Income Level Analysis Using Logistic model

Xinyi Chen, Zehui Yu, Yuxin Xie

October 19, 2020

Income Level Analysis Using Logistic model

Xinyi Chen, Zehui Yu, Yuxin Xie

October 19, 2020

Abstract

With the rapid development of the world society, in order to fully embracing the economic globalization and seize the chances, people have to work harder, make more money and getting themselves more prepared to face the challenges and opportunities. One major goal for most of the people pursing successful life is earning more and making more money, as income level is one of the most classical symbol to identify a person's achievement. However, income level may be affected by many aspects. Higher educational degrees are typical prerequisites for highly compensated work. What's more, labor market economics imply a positive relationship between hours worked and income level. However, despite decades of progressive efforts, there's still some inequality in workplaces across gender, age and working location. This lead us to investigate those factors that may affect a person's income level, modeled using data set from Canadian General Social Study of year 2017's Family results. Logistic regression tested the log odds of personal income level against gender, age, education level, working location and working hours. Employee could use this model as a guideline to know how to improve their income with least effort. It is also useful for students to make of decisions after they graduate, such as which provinces to go for working and which kind of education level to achieve.

Introduction

Our goal is to find out what are the attributes that impact a person's income level, and how they impact on it. The data set is obtained from Canadian General Social Study - 2017 General Social Survey (GSS) on the Family. Then the data is further narrow down our scope to 6 variables focus only on age, gender, province, education, location and income level. The main question is whether or not those people of highest income level work more, have higher educational level of lower, in a particular province. Or are the gender and age have influences. Perhaps the well-paid choose to work more because each additional hour worked is highly lucrative, thus earning more simply because they work more. On the other hand, they may not work as much because the high educational level guarantees them a relatively higher hourly payment thus they can earn enough to fulfill their needs in a shorter amount of time without gender discrimination. Or those earning low salaries must work more because they live in a smaller province without much opportunities and they are old hence that is the only way to earn an adequate income to meet the ends. The answer to this question can be boiled down to many incentives. The relationship between income and many other factors is a question of great practical and significance. Therefore, we perform demographic analysis to see which variables are correlated and how they influence on income level. We classified the income level as two categories, those earned a yearly income higher than 125000 are identified as high income and the rest are classified as Regular. A logistic model was built to predict income level and analyze how the attributes affect income level. Our hypothesis is that the income level is generally be determined by age, gender, hours you worked, education background and the location of your workplace.

Data

We obtained the data set from Canadian General Social Study - 2017 General Social Survey (GSS) on the Family, the data set contains family's basic background information(such as, age, sex, education and income), social-economic conditions, living conditions and well-being using stratified sampling method. The target population includes all non-institutionalized persons with 15 years of age and older, living in the 10 selected provinces of Canada. It uses a new frame, created in 2013, that combines telephone numbers (landline and cellular) with Statistics Canada's Address Register, and collects data via telephone. Data was collected via computer assisted telephone interviews (CATI). In order to carry out sampling, each of the ten provinces were divided into strata, each record in the survey frame was assigned to a stratum within its province which is good. A simple random sample without replacement of records was next performed in each stratum. The total number of variables is 81 with 20602 observations. The data set is good because it contains 20,602 responses which is large, and many variables ensures that it is representative. However, the drawbacks are only based on Canadian data which may not be true for other countries, and it focus on just one year instead of a period of time to capture the change.

Model

We continue our analysis by building a logistic model to predict the probability of earning the high income level against age, gender, province, education, and average hours worked. It contains 6 variables and could be applied in Canada but not other countries. Age is a numerical variable. By the cleaning process, we obtain a variable named "is_male", 1 indicates the person is male and 0 means female. Variable education has 7 categories, ranged from below High school diploma or a high school equivalency certificate, College, Less than high school diploma or its equivalent, Trade certificate or diploma, University certificate or diploma below the bachelor's level, University certificate, diploma or degree above the bachelor's level and higher education, to Bachelor's degree. These represents their level of education. Provinces are sort alphabetically, which means Alberta is level 1 and Saskatchewan is level 10. The average working hour have 5 levels, from 1 to 5 they are 0 hours, 0.1 to 29.9 hours, 30.0 to 40.0 hours, 40.1 to 50.0 hours, 50.1 hours and more. We classified a person with high income level as his or her yearly income is greater or equal than $125000. The reason we choose the above variable it is because we believe that a person's income level is not only depend on its number of hours worked but also other social influences with objective and subjective reasons. Using the data selected we build a logistic model with the following model.

```
## # A tibble: 12,895 x 7
##      age is_male province average_hours_w~ education income_responde~
##    <dbl>   <dbl> <fct>    <fct>            <fct>     <fct>
##  1  52.7       0 Quebec   30.0 to 40.0 ho~ High sch~ $25,000 to $49,~
##  2  51.1       1 Manitoba 50.1 hours and ~ Trade ce~ Less than $25,0~
##  3  28         1 Quebec   30.0 to 40.0 ho~ College,~ Less than $25,0~
##  4  63.8       0 British~ 30.0 to 40.0 ho~ High sch~ Less than $25,0~
##  5  15.7       1 British~ 0.1 to 29.9 hou~ Less tha~ Less than $25,0~
##  6  40.3       0 Manitoba 50.1 hours and ~ Universi~ $25,000 to $49,~
##  7  26.8       0 Quebec   30.0 to 40.0 ho~ Trade ce~ $25,000 to $49,~
##  8  30.6       0 British~ 30.0 to 40.0 ho~ Bachelor~ $25,000 to $49,~
##  9  68.8       1 Ontario  30.0 to 40.0 ho~ High sch~ $125,000 and mo~
## 10  33.8       0 British~ 40.1 to 50.0 ho~ Bachelor~ $75,000 to $99,~
## # ... with 12,885 more rows, and 1 more variable: incomelevel <fct>
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```

```
##
## Call:
## svyglm(formula = incomelevel ~ age + is_male + province + average_hours_worked +
##     education, design = example.design.strs, family = "binomial")
##
## Survey design:
## svydesign(id = ~1, strata = ~province, data = df, fpc = ~fpc)
##
## Coefficients:
```

```
##                                                                         Estimate
## (Intercept)                                                           -16.715010
## age                                                                     0.043458
## is_male                                                                 1.037374
## provinceBritish Columbia                                               -0.718482
## provinceManitoba                                                       -1.215515
## provinceNew Brunswick                                                  -1.227435
## provinceNewfoundland and Labrador                                      -0.358438
## provinceNova Scotia                                                    -1.279168
## provinceOntario                                                        -0.559143
## provincePrince Edward Island                                           -1.674830
## provinceQuebec                                                         -0.949984
## provinceSaskatchewan                                                   -0.578832
## average_hours_worked0.1 to 29.9 hours                                  11.391138
## average_hours_worked30.0 to 40.0 hours                                 12.219336
## average_hours_worked40.1 to 50.0 hours                                 12.992144
## average_hours_worked50.1 hours and more                                13.241234
## educationCollege, CEGEP or other non-university certificate or di...   -0.918601
## educationHigh school diploma or a high school equivalency certificate  -1.593074
## educationLess than high school diploma or its equivalent               -2.203226
## educationTrade certificate or diploma                                  -0.889914
## educationUniversity certificate or diploma below the bachelor's level  -0.725763
## educationUniversity certificate, diploma or degree above the bach...    0.775574
##                                                                       Std. Error
## (Intercept)                                                             0.536557
## age                                                                     0.003076
## is_male                                                                 0.097847
## provinceBritish Columbia                                                0.158970
## provinceManitoba                                                        0.241083
## provinceNew Brunswick                                                   0.230462
## provinceNewfoundland and Labrador                                       0.200576
## provinceNova Scotia                                                     0.223329
## provinceOntario                                                         0.128933
## provincePrince Edward Island                                            0.334526
## provinceQuebec                                                          0.152249
## provinceSaskatchewan                                                    0.207215
## average_hours_worked0.1 to 29.9 hours                                   0.509744
## average_hours_worked30.0 to 40.0 hours                                  0.491480
## average_hours_worked40.1 to 50.0 hours                                  0.497217
## average_hours_worked50.1 hours and more                                 0.501616
## educationCollege, CEGEP or other non-university certificate or di...    0.130334
## educationHigh school diploma or a high school equivalency certificate   0.151806
## educationLess than high school diploma or its equivalent                0.291902
## educationTrade certificate or diploma                                   0.172510
## educationUniversity certificate or diploma below the bachelor's level   0.248253
## educationUniversity certificate, diploma or degree above the bach...    0.113730
##                                                                       t value
## (Intercept)                                                           -31.152
## age                                                                    14.129
## is_male                                                                10.602
## provinceBritish Columbia                                               -4.520
## provinceManitoba                                                       -5.042
## provinceNew Brunswick                                                  -5.326
## provinceNewfoundland and Labrador                                      -1.787
## provinceNova Scotia                                                    -5.728
## provinceOntario                                                        -4.337
## provincePrince Edward Island                                           -5.007
## provinceQuebec                                                         -6.240
## provinceSaskatchewan                                                   -2.793
## average_hours_worked0.1 to 29.9 hours                                  22.347
## average_hours_worked30.0 to 40.0 hours                                 24.862
## average_hours_worked40.1 to 50.0 hours                                 26.130
## average_hours_worked50.1 hours and more                                26.397
## educationCollege, CEGEP or other non-university certificate or di...   -7.048
## educationHigh school diploma or a high school equivalency certificate -10.494
## educationLess than high school diploma or its equivalent               -7.548
## educationTrade certificate or diploma                                  -5.159
## educationUniversity certificate or diploma below the bachelor's level  -2.923
## educationUniversity certificate, diploma or degree above the bach...    6.819
##                                                                       Pr(>|t|)
## (Intercept)                                                             < 2e-16
## age                                                                     < 2e-16
## is_male                                                                 < 2e-16
## provinceBritish Columbia                                              6.25e-06
## provinceManitoba                                                      4.67e-07
## provinceNew Brunswick                                                 1.02e-07
## provinceNewfoundland and Labrador                                      0.07395
## provinceNova Scotia                                                   1.04e-08
## provinceOntario                                                       1.46e-05
## provincePrince Edward Island                                          5.61e-07
## provinceQuebec                                                        4.52e-10
## provinceSaskatchewan                                                   0.00522
## average_hours_worked0.1 to 29.9 hours                                   < 2e-16
## average_hours_worked30.0 to 40.0 hours                                  < 2e-16
## average_hours_worked40.1 to 50.0 hours                                  < 2e-16
## average_hours_worked50.1 hours and more                                 < 2e-16
## educationCollege, CEGEP or other non-university certificate or di...  1.91e-12
## educationHigh school diploma or a high school equivalency certificate  < 2e-16
## educationLess than high school diploma or its equivalent              4.72e-14
## educationTrade certificate or diploma                                 2.52e-07
## educationUniversity certificate or diploma below the bachelor's level  0.00347
## educationUniversity certificate, diploma or degree above the bach...  9.55e-12
##
## (Intercept)                                                           ***
## age                                                                   ***
## is_male                                                               ***
## provinceBritish Columbia                                              ***
## provinceManitoba                                                      ***
## provinceNew Brunswick                                                 ***
## provinceNewfoundland and Labrador                                     .
## provinceNova Scotia                                                   ***
## provinceOntario                                                       ***
## provincePrince Edward Island                                          ***
## provinceQuebec                                                        ***
## provinceSaskatchewan                                                  **
## average_hours_worked0.1 to 29.9 hours                                 ***
## average_hours_worked30.0 to 40.0 hours                                ***
## average_hours_worked40.1 to 50.0 hours                                ***
## average_hours_worked50.1 hours and more                               ***
## educationCollege, CEGEP or other non-university certificate or di...  ***
## educationHigh school diploma or a high school equivalency certificate ***
## educationLess than high school diploma or its equivalent              ***
## educationTrade certificate or diploma                                 ***
## educationUniversity certificate or diploma below the bachelor's level **
## educationUniversity certificate, diploma or degree above the bach...  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.8896743)
##
## Number of Fisher Scoring iterations: 13
```

Formula is

$$log(P/(1-P)) = \beta_0+\beta_1 age+\beta_2 ismale+\beta_3 provinceBC+...+\beta_9 provinceSa+\beta_{10} worked0.1to29.9hrs+...+$$

$$\beta_{13} worked50.1hrsandmore+\beta_{14} eduCollege+...+\beta_{19} eduuniveristyaboveBachelor$$

P is the probability of a person's income level is High

Variables is_male to education University certificate, diploma or degree above the bach… are dummy variables. When a person is male, is_male equals to 1 and 0 otherwise. This logic applies to all other dummy variables.

$\beta_0 = -16.715010$ means when a person's age=0, gender is female, province is Alberta, average hours worked is 0 hours, education is Bachelor's degree, log odds of income level is High equals to -16.715010.

$\beta_1 = 0.043458$ means when age increase 1 year, log odds of income level is High will increase by 0.043458.

$\beta_2 = 1.037374$ means when a person is a male, log odds of income level is High will increase by 1.037374 than when is a female.

β_3 to β_9 are dummy variables of provinces. $\beta_3 = -0.718482$ indicates that when a person's province changed from Alberta to British Columbia, log odds of income level is High will reduce by 0.718482. β_{4-9} have the same meanings as β_3.

β_{10} to β_{13} are dummy variables of average hours worked. $\beta_{10} = 11.391138$ indicates that when a person's worked hours changed from 0 hours to 0.1-29.9 hours, log odds of income level is High will increase by 11.391138. β_{11} to β_{13} have the same meanings as β_{10}

β_{14} to β_{19} are dummy variables of education. $\beta_{14} = -0.918601$ indicates that when a person's education level changed from Bachelor's degree to College, CEGEP or other non-university certificate or di… , log odds of income level is High will reduce by 0.918601. β_{15} to β_{19} have the same meanings as β_{14}

Results

Figure1: Income level based on age

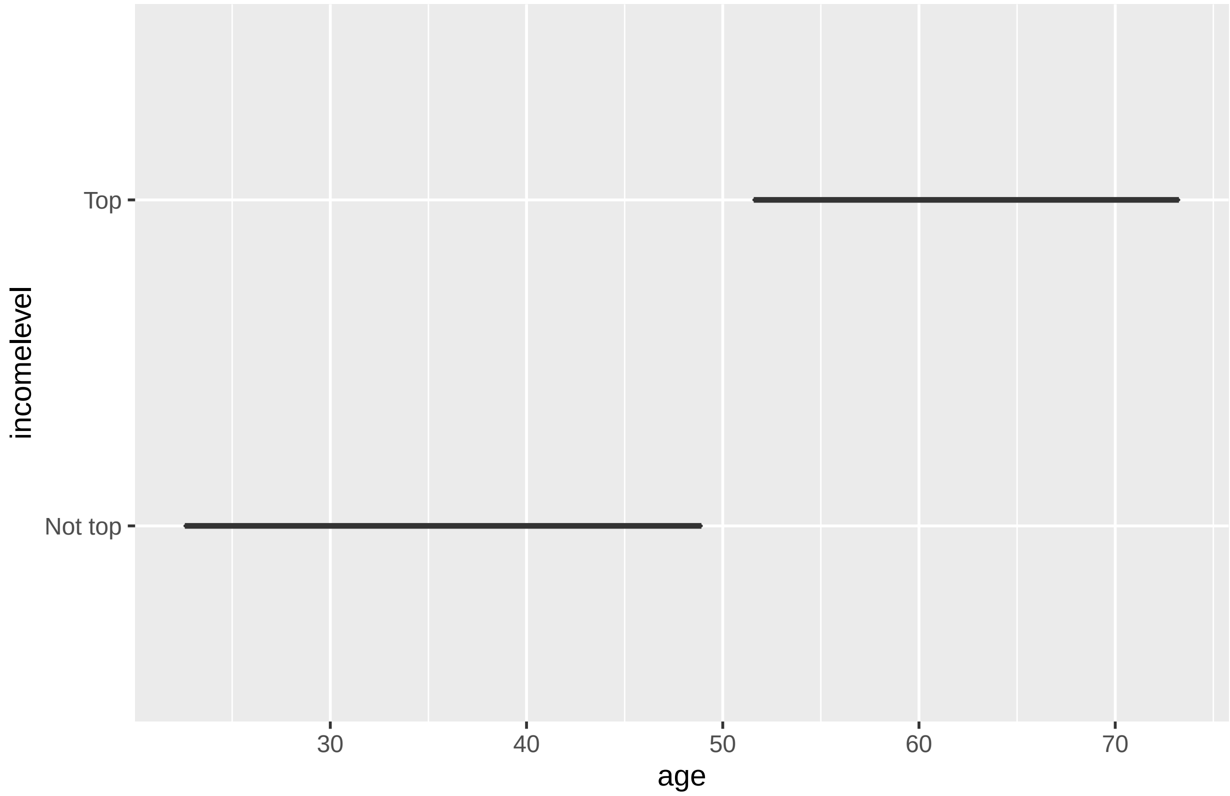


Figure2: Income level based on average_hours_worked

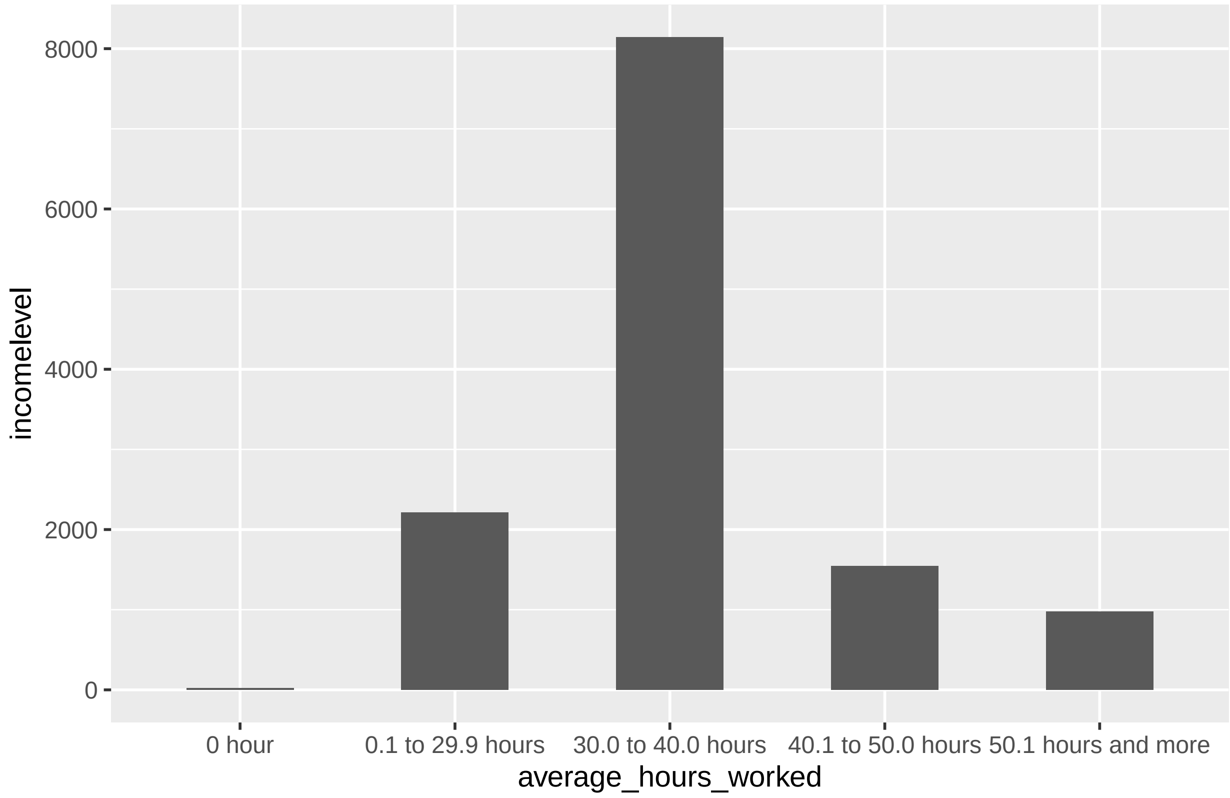


Figure3: Income level based on education level

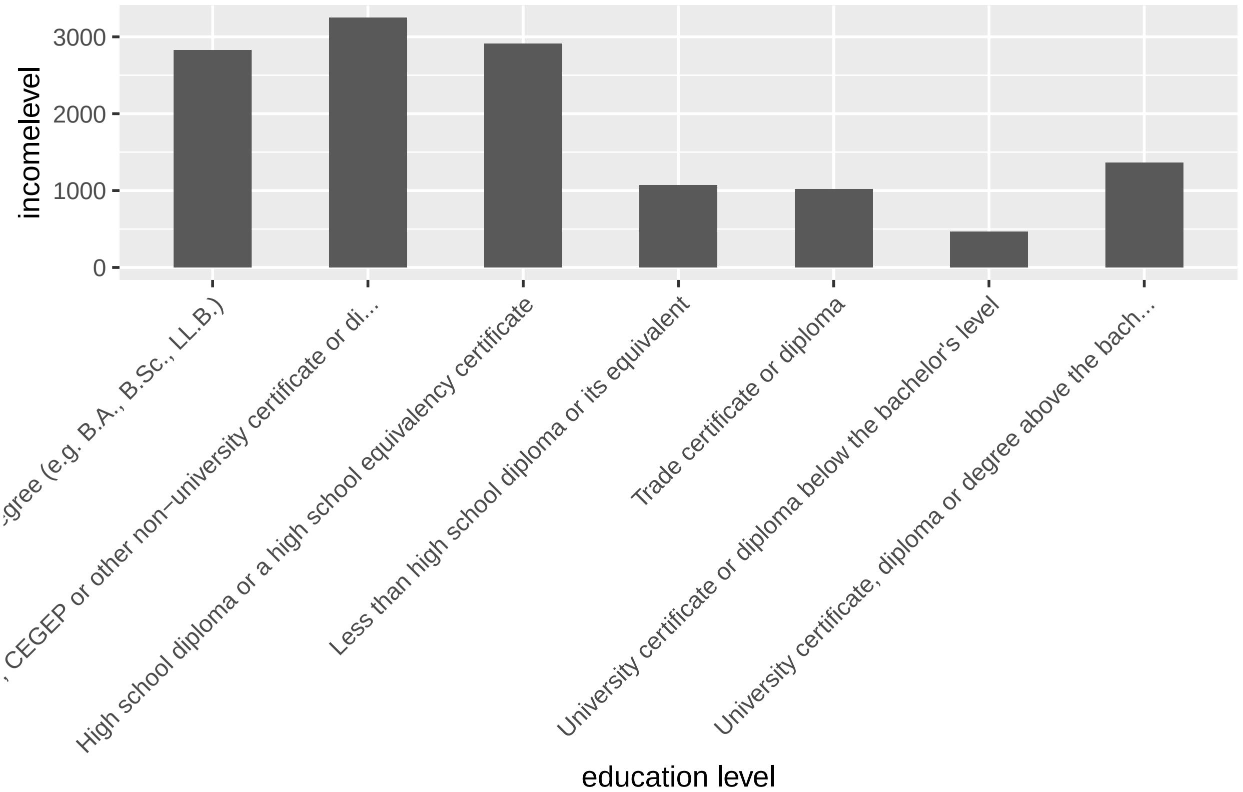
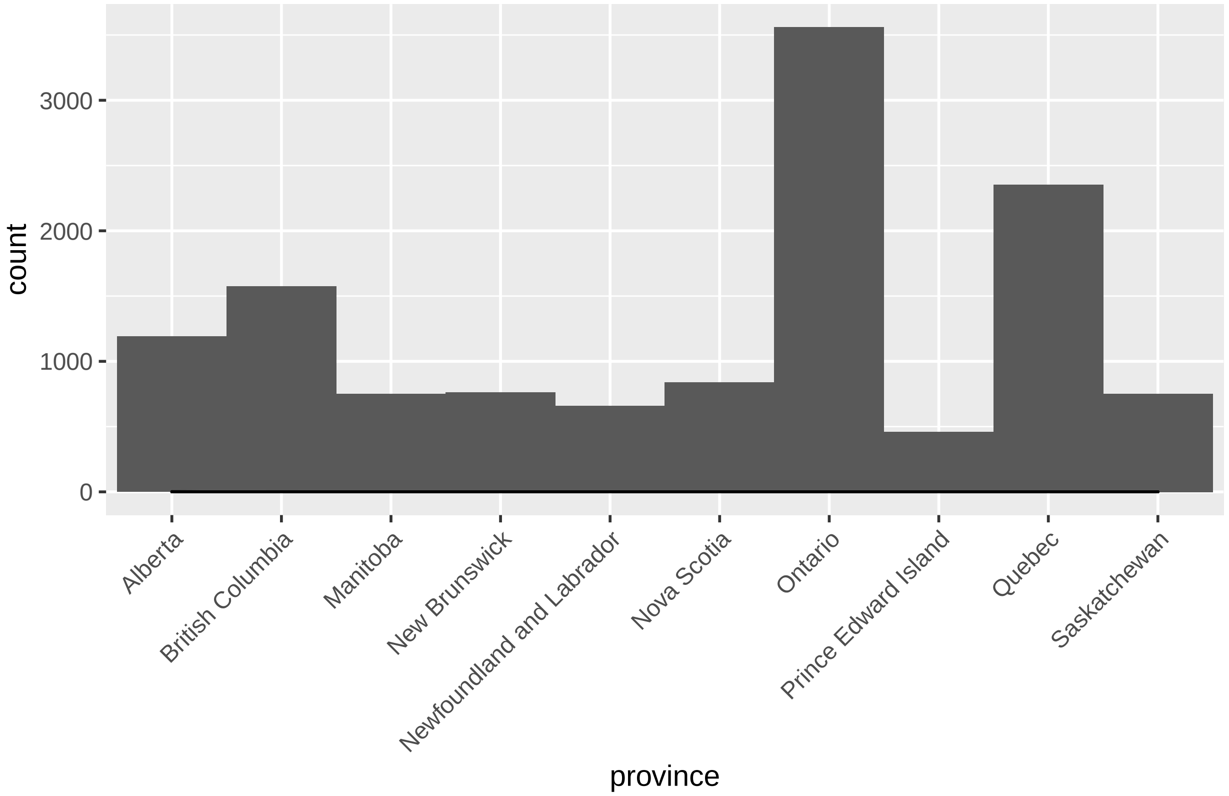


Figure4: Income level based on provinces



The dataset contains 12895 variables and 8 variables. By referring to Figure2, it is clear to see that for those who earn a higher salary are typically older, with a median age of 51, and the logistic model gives as the same feedback as when age increase by 1 year, log odds of income level is top will increase by 0.043458. However, given the same age, the income level can be vary different. We can tell while age ensure a boundary for a person's income level, but it is not the one and only one factor. What's more, the income level could also change as the result of number of hours working per week. It is easy to observe that for people who earn a higher salary usually works 30 to 40 hours per week from Figure 2. Even though, given the same income level, the same age and number of hours working per week, other factors could also affect a person's earning. By referring to Figure 3, higher income levels are typically from high educational level, which also is consistent with the results from our logistic model that with all other conditions remained the same. Besides, moving from small city to urban area, such as moving from Province Edward to Province British Columbia, more job opportunities provides a higher income. According to our logistic model, when a person's province changed from Alberta to British Columbia, log odds of income level is High will reduce by 0.718482. On the other hand, gender seems like a big problem as an important factor to decide a person's income level, which can also be validate by model result. Therefore, we can conclude that a person from high income level is typically a male that works 30 to 40 hours per week from higher education background.

Discussion

This model provides a guideline for people who want to improve their income level. The model proves that the factors including age, gender, province, education, and average hours worked do have impacts on the income level. Any changes in one of the factors can increase or decrease the probability of getting a higher income. For instance, the model indicates that a higher education level will increase the probability of obtaining a higher income, and moving from Province Alberta to Province British Columbia may also lead to a higher income, etc. Therefore, these results obtained from the analysis of the model help us achieve the original goal of the study which is to is to explore the relationship between income and many other factors(age, gender, province, education, and average hours worked). Hence by using this sample, people can find a way to improve their income and know which factor has a more influential impact on the income level. Moreover, the employer can also use it as a reference to set their employees' wage payments according to employees' education level, their working hours, etc. Students could also benefit from this model by referring it and make better decisions after they graduate, to decided whether to further their study for a higher degree level or going to work directly and where should they go for a work. However, as this model is predicted by using data from the Canadian General Social Study of the year 2017's Family results, the results concluded from the model may not be suitable to use in another country.

Weaknesses

By using a logistic regression model, we are assuming that each observations are independent from each other. However, this may not be true in real life thus we ignore some random variability. Besides, the data set is not large enough to represent the situation within the whole country. There is no job data, different jobs always lead to different levels of income, thus the factors contributing to the changes in income level are limited in our model. Besides, during an economic booming period versus an economic recession period, the income level may show a different changing pattern due to the change in real-life labor supply and demand.

Next Steps

We can increase the sample size to get more data, getting more people involved and collect more information, thus make a more reliable and representative conclusion based a relatively larger sample size. Also, we will collect more detailed factors such as job type to better explain what factors will influence the income level and how these factors influence, since different job type has different income settings that may affect a person's salary. From the economic perspective, the income level will be determined by the demand and supply of labor, therefore we may also apply the basic demand and supply model to improve this analysis. The last but not the least, adding Bayesian inference is also helpful for our study.

References

Bibliography:

1. Cycle 31, General Social Survey: Families, Public Use Microdata File Documentation and User's Guide. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf
2. Government of Canada, Statistics Canada. "Canada at a Glance 2018 Population." Population - Canada at a Glance, 2018, 27 Mar. 2018, www150.statcan.gc.ca/n1/pub/12-581-x/2018000/pop-eng.htm.
3. 2017 datafile, https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/subsda3
4. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
5. https://homepage.divms.uiowa.edu/~luke/classes/STAT4580/catone.html
6. https://rkabacoff.github.io/datavis/Bivariate.html
7. Samantha-Jo Caetano(2020), data cleaning code (gss_cleaning.R).https://q.utoronto.ca/courses/184060/files/9422740?mo
8. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Appendix

GitHub link: https://github.com/xinyi-chen-16/Sta-304-PS2/tree/main