A logistic analysis of income status based on age, ethnicity and education

Gen Cao 1005109696, Jia Yuan Liu 1004793841, Yuanjie Ji 1004768469, Yunhan Chen 1004710583

19/10/2020

Abstract

The research objective of this project is to classify an individual's income status based on their age, ethnicity and maximum education level obtained and study the impact of each variable on an individual's income status. A binary response variable "income_status" is constructed and is used as the response variable in our logistic regression model. Results from our model show that while age is not a meaningful factor in determining an individual's income status, ethnicity and education are significant factors. Further improvements include limiting the sampling population to working individuals to account for more accurate interpretation for the age variable, sampling more visible minority respondents, and including other potential factors of influence in our model to avoid underfitting. The findings of this project can provide insights on social values and trends on wealth distribution.

Introduction

For contemporary society, income is an important indicator and embodiment of estimating a person's workability and social status. We first built a binary variable, defined the income state variable with the boundary of greater than 50,000 and less than 50,000 as our y response variable. Then we choose 'age', 'vis_minority', 'education' for our x variables.

The first is about age. In the study of the relationship between age and income, we are based on curiosity and analysis of the following points. The first is the acceptance of technological and social development brought about by age. For older workers, their rich social experience is a good indicator of increasing their wealth. At the same time, the rapid development of society and technology is a huge challenge to the rising trend of the age. For the same things, generally speaking, the acceptance and acceptance speed of young people will be faster. The overall social impact is the key to our next data research by carrying out our (GSS) data set.

Another equally significant factor we noticed is the visible minority group in society. According to the estimated data analysis of Statistics Canada, in the next 2031, 29% to 32% of the Canadian population will be visible minorities. Visible minorities will account for 63% of Toronto's population and 59% of Vancouver's population (Statistics Canada, 2010). Because of these trends, in the 20th and 21st centuries, Canada will become a country rich in ethnic diversity and race. This visible difference is also an intuitive judgment and impression that we usually accept.

Last but not the least, the effect of education level on income is also the key to our research. First, the cost of receiving education is very high. We are curious about the relationship ratio between education investment and income, which is a direct output. Men with apprenticeship certificates had particularly high incomes in 2015. This reflects the strong demand for workers with apprenticeship certificates in the labor market as a whole. Their average income is $72,955, which is 7% more than men with a college degree, 31% more than high school men with the highest degree, and 11% less than men with a bachelor's degree. The growth rate is faster than the income of men with all other educational qualifications. The income of men with a bachelor's degree increased by 6%, and the income of men with a college degree increased by 8%. (Statistics Canada)

Data

We extracted four variables "income_respondents", "age", "vis_minority" and "education" from the Canadian General Social Survey (GSS) 2017 cleaned data set (Alexander, 2020). Data for 2017 GSS on Families was collected from February 2 to November 30, 2017. The target population for the 2017 GSS include individuals who are 15 years or older, excluding residents of the Yukon, Northwest Territories and Nunavut and full-time residents of institutions (Statistics Canada, 2020). Based on Statistics Canada 2018 estimates, the population for Canada excluding the Yukon, Northwest Territories and Nunavut is 23,812,605 in 2017 (Statistics Canada, 2020). The frame of GSS was the linkage of a list of telephone numbers in use (compiled from various sources including telephone companies, Census population etc.) and the address register GSS's data collection method is based on telephone interviews (Statistics Canada, 2020). 91.8% of reached households were eligible (containing one or more 15 years or older respondents), and those ineligible were terminated from the study (Statistics Canada, 2020). The target sample size was 20,000, while the actual sample size was 20,602 (Statistics Canada, 2020).

The sampling strategy was based on a stratified sampling method, meaning that the frame is distributed to each stratum in every province, and a simple random sampling without replacement (Statistics Canada, 2020). A minimum sample size had to be reached within each stratum to ensure the collected sample is not biased (Statistics Canada, 2020).

Our interested variable of study is respondents' income. Based on the GSS Cycle 31 Microdata and Documentation and User's Guide, we discovered that income information was obtained through linkage with respondents' tax data. This method of data collection minimizes respondent bias, where respondents inaccurately report information. The decrease of respondent bias can increase the validity of the questionnaire.

Data was collected via computer assisted phone interviews (CATI). Phone interview data collection methods and long survey time will cause a lot of data missing. When doing data analysis, we observed a lot of missing values. Some respondents will be reluctant to provide relevant data. The survey results we received are only one result that the public is willing to provide.

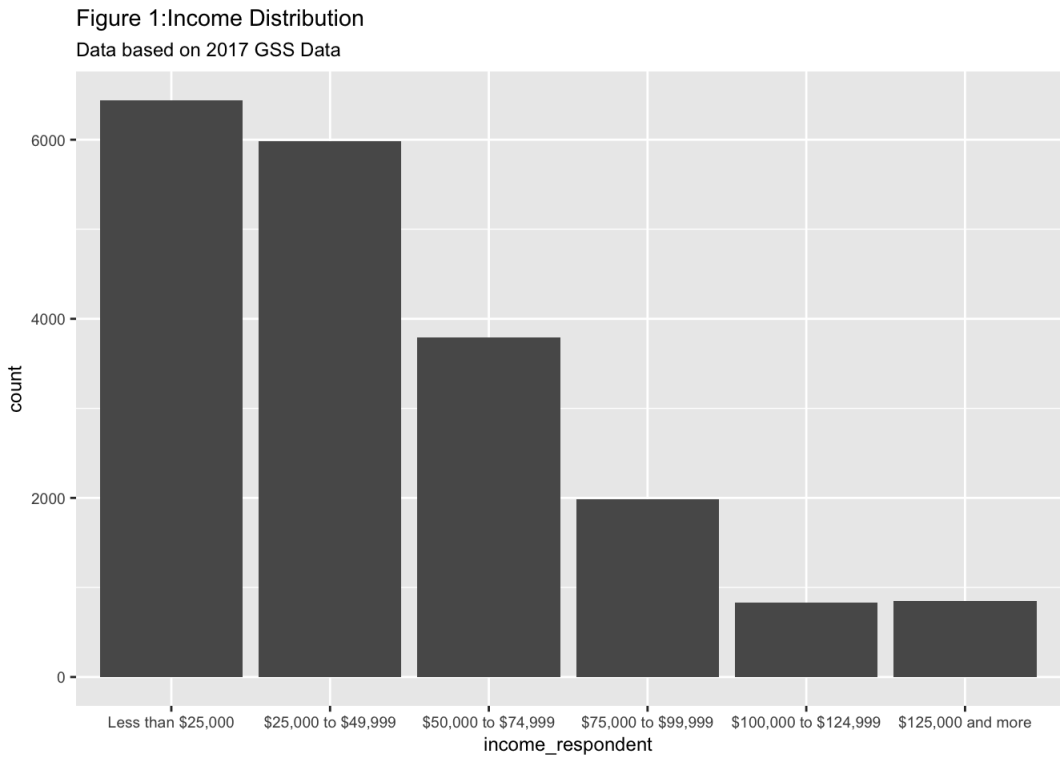
Exploratory analysis is performed on the income distribution of respondents.Figure 1 shows the income distribution of respondents. From the figure, we can see that approximate 60% respondents' annual earning is less than /$49,999, and this forms our basis in setting the boundary for determining income status.

```
## [1] 0.01655179
```

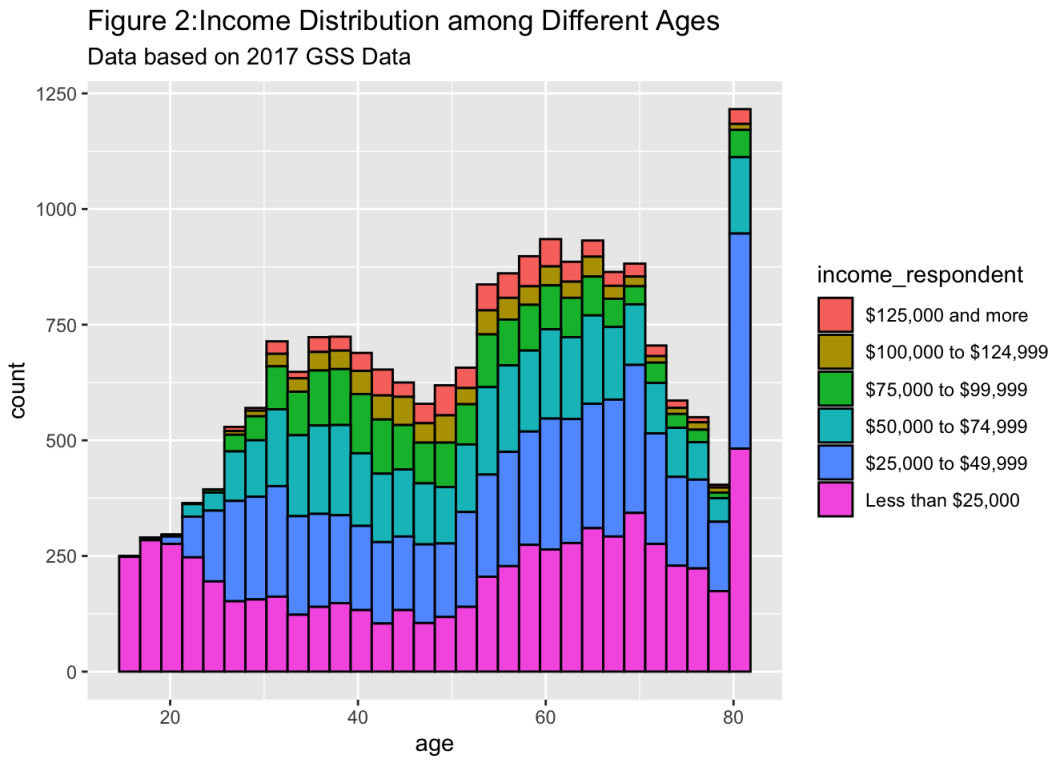
```
## [1] 0.006795457
```

```
## [1] 0
```

```
## [1] 0
```



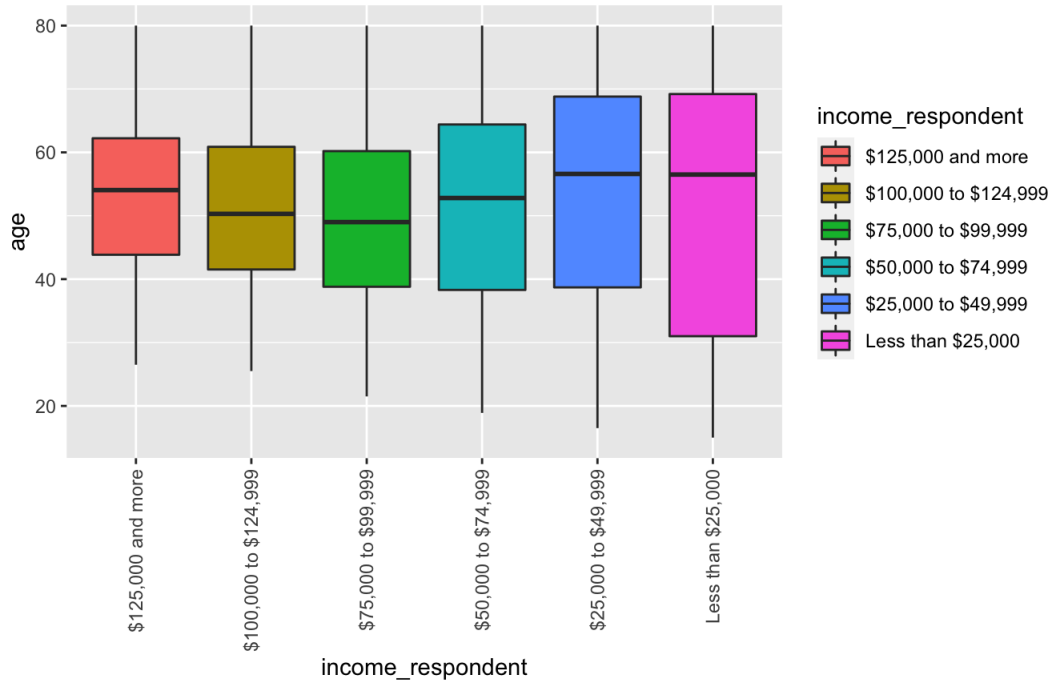
Barplot of Income distribution



Income Distribution among Different Ages

Figure 3:Boxplot of Income Distribution among Different Ages

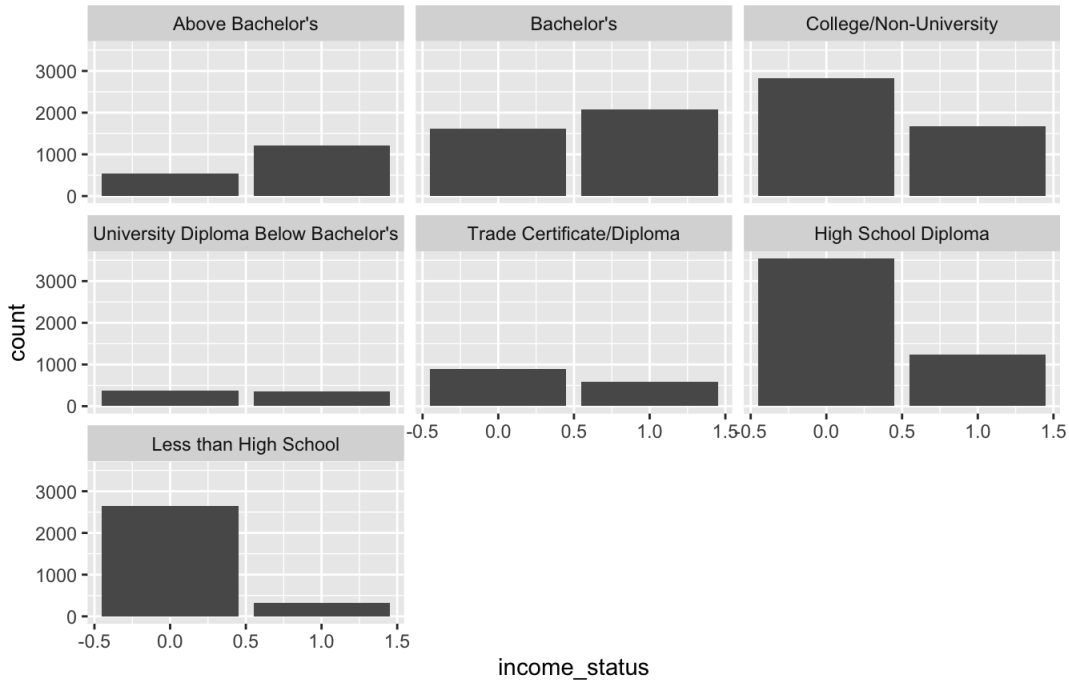
Data based on 2017 GSS Data



Boxplot of The Income Distribution among Different Ages

Figure 4:Income Status of Different levels of Education

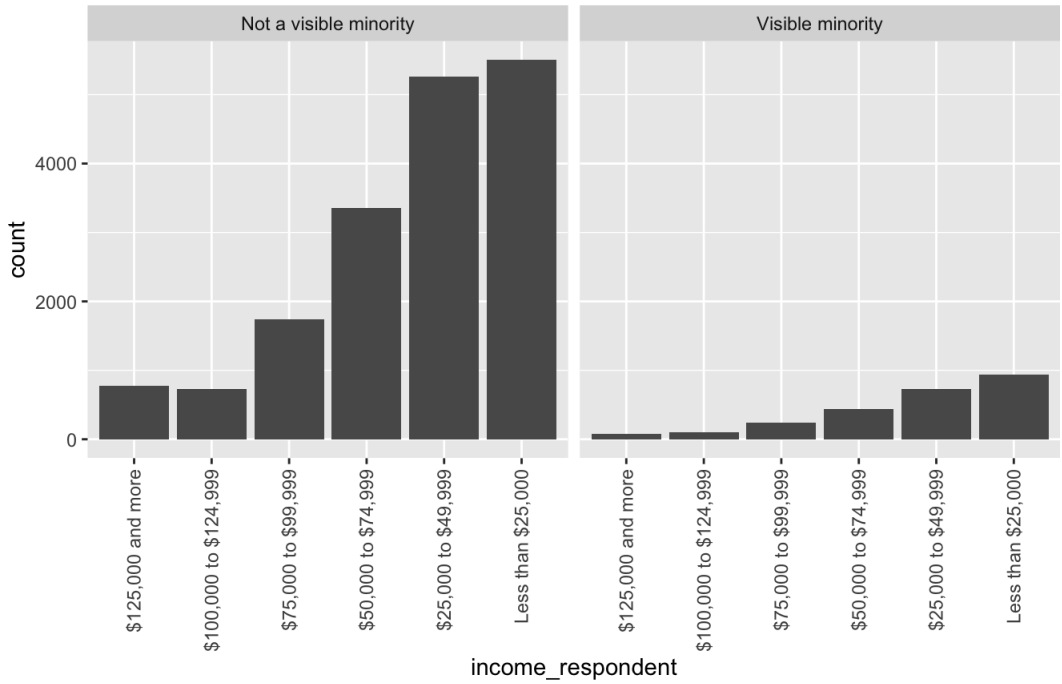
Data based on 2017 GSS Data



Income Status of Different levels of Education

Figure 5:Income level for Visible Minority and Non Visible Minority

Data based on 2017 GSS Data



Income level for Visible Minority and Non Visible Minority

```
##
## Call:
## glm(formula = df$income_status ~ age + vis_minority + education,
##     family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5820  -0.9916  -0.4940   1.0239   2.3142
##
## Coefficients:
##                                             Estimate Std. Error z value
## (Intercept)                                 9.156e-01  7.313e-02  12.521
## age                                        -7.315e-05  9.409e-04  -0.078
## vis_minorityVisible minority               -5.602e-01  4.921e-02 -11.383
## educationBachelor's                        -5.384e-01  6.180e-02  -8.712
## educationCollege/Non-University            -1.365e+00  6.070e-02 -22.491
## educationUniversity Diploma Below Bachelor's -9.107e-01  9.111e-02  -9.996
## educationTrade Certificate/Diploma         -1.316e+00  7.505e-02 -17.536
## educationHigh School Diploma               -1.920e+00  6.203e-02 -30.950
## educationLess than High School             -2.956e+00  7.869e-02 -37.572
##                                            Pr(>|z|)
## (Intercept)                                  <2e-16 ***
## age                                           0.938
## vis_minorityVisible minority                 <2e-16 ***
## educationBachelor's                          <2e-16 ***
## educationCollege/Non-University              <2e-16 ***
## educationUniversity Diploma Below Bachelor's <2e-16 ***
## educationTrade Certificate/Diploma           <2e-16 ***
## educationHigh School Diploma                 <2e-16 ***
## educationLess than High School               <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26303  on 19879  degrees of freedom
## Residual deviance: 23540  on 19871  degrees of freedom
## AIC: 23558
##
## Number of Fisher Scoring iterations: 4
```

Table 1: Summary of GLM

```
##                                                 2.5 %       97.5 %
## (Intercept)                                 0.772313196  1.058965615
## age                                        -0.001917382  0.001771085
## vis_minorityVisible minority               -0.656606250 -0.463704761
## educationBachelor's                        -0.659502599 -0.417251566
## educationCollege/Non-University            -1.484154346 -1.246221007
## educationUniversity Diploma Below Bachelor's -1.089294168 -0.732165439
## educationTrade Certificate/Diploma         -1.463059292 -1.168887210
## educationHigh School Diploma               -2.041271141 -1.798131801
## educationLess than High School             -3.110693996 -2.802240417
```

Table 2: Confidence Intervals

	2.5 %	97.5 %
(Intercept)	0.7723132	1.0589656
age	-0.0019174	0.0017711
vis_minorityVisible minority	-0.6566063	-0.4637048
educationBachelor’s	-0.6595026	-0.4172516
educationCollege/Non-University	-1.4841543	-1.2462210
educationUniversity Diploma Below Bachelor’s	-1.0892942	-0.7321654
educationTrade Certificate/Diploma	-1.4630593	-1.1688872
educationHigh School Diploma	-2.0412711	-1.7981318
educationLess than High School	-3.1106940	-2.8022404

Table 3: ANOVA Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	19879	26302.73	NA
age	1	16.01155	19878	26286.72	6.3e-05
vis_minority	1	23.15978	19877	26263.56	1.5e-06
education	6	2723.78491	19871	23539.78	0.0e+00

```
##            1            2            3            4            5            6
## "HighIncome"  "LowIncome" "HighIncome"  "LowIncome" "HighIncome" "HighIncome"
```

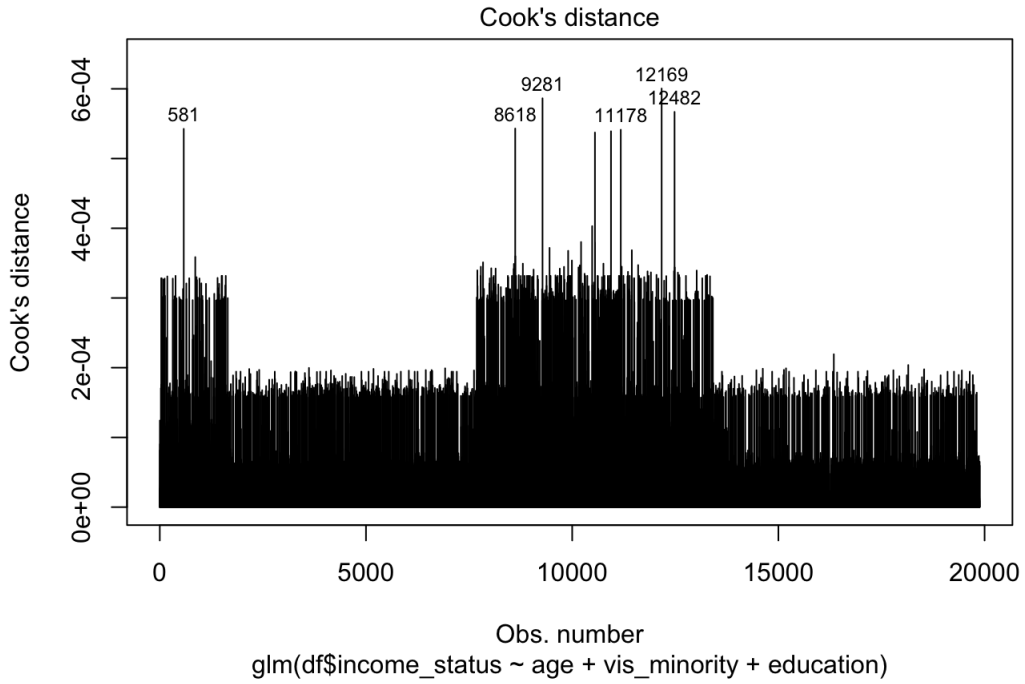


Figure 6: Influential points

Table 4: Variance Inflation Factors

0 rows | 1-1 of 11 columns

```
##                  GVIF Df GVIF^(1/(2*Df))
## age          1.046700  1        1.023083
## vis_minority 1.067699  1        1.033295
## education    1.059121  6        1.004798
```

Model

We use a logistic model for this study. While a linear regression model analyzes the linear association between the response variable Y and the predictor variable X, a logistic regression model is helpful in outputting well-calibrated probabilities of the classification of the outcome and the predictor variables. The objective of the study is to classify respondents into high-income and low-income groups based on their age, ethnicity and education and study the level of effects of each variable on predicting the response (high/low income), which is more suitable with the purpose of a logistic regression model. In addition, the response variable Y for linear regression must be numerical, but our interested “income_respondent” variable from the GSS dataset is categorical. Thus, a logistic regression model is more suitable for our purpose of study and also easier to implement and interpret.

Logistic regression model requires the assumption of the response variable to be a binary variable. We constructed the y response variable “income_status” based on the “income_respondent” variable separated into two groups with the boundary of 50,000 dollars annual income, define y_i = 1 or 0. We estimated a logistic regression model based on the generalized linear model function, with a “binomial” family indicating a binary outcome.

The data was collected from simple random sampling without replacement in each geographic area in 2017, meaning that each respondent of the sampled data is only being recorded in this study once and has no relationship with other respondents. These independent observations fulfills an important requirement of logistic regression Statistics Canada, 2020). In our model,individual survey respondents are denoted by $i \in [1, 2, \ldots, N]$; the log odds of our outcome “high/low income” is modeled as a linear combination of the predictor variables $X_i = (X_{i1}, X_{i2}, X_{ik})' $. We chose to use the age predictor variable as a numerical variable instead of grouping respondents into different age groups. Our decision stems from the reason that the use of age group categorization can vary by culture and context, which may not be representative of every individual within the age range. Keeping the numerical data enables our model to take into effect of every individual’s age at the individual-respondent level. The predictor “vis_minority” is a categorical variable since it has two categories “Visible Minority” and “Not a Visible Minority”, and there is no intrinsic order to the categories. The predictor “education” is an ordinal variable, where it has 7 categories of educational experience that can be ordered. The predictors “vis_minority” and “education” are included in the logistic model using “dummy” variable coding. Each value of a category is represented by numerical representation, with 1 indicating the presence and 0 indicating the absence.

Let $p_i = P(y_i = 1 \mid X_i)$. The following equation states that the probability that some individual i is either high-income (>$50,000) or low-income (<$49,999) depends on their age, ethnicity and highest level of education. $beta_0$ is the intercept of the model, and $beta_1, beta_2, beta_3$ are coefficients for each predictor variable.

Model diagnostics are performed to ensure the analysis is informative. By plotting Cook's Distance, we are able to detect observations that strongly influence fitted values of the model. The Variance Inflation Factor test is also performed to detect the problem of multicollinearity; the predictor variables age, visible minority and education have VIF close to 1 reflects that the regressors are not correlated with each other. We estimated our model using glm via the R "stats" Package (R Core Team, 2019b).

$$\log (p_i / (1 - p_i) ,= B_0 + B_1 Age_i + B_2 VisMin_i + B_3 Education_{i1} + B_4 Education_{i2} + B_5 Education_{i3} + B_6 Education_{i4} + B_7 Education_{i5} + B_8 Education_{i6}$$

Results

Figure 2 and Figure 3 shows the income distribution for the different age groups. We could see the interquartile range (IQR) for people who are earning less than $25,000 are much larger than other respondents?? groups, which means the ages of respondents in the group whom earning less than $25,000 are very diverse.

Figure 4 shows the Income Status of Different levels of Education. The left bar in each plot shows the number of respondents who earn the income below $50,000 per year and the right bar in each plot shows the number of respondents who earn the income over $50,000.

Figure 5 shows the income level for visible minority and non visible minority, the income level for visible minority and non visible minority follows the same trend.

Figure 6 shows the measurement of the influence for model.

Table 1 shows the results of the logistic regression model. The intercept B_0 is 9.156e-01. The β_1 is -7.315e-05 The β_2 is -5.602e-01 The β_3 is -5.384e-01, it is the estimate parameter for people who have Bachelor's degree. The β_4 is -1.365e+00, it is the estimate parameter for people who have College/Non-University degree. The β_5 is -9.107e-01, it is the estimate parameter for people who have University Diploma Below Bachelor. The β_6 is -1.316e+00, it is the estimate parameter for people who have Trade Certificate/Diploma. The β_7 is -1.920e+00 , it is the estimate parameter for people who have High School Diploma. The β_8 is -2.956e+00, it is the estimate parameter for people who have education Less than High School.

Table 2 shows the confidence intervals for the parameters in the fitted model with a 95% confidence level. The confidence intervals of intercept B_0 is (0.772313196 , 1.058965615). The confidence intervals for β_1 is (-0.001917382 , 0.001771085) The confidence intervals for β_2 is (-0.656606250, -0.463704761) The confidence intervals for β_3 is (-0.001917382, 0.001771085) The confidence intervals for β_4 is (-1.484154346, -1.246221007) The confidence intervals for β_5 is (-1.089294168, -0.732165439) The confidence intervals for β_6 is (-1.463059292, -1.168887210) The confidence intervals for β_7 is (-2.041271141, -1.798131801) The confidence intervals for β_8 is (-3.110693996, -2.802240417) The respondents have the eduction degree above bachelor when $\beta_3,\beta_4,\beta_5,\beta_6,\beta_7,\beta_8$ all equal to 0.

Table 3 is the ANOVA table for the fitted model The degrees of freedom for age is 1, the p-value is 6.3e-05, it is less than 0.05. The degrees of freedom for parameter "vis_minority" is 1, the p-value is 1.5e-06. The degrees of freedom for parameter "education" is 6, the p-value is very close to 0.

Table 4 is calculating variance inflation factor to investigate correlation among predictors

Resulting Equation:

$$\log (p_i / (1 - p_i)) ,= 0.9156 + (-0.00007315) Age_i + (-0.5602) VisMin_1 + (-0.5384) Education_{i1} + (-1.365) Education_{i2} + (-0.9107) Education_{i3} + (-1.316) Education_{i4} + (-1.92) Education_{i5} + (-2.956) Education_{i6}$$

Discussion

Considering the age's effect on income, the boxplot of age and income respondent does display a certain trend. Initially, for the lowest income group, the distribution of age is quite dispersed. When we come to the groups with higher incomes, we can see that the distribution of age gets more concentrated to the range of [40, 60]. It seems that people under this range of age tend to get highly paid. To reach a plausible conclusion, we will analyze people younger than 40 and people older than 60 respectively. For the age group under 40, most people of this group are making advance in their career. Some may get promoted while others may seek higher salary by switching jobs. Overall, income increases with age within this group. For the age group above 60, although there are still people working and getting highly paid or receiving high pensions, a great number of people's income decreases greatly due to retirement. As a result, the income of this age group on the whole shows a downward pattern. The coefficient of variable "age" is not meaningful. The p-value of it is greater than 0.05, so this coefficient should be zero according to t-test. Ages (full range) and income are not linearly related overall.

Compared with age, the influence of education level on income is significant. Intuitively, we infer that highly educated people are highly paid. This inference can be well demonstrated by both the bar chart of income status and the logistic regression model. First, we will look at the income status bar charts. For "Less than High School" and "High School Diploma" categories, it is obvious that the size of the high-income group dominates that of the low-income group (the number of count of "1" is less than half of that of count "0"). When the level education rises to "Trade Certificate/Diploma", "University Diploma Below Bachelor's", and "College/Non-University" categories, the gap between the two income group gets smaller (the number of count of "1" is less than that of count "0" but more than half of it). For the two highly educated

categories “Bachelor’s” and “Above Bachelor’s”, there is an inverse trend: the majority of people enter the high-income group (the number of count “1” is greater than that of count “0”). Then we go to our logistic regression model. It is easy to see that the coefficients are negative, and the absolute value of them are quite large. These coefficients mean that a lower level education will significantly decrease the probability of earning a high income. Both the model and graphs convey the idea that higher levels of education imply higher probabilities of getting highly paid, and, as we all know, working hard is the major way to get access to higher levels of education. To conclude, our analysis corresponds to the saying that “more pains, more gains” (at student ages), and can be used to encourage students to work hard at school.

Interestingly, our logistic regression model indicates that whether being of visible minority plays a huge role in the chance of earning a high-income. The coefficient of being a visible minority is -0.56015551. However, this coefficient can only reflect the change of log odds if being a visible minority. To more directly illustrate this coefficient, let us consider an example of two young individuals with similar backgrounds. Both of them are 40 years old, and have a bachelor degree. The only difference is that one is a visible minority while the other one is not. We calculate the chance by our logistic regression model. The result is shocking: the one who is a visible minority only has a chance of 39.6% of getting into the rich group whereas the one who is not has a chance of 70.4%, which is almost twice as much as its counterpart. Even though Canada has long been considered a multicultural and friendly country, and discrimination events on ethnicities are seldom reported, it seems that visible minorities do not get as many opportunities as others may remain an unspoken rule nowadays.

Weakness and Next Steps

One weakness of our analysis is that the impact of education on income may be undermined by people’s retirement. For the retired well educated people (Bachelor’s and above Bachelor’s), it is highly possible that they once had an annual income higher than $50,000. However, after their retirement, their income is not as high as it once was or even experiences a significant decrease. If we want a more accurate model of the relationship between education and income, we can limit our sampling population to those of individuals who are still working.

Another weakness noticed is that the size of visible-minorities is much smaller than that of non-visible-minorities. As a result, the coefficients of our logistic regression model is more suited to people who are not visible minorities. One possible future step is that we can do age and education parts separately for the visible-minority group and non-visible-minority group to get more accurate coefficients for both groups. Then we fix all other “X variables” and compare the difference of being a visible minority or not only. In this way, we can see the impact of being a visible minority directly while not influencing the accuracy of other coefficients.

The third weakness is that there may be other influential “X variables” affecting people’s income, such as gender, working hour, and disability. For future improvements, we can elaborate our model by adding new “X variables” into our model and to avoid underfitting.

References

#ggplot2 citation @Book{, author = {Hadley Wickham}, title = {ggplot2: Elegant Graphics for Data Analysis}, publisher = {Springer-Verlag New York}, year = {2016}, isbn = {978-3-319-24277-4}, url = {https://ggplot2.tidyverse.org}, }

tidyverse citation

#> @Article{, #> title = {Welcome to the {tidyverse}}, #> author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D’Agostino McGowan and Romain Fran??ois and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill M??ller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani}, #> year = {2019}, #> journal = {Journal of Open Source Software}, #> volume = {4}, #> number = {43}, #> pages = {1686}, #> doi = {10.21105/joss.01686}, #> }

R citation

#> @Manual{, #> title = {R: A Language and Environment for Statistical Computing}, #> author = {{R Core Team}}, #> organization = {R Foundation for Statistical Computing}, #> address = {Vienna, Austria}, #> year = {2019}, #> url = {https://www.R-project.org/}, #> }

car citation

@Book{, title = {An {R} Companion to Applied Regression}, edition = {Third}, author = {John Fox and Sanford Weisberg}, year = {2019}, publisher = {Sage}, address = {Thousand Oaks {CA}}, url = {https://socialsciences.mcmaster.ca/jfox/Books/Companion/}, }

Knitr citation

@Manual{, title = {knitr: A General-Purpose Package for Dynamic Report Generation in R}, author = {Yihui Xie}, year = {2020}, note = {R package version 1.30}, url = {https://yihui.org/knitr/}, } #tinytex citation @Manual{, title = {tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents}, author = {Yihui Xie}, year = {2020}, note = {R package version 0.26}, url = {https://github.com/yihui/tinytex}, }

#broom citation @Manual{, title = {broom: Broom provides three verbs that each provide different types of information about a model.},

author = {David Robinson}, year = {2020}, note = {R package version 0.26}, url = { https://broom.tidymodels.org/, https://github.com/tidymodels/broom}, }

Wu, Changbao, and Mary E. Thompson. “Basic Concepts in Survey Sampling.” Sampling Theory and Practice. Springer, Cham, 2020. 3-15.

General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User??s Guide Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org. R Core Team (2019b). R: A Language and Environment for Statistical Computing. R Founda- tion for Statistical Computing, Vienna, Austria.

Government of Canada, Statistics Canada. General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User’s Guide. Apr. 2020, www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey.

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.