data points in higher total number of children, as it makes sense for those to form family with their partners. From the result, we can conclude several trends. First, most of participants who are married gradually have more children at around the same time, regardless of income. Marital status also seems to affect people with less than 25,000 annual income the most, as the regression lines have significant difference in Figure 3.

A significant result from Figure 1: "Single, never married" is that those with less than 25,000 annual income are likely to have children at significantly older age. However, this regression line indicates a weakness of our model,there are little amount of data points in "Single, never married" with 4 or more children, and the prediction line for "less than 25,000" seems to be influenced by these small data points that can be potentially outliers. In addition, we can see that there is not enough information at all to make a regression line for "125,000 and more" in the same graph. On the other hand, for "Married", generally there are enough data points across each categories and discrete numbers. This suggests the GSS data does not have equal representation in each category, and with the lack of data points, this is certainly one of the weakness in the survey as well.

There are many areas that can be improved on the study, data, and the model. First, the study can investigate similar questions through sampling the data. Then, the scatter plot will not be overload with thousands of data points. It is also possible to reduce and simplify the model to investigate each variable's effect. For example, we can have $y(i) = \beta_0 + \beta_1 X_{mar:Married} + \beta_2 X_{children}$, to predict the age of the participants through their marital status and number of children. The model will have less constraint and can potentially deliver new results. Further, the model assumes each variables are independent of each other; however, it is possible that there are confounding variables that influences the results of this study. The GSS data offers large variability that many other variables, such as aboriginal group, province, education level, can all be potential variables to explore the wellness of Canadians.

References

Alexander, R. (2020, May 17). Telling Stories With Data. Retrieved from https://www.tellingstorieswithdata.com/

Faculty of Arts & Science, University of Toronto. (26, January 2017). Computing in the Humanities and Social Sciences. Retrieved from http://www.chass.utoronto.ca/

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

Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.".

Xie, Y., Dervieux, C., & Riederer, E. (2020). R Markdown Cookbook. CRC Press.