What are the reasons to cause low-income in Canada?

Xu Yang 1004234740 Wei-chieh Li 1004548741 Mengyuan Wang 1005239341
Hanyang Liao 1003811930

Group 114

19/10/2020

Abstract

How to help people get rid of low incomes is a hard question worldwide, even for developing countries like Canada. Even if the Canada Government has a relatively sufficient budget to help low-income people, it is more important to allocate those resources properly. In this study, we tried to figure out reasons that cause low personal incomes. Thus, we used data from the General Social Survey (GSS main survey) in 2011 and built models to find the relation between individual annual income with both physical aspects (like ages) and mental aspects like mental health. After our study, we hope to make useful directions for the government to consider allocating the government spending wisely and sufficiently and helping people who are facing low income or bad living well-being to get rid of this trap. As a result, Canadian citizens would have better living standards. It is very important to guarantee basic life needs if the government hopes to ensure human rights for its citizens. This study might also be a good guide for countries other than Canada and face the same situation.

Introduction

Even in developed countries like Canada, there are still people struggling in a low-income line or facing low life well-being. Thus, our survey aims to explore any reason that might cause the low income in Canada. Therefore, we obtained the data, the twenty-fifth cycle of the General Social Survey (GSS main survey) in 2011. This data shows the whole situation of Canadian families, and there are also several variables about incomes.

However, instead of exploring family incomes, we would explore incomes for individuals. In other words, we want to explore what might be reasons that might cause a person with low income. The reason we did the study about personal incomes is that we think it is more efficient for a person with higher income in a family, instead of expecting more people to work but with lower income. In reality, there must be a case that most of the family members cannot work. For example, wives have to take care of the family, and kids are still young. This study is also beneficial for the government to make more valid policies to avoid low-income individuals rather than make subsidies for every family.

According to the minimum wage which is $10.25 per hour in 2011, we also considered the low-income cut-off which is $20,500 in Canada. In the data set, we only have categorical data of individual incomes, so we set people whose individual income is less than 20,000 dollars as a low-income group.

We picked variables which might be possible to cause income not only from physical aspects like the age but also other mental aspects. And then, we built a logistics model to predict the probability for a person with a low income.

Data

The data we used is called the public use microdata data file (PUMF) for the twenty-fifth cycle of the General Social Survey (GSS main survey) conducted from Feb. to Nov. in 2011. It contains many different useful aspects to overview the well-being of Canadian families. It contains the basic background of Respondents like birth date, personal history, marriages, main activities, work experience, educational background, and family situations like the number of children, unions, organizations and any financial support.

The target population for this survey includes all persons 15 years of age and older in Canada but excluding several specific areas and full-time residents of institutions. It used a stratification method, which divided the frame by 10 provinces based on their geographic area and formed by grouping in several areas for a total of 27 strata.

Moreover, it used the Random digit dialling (RDD) method, which randomly generates a list of phone numbers used to reach households. The RDD frame comprises all possible 10-digit phone numbers based on the area codes and 3-digit prefixes.

The survey used computer-assisted telephone interviewing (CATI) for data collection, which means that all respondents are interviewed by telephone. Overall, it gets 22,435 respondents. This sampling method is good because it divided different frames by their telephone number, which is accurate. Also, considering populations in several areas are larger, they were divided by groups that are more reasonable for data collecting.

The questionnaire is formed by 16 sections with an extra control form, and there are more variables set in each section. All questions in the questionnaire are very brief and obvious, which are basic personal information. I think it is a very good questionnaire because it is very user-friendly in telephone interview communication. Also, considering respondents from all age groups and different, clear questions are easy to understand and answer.

Overall, it is enough to say that this is a very clear and useful data because it contains many different variables to imagine the well-being of families from all areas of Canada. Moreover, the proper sampling method is very important to make the dataset accurate. The frame was divided by geographic areas, which make sure

it covered all areas of Canada. The estimation weights were adjusted using a post-stratification technique to cover individuals from different sex-age groups. It also had an error control system to reduce its potential effects.

Also, it had a lot of references to avoid sampling errors since this is the twenty-fifth cycle already. As a result, it can be collected as a very large number of respondents and relatively accurate answers, which is very useful and valuable for further research.

However, there are several potential drawbacks, as well. For example, since it is based on telephone interviews, those without a telephone are excluded. Also, it is good for a questionnaire that is very detailed, but it might cause the interviews longer, and the participation rate will be lower.

Model

For this section, we continued our analysis by building a logistic regression model to figure out the relationship between the response variable (annual personal income of the respondent) and predictor variables (age, the number of children, education status, marriages, health status, birth county, born parents, unable to pay rent, number of unions, and work hours per week). Besides, this model is only suitable in Canada because the data comes from the General Social Survey. The target population is all persons 15 of age and older in Canada.

We choose this logistic regression model because we believe the predictor variables that we listed above are the main factors that can affect the individual incomes. Also, we are very curious about which kind of respondents can get rid of low income.

We divided these variables into two categories. One is called numerical variables; they are age, number of children, and work hours per week, with specific numbers to value. Most of the variables belong to categorical variables; they are education status, health status, birth country, born parents, unable to pay rent, and the number of unions.

For categorical variables, it's described very detailedly in the survey. For our study convenience, we tried to shrink categories for variables. For example, there are 7 categories for the annual individual income variable, and we only classified income below or above $20,000 to distinguish a person who is in low income or not.

(Figure 1)

```
##          (Intercept)                   age     childrern_number
##          -5.86367108           0.06067158          -0.04296761
##    education_status2    education_status3     marriage_status1
##          -0.65990195          -1.21429584           0.18428526
##       health_status1       birth_country1        born_parents1
##           0.53194442           0.51193925          -0.14630234
## unable_to_play_rent1    number_of_unions1    number_of_unions2
##           0.88601264           1.02249560           1.03191684
```

```
##    number_of_unions3    number_of_unions4    number_of_unions5
##           0.94814972           0.72914920          -0.87088633
##  work_hours_per_week
##           0.09342586
```

Using the data selected we build a logistic regression model with the following formula:

$$\log(p/1\text{-}p) = B_0 + B_1 age + B_2 children\ number + B_3 education\ status2 + B_4 education\ status3 + \cdots + B_{15} work\ hours\ per\ week$$

P stands for the probability of a person who does not face the low annual personal income which we set the line at $20,000 here. For categorical variables here, we set dummy variables to distinguish different types in their variables. For example, the variable education_status2 shows that for a person who has a diploma/certificate from community college or trade/technical, the value would be 1. Otherwise, it would be zero–the same logic for all dummy variables from B_3 to B_{14}.

B_0 is the intercept, which equals -5.863671, which means when the age, number of children, work hours per week are equal to 0 while dummy variables education status, health status, birth country, born parents, unable to pay rent, and the number of unions are equal to "No", the log odds of annual personal income equal to -5.863671. In other words, the probability for a person who is 0 years old, with a doctorate/masters/bachelor's degree, without any children and legally married and with an excellent health condition, not born in Canada, able to pay rent, not participating in any union and working zero-hour per week, being able to get rid of low income will be 0.2833% (calculated by log-odds above). In reality, there might be limitations for input. For example, ages must be greater than 15.

B_1 to B_{15} all represent coefficients in our equation. For B_1 representing ages, which is 0.0606, it means that when the age increases by 1, the odds of annual personal income greater than 20,000 dollars will increase by 1.06 times. B_2= -0.042968 means when the number of children increases one, the odds of annual personal income which is greater than 20,000 will decrease by 4.2%.

For B_3 to B_4, they are all about the educational background of the respondents. When respondents have a diploma/certificate from community college or trade/technical instead of obtaining a doctorate/master/bachelor's degree, the odds of not being low income will decrease by 48%. Similarly, when respondents neither have a diploma/certificate from community college or trade/technical nor obtaining a doctorate/master/bachelor's degree, the odds of not being low income will decrease by 70%.

For B_5 and B_6, it means that when a person is legally married and with an excellent health condition, the odds of not being low income will grow by 1.2 times and 1.7 times, respectively. The birth country and born parents also affect personal income.

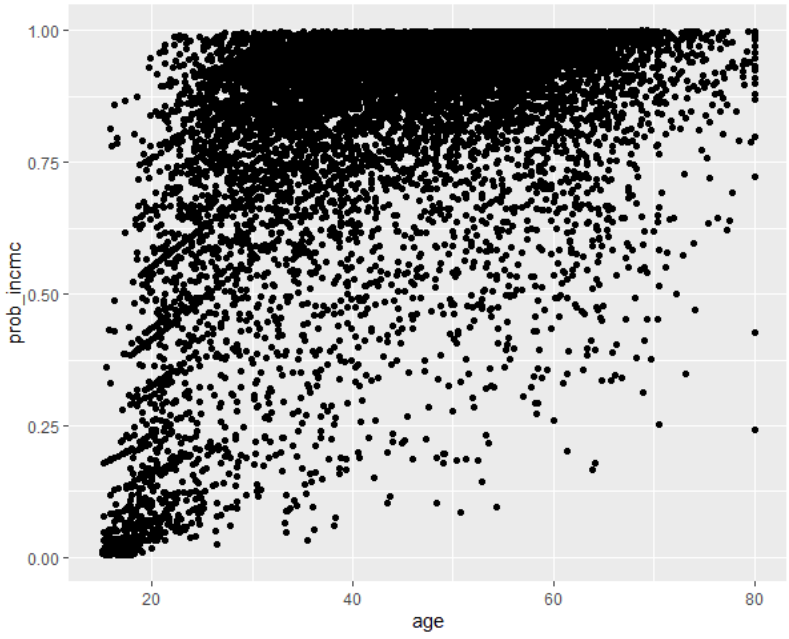
When a person is born in Canada, the odds of not being low income will grow by 1.669 times.

For B_8, it means that when respondents were born or adopted at birth, if they were living with both their birth or adoptive mother and their birth or adoptive father, then the odds of not being low income will decrease by 13.6%. For B_9, if a person cannot pay rent at any time during the past 12 months, the odds of not being low income will grow by 2.42 times. For B_{10} to B_{14}, they all relate to the number of unions that a respondent participates in. Generally, if people participate in more unions, the odds of not being low income will grow. Only if a person participates in more than 5 unions, the odds of not being low income will decrease by 58%. The last variable measurement of work-hour per week indicates that when people work one more per week, the odds of not being low income will grow by 1.0979 times.

The caveats are the data only suitable in Canada, cannot represent worldwide, and some respondents didn't answer all the questions. Also, please note that for the logistics regression model, the p-value or t-value in the summary table does not have any statistical meaning. Thus we do not make any evaluation based on R square here.

Results

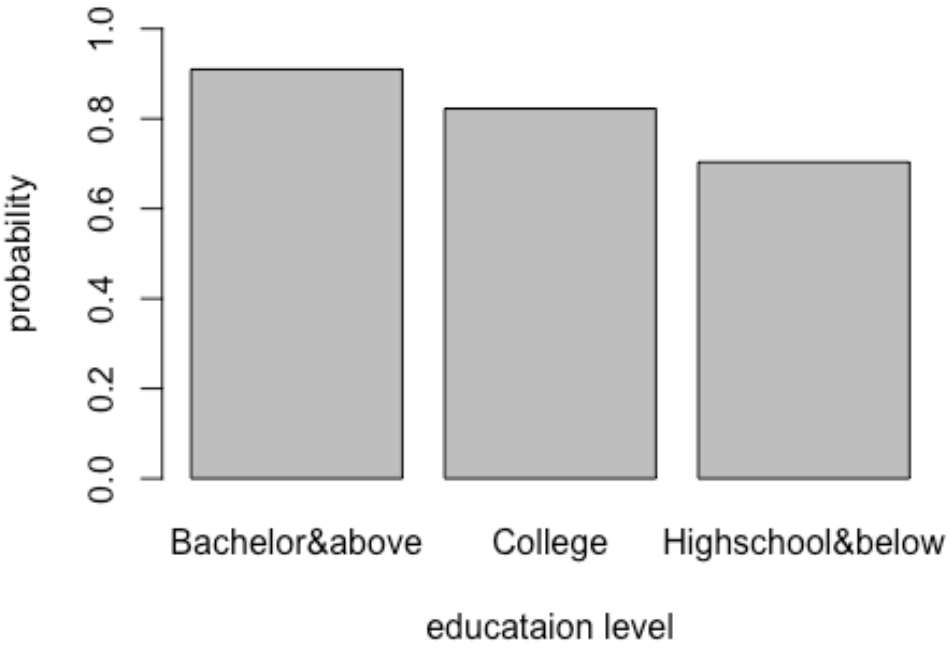
(Figure 2)



By looking at the scatter plot of probability by age, we find out that as age increases, the probability of respondents that are above the low-income line increases. As our model shows, respondents increase 1 unit of age. The odds of annual personal income, which is greater than $20,000 will increase by 1.06 times. However, the probability may vary in the same age of respondents. For example, at the age of 40,

the minimum probability is 0.63, but the maximum is 0.99. Combined with the scatter plot, we can notice that the age affects probability a lot. Respondents under 20 years old may still be in high school and cannot earn money, so the mean probability of respondents under 20 years old is only 0.16. However, respondents between 30 to 40 years old have the ability and responsibility to earn money for the family. The mean probability of respondents between 30 to 40 years old is 0.88, which means the mean of probability is affected by age. Most respondents under 20 years old have low incomes. They distribute in the left down corner of the graph. As we can see, between ages 20 to 30, the probability goes up quickly. After 30 years old, most respondents have a high probability of earning more than 20000 dollars annually. Age affects probability in each age gap. Hence age is a significant variable in our model.

(Figure 3)



By looking at the barplot of the probability of respondents that are above the low-income line and the educational background of the respondent, the respondent who has a bachelor's and above education background has the highest probability of earning more than $20000 annually. Respondents with a college diploma in the middle and respondents with high school and below have the lowest probability. In our model, compared to the bachelor and above diploma, respondents with a college diploma have odds of probability above the low-income line 48% lower.

Respondents with high school and below diplomas have 70% lower odds of probability.

Based on our analysis, we can conclude that age and education background are two essential elements that will impact personal incomes. If a person is relatively elder and with a bachelor's degree or above, it is more likely for her/him to get rid of low-income financial traps. This conclusion is also highly related to the model which we built before, and it also shows that this model is reasonable and strong. According to our other research, we also have strong references to support that for young people who are less educated, they have higher odds to get into financial trouble.

Overall, the government can consider allocating more spending on those two aspects, and we discussed more possible policies in detail in the discussion section.

Discussion

According to the modeling and results analyzed above, we detected several essential elements for an individual to get rid of low-income traps. Firstly, we found that age is a very significant aspect which would affect personal incomes. Even though we could tell from the graph that the income difference for people from the 40-aged group is huge, generally, younger people (from 20 to 30) are more likely to face low income. It makes sense intuitively because most of them are not well-experienced, and they just graduated from colleges. In that case, they are facing huge financial pressure. To help those young people, the government can give more subsidies for them or reduce tuition loan payment if they just graduated from school. Moreover, the government can also encourage colleges to create more coop or PEY programs so that young people can obtain experience and help them to get involved in society quickly.

Secondly, we can find from the histogram that education background also matters a lot for individual incomes. Those of people who obtained bachelor's degrees or above (including doctorate and master's degrees) are more likely to earn more than $20,000 annually. On the contrary, for those who graduated from high school or below, they are likely to face low-income. This clearly shows the importance of higher education. In this case, the government should allocate more funds to develop higher education like colleges and universities so that those institutions have more ability to adapt more students. On the other hand, the government can spend more money to support students who cannot pay for college tuition.

Moreover, we also found that more children in a family will also impact on personal incomes. Thus, the government could give more financial supports for more-children family.

Finally, other elements are significant to incomes as well, but they are more about personal preference. For example, marriage status or mental health will also impact incomes. However, the government may not react positively to those aspects because they depend on personal decisions. However, it would be very interesting to do further social science researches, and they provide important references.

Weaknesses

Firstly, as we mentioned before, this dataset is based on Canada, so results might cause bias if this study would be used in other countries. However, it is still efficient for the Canadian government to make decisions already.

Secondly, the most important variable called INCMC （Annual personal income of the respondent） is a categorical variable. Thus we can't find an exact linear relationship between income and other aspects but divide the income by low-income and non-low-income two groups to build logistics. However, if we can find exact numbers of income or the survey could collect this data by an exact number, the study would be much more reliable.

Third, the reason we picked this dataset is that it has individual income data which we are looking for instead of general family incomes. It also has a very detailed explanation for each variable. However, this dataset is from 2011 which is a little far away from now. It is better to find a closer dataset so that we can avoid time series bias.

Fourth, other personal lifestyles might cause contradictory results. For example, those of people who are unable to pay rental rent or mortgage payments are less likely to be low-income. I think the reason that a certain group of people would be unable to pay rent or mortgage payments is that they get used to living beyond their means. They might not be low-income, but they are still facing financial traps.

Moreover, this survey is conducted through telephone, which means some drawbacks are shown by telephone surveys. For example, some people might not use phones in their daily lives, which means they cannot be interviewed. Sometimes, telephone calls are perceived as telemarketing and thus negatively received by potential respondents. This situation increases the difficulty of investigation, hard to reach respondents. The timing must be carefully considered. Some of the questions are too long. Maybe some of the respondents don't have the patience to answer all of them. Besides, some of the questions are too specific and related to personal privacy, which means the non-answered situations increase.

Overall, in social research, it is ideal for conducting a face-to-face survey than a telephone survey because better responses can be recorded when the respondents could have a physical interview.

Next Steps

For the next steps, we can follow up on the exact incomes we were looking for previously. Also, since the Internet became more important in our lives, we would be able to use the Internet to distribute surveys, which is more efficient for collecting more reliable data. We could also build a system to follow up personal incomes by time series to indicate the efficiency of related policies. Moreover, we could do more research on the government policies for low-income people and consider applying financial models into the study so that we could have a better understanding. It will help us have a more reasonable process to select related variables so that we can build a proper model. Finally, we might also use other algorithms to evaluate whether the model is good or not since we did not do that in this study.

References

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

T. Lumley (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1): 1-19

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

General social survey on family (cycle 25) (2011), https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda3+gss25