# Determining the effect of the Income Bracket of a Canadian Citizen by Naturalization on that Individual's Probability of Having Recieved a Bachelor's Degree

Ilke Sun

2020-10-09

**Determining the effect of the Income Bracket of a Canadian Citizen by Naturalization on that Individual's Probability of Having Recieved a Bachelor's Degree**

**Ilke Sun**

**19th of October, 2020**


## Abstract

In this study we aim to determine the relationship between individual's income bracket and the probability of that individual having a Bachelor's degree. The population of interest is Canadian citizens by naturalization who have received their citizenship through the application process. This paper aims to obtain the result that if a Canadian by naturalization is located in higher income brackets, that individual should have a higher chance of having a Bachelor's. Many individuals immigrate to Canada with of receiving a degree and being employed. In addition, educational attainment is expected to result in higher salaries or job opportunities. Assessing for three different logistic regression models with different explanatory variables, we observe that being in the lowest, third highest, second highest and the highest personal income brackets have statistically significant postive effect on probability of an individual having received a Bachelor's degree.


## Introduction

Canada is often referred as land of immigrants because of the number of foreign-born people it welcomes each year. Foreign people come to Canada for various reasons including job opportunities and better education, often times with hopes of becoming a Canadian citizen. Statistics Canada indicate that immigrants make up almost 22% of the Canadian population in 2016 Census and roughly 16% were Canadian by naturalization . Although, people immigrate due to various reasons, about 60% of the immigrants to Canada were admitted under the economic category. Meaning that their prior goal is to improve their welfare by being involved in higher education and/or job opportunities in Canada.

There are various prior studies regarding education attainment and income. In one of the prior studies done in Canada (Paquatte, 1999), researchers observe that post secondary school attainment yields in more employment opportunities and possibly higher income. In this paper we focus on people who have immigrated to Canada and received their citizenship through the application process rather than birth. In addition, we focus on receiving a Bachelor's degree, instead of post secondary school attainment.

In the data set there are total of four variables regarding individual's and his family's income and work habits which are average hours worked, worked last week, income bracket of the family, and income of the respondent. Our main variable of interest is income bracket of the respondent. We aim to find a statistically significant relationship between one's income bracket and the probability of the individual having received

a Bachelor's degree. We expect that people with Bachelor's degree to be in higher income brackets and to assess this probability we use logistic regression.

## Data

In this paper, we use a subset of General Social Survey on Family (2017) data. The GSS 2017 data has 80 variables with over 20000 observations. This data set was built through a dictionary with code written by Rohan Alexander and Sam Caetano with MIT License. This is a large data set with values that are unrelated to the aim of this paper. Hence, a subset of the data was created. The study focuses on Canadian citizens who have received their citizenship by naturalization, meaning that they were born in a foreign country and became Canadian through the application process. We further want to investigate the relationship between economical characteristics regarding these individuals and whether or not they have received a degree equal or higher than Bachelor's Degree.

In the original data set, there are 4 indicators regarding economical characteristics: average hours worked, if an individual has worked last week, income of the family and income of the respondent. These variables are included in the subset. In addition, we aim to control for age and sex, hence, they are also included in the subset. The response variable, receiving Bachelor's Degree, was added to the subset because it was not present in the original data set. In the original data set education was given as a categorical variable with seven different levels, out of this seven categories two indicated that an individual has received a Bachelor's Degree. Thus, our response variable was created as a dummy variable, 1 indicating person has received the degree and 0 indicating otherwise. Overall, the data set has 1748 observations and 7 variables where age is a numeric variable reported to the single decimal point ranging from 15.0 to 80.0, further on, Bachelor's Degree, worked last week and sex are dummy variables, and finally, average hours worked, income of the respondent and income of the family are categorical variables with 5, 6 and 6 categories, respectively.

There are several drawbacks from the GSS 2017 data. There are many missing values and some values that are not sensible. For example, for the question whether or not the respondent worked last week some answers include "Don't know". These values and values that are not available (i.e. N/A) were removed from the data. Although, this is done so that accuracy and sensibility improves, it costs a decrease in the number of observations. Another, drawback was that lack of economic variables in the data set. Only 2 out of 80 variables in the original data set are related to income, and there is some data about occupation but most of these values are unreported or missing. In addition to those 2 variables, only 2 more variables indicate individuals work habits, overall, scarce amount of variables limits the ability to make a more detailed analysis. The third and final drawback is that the most variables in the original data set, also in the subset, are categorical. Working with numeric variables for income, especially continuous variables, might have enhanced our analysis and might have caused us to better understand the relationship between individual's income and the probability that they have received a Bachelor's Degree.

## Model

For the categorical variables we set reference categories to be the lowest income brackets for both personal income and family income, and we set longest working hours (50.1 hours and more) to be the reference category in average hours worked. We expect that working shorter hours and being in the higher income category may increase the chances of having received a Bachelor's degree. Hence, by referencing the most unlikely categories that a person with a Bachelor's degree to be in, we may observe same direction for within each categorical variable.

The X variables below are:

- X1: X-PERSONAL INCOME = $25,000 to $49,999
- X2: X-PERSONAL INCOME = $50,000 to $74,999
- X3: X-PERSONAL INCOME = $75,000 to $99,999
- X4: X-PERSONAL INCOME = $100,000 to $124,999
- X5: X-PERSONAL INCOME = $125,000 and more
- X6: X-FAMILY INCOME = $25,000 to $49,999

- X7: X-FAMILY INCOME = $50,000 to $74,999
- X8: X-FAMILY INCOME = $75,000 to $99,999
- X9: X-FAMILY INCOME = $100,000 to $124,999
- X10: X-FAMILY INCOME = $125,000 and more
- X11: X-AGE
- X12: X-SEX = MALE
- X13: X-WORKED LAST WEEK = YES
- X14: X-AVERAGE HOURS WORKED = 0 hours
- X15: X-AVERAGE HOURS WORKED = 0.1 to 29.9 hours
- X16: X-AVERAGE HOURS WORKED = 30.0 to 40.0 hours
- X17: X-AVERAGE HOURS WORKED = 40.1 to 50.0 hours

In this paper logistic regression is used to analyze the relationship between individual's economical characteristics and the probability that they have received a Bachelor's degree. In the first model we include all of the seven variables in our data set:

log(p-hat/(1-p-hat)) = B-hat0 + B-hat1(X1) + B-hat2(X2) + B-hat3(X3) + B-hat4(X4) + B-hat5(X5) + B-hat6(X6) + B-hat7(X7) + B-hat8(X8) + B-hat9(X9) + B-hat10(X10) + B-hat11(X11) + B-hat5(X12) + B-hat13(X13) + B-hat14(X14) + B-hat15(X15) + B-hat16(X16) + B-hat17(X17)

where:

- p-hat is the estimated probability that the individual received a Bachelor's degree.
- B-hat0 is the intercept estimate
- B-hat1 is the estimated difference in probability between individuals with an personal income less than $25,000 and income from $25,000 to $49,999.
- B-hat2 is the estimated difference in probability between individuals with an personal income less than $25,000 and income from $50,000 to $74,999.
- B-hat3 is the estimated difference in probability between individuals with an personal income less than $25,000 and income from $75,000 to $99,999.
- B-hat4 is the estimated difference in probability between individuals with an personal income less than $25,000 and income from $100,000 to $ 124,999.
- B-hat5 is the estimated difference in probability between individuals with an personal income less than $25,000 and income of $125,000 and more.
- B-hat6 is the estimated difference in probability between individuals with an family income less than $25,000 and income from $25,000 to $49,999.
- B-hat7 is the estimated difference in probability between individuals with an family income less than $25,000 and income from $50,000 to $74,999.
- B-hat8 is the estimated difference in probability between individuals with an family income less than $25,000 and income from $75,000 to $99,999.
- B-hat9 is the estimated difference in probability between individuals with an family income less than $25,000 and income from $100,000 to $124,999.
- B-hat10 is the estimated difference in probability between individuals with an family income less than $25,000 and income of $125,000 and more.
- B-hat11 is the estimated effect of age.
- B-hat12 is the estimated difference in probability between a female and a male.
- B-hat13 is the estimated effect of not working last week compared to working.
- B-hat14 is the estimated difference in probability between individuals who work, on average, over 50.1 hours and 0 hour, on average.
- B-hat15 is the estimated difference in probability between individuals who work, on average, over 50.1 hours and 0.1 to 29.9 hours, on average.
- B-hat16 is the estimated difference in probability between individuals who work, on average, over 50.1 hours and 30.0 to 40.0 hours, on average.
- B-hat17 is the estimated difference in probability between individuals who work, on average, over 50.1 hours and 40.1 to 50.0 hours , on average.

In the second model, variables with high p-values are removed. These removed variables are sex, average hours worked and income of the family. Hence, the second model is:

log(p-hat/(1-p-hat)) = B-hat0 + B-hat1(X1) + B-hat2(X2) + B-hat3(X3) + B-hat4(X4) + B-hat5(X5) + B-hat11(X11)

where:

- p-hat is the estimated probability that the individual received a Bachelor's degree.
- B-hat0 is the intercept estimate
- B-hat1 is the estimated difference in probability between individuals with an income less than $25,000 and income from $25,000 to $49,999.
- B-hat2 is the estimated difference in probability between individuals with an income less than $25,000 and income from $50,000 to $74,999.
- B-hat3 is the estimated difference in probability between individuals with an income less than $25,000 and income from $75,000 to $99,999.
- B-hat4 is the estimated difference in probability between individuals with an income less than $25,000 and income from $100,000 to $124,999.
- B-hat5 is the estimated difference in probability between individuals with an income less than $25,000 and income of $125,000 and more.
- B-hat11 is the estimated effect of age.

The third and the final model is the relationship between, our main interest variable, the personal income bracket and probability of having received a Bachelor's degree. This is included so that we can observe the direct relationship between our explanatory and response variables to assess the fit of the log function. Hence, the third model is:

log(p-hat/(1-p-hat)) = B-hat0 + B-hat1(X1) + B-hat2(X2) + B-hat3(X3) + B-hat4(X4) + B-hat5(X5)

- p-hat is the estimated probability that the individual received a Bachelor's degree.
- B-hat0 is the intercept estimate
- B-hat1 is the estimated difference in probability between individuals with an income less than $25,000 and income from $25,000 to $49,999.
- B-hat2 is the estimated difference in probability between individuals with an income less than $25,000 and income from $50,000 to $74,999.
- B-hat3 is the estimated difference in probability between individuals with an income less than $25,000 and income from $75,000 to $99,999.
- B-hat4 is the estimated difference in probability between individuals with an income less than $25,000 and income from $100,000 to $124,999.
- B-hat5 is the estimated difference in probability between individuals with an income less than $25,000 and income of $125,000 and more.

## Results

Figure 1

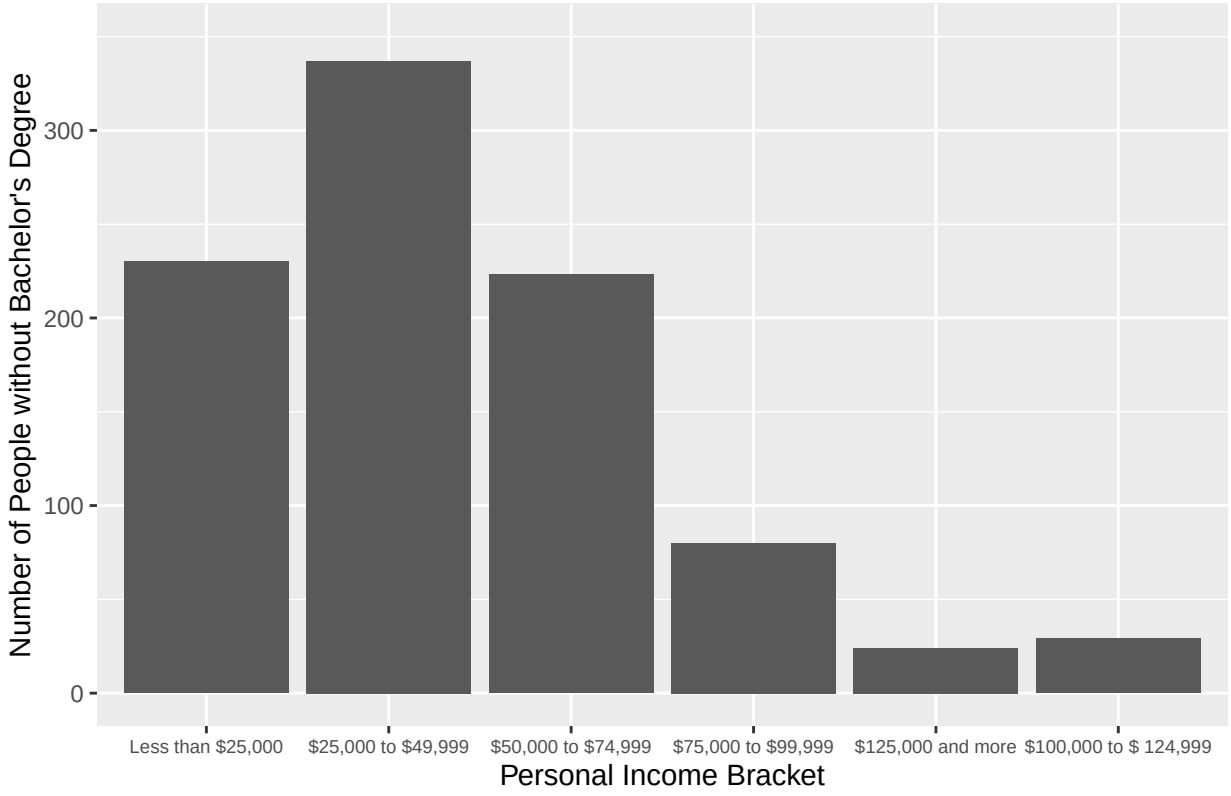
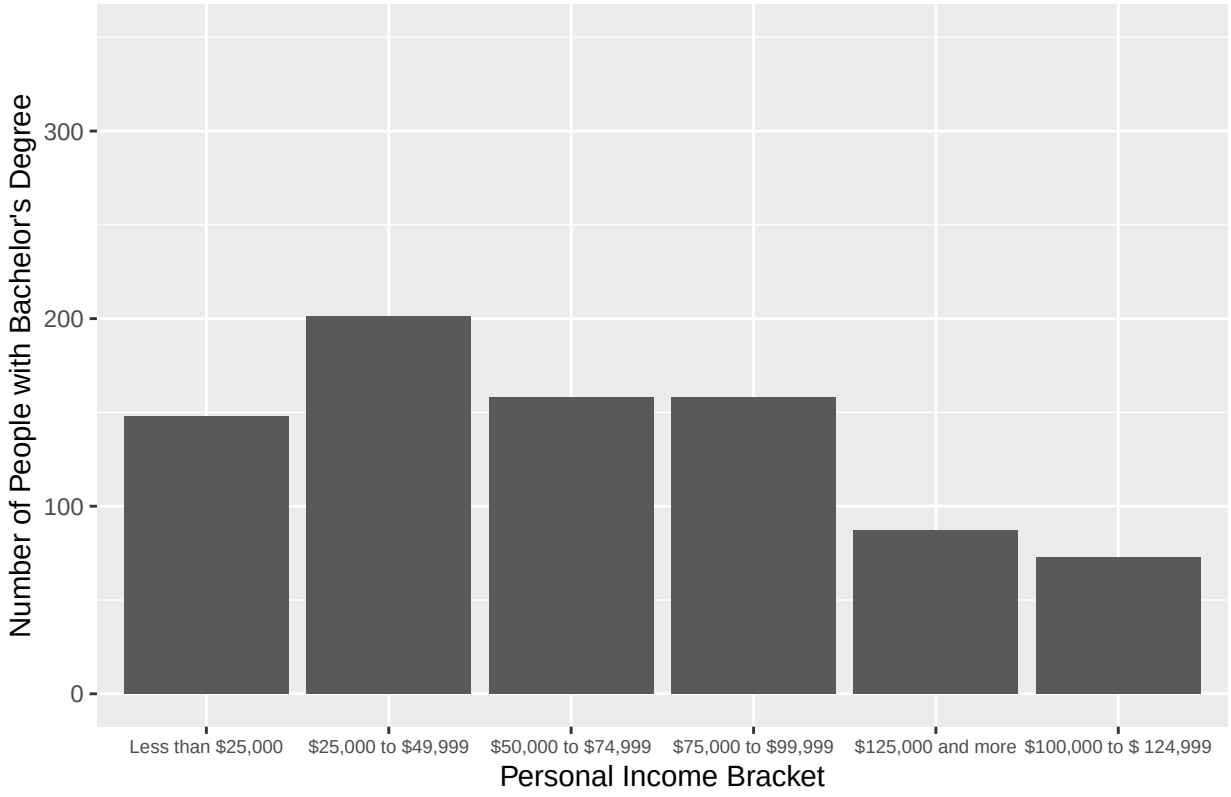



Figure 2



As mentioned we have a total of 1748 observations in our sample which are the individuals who are Canadian citizens by naturalization. Out of these 1748 individuals, 825 have a Bachelor's degree and 923 don't. From Figure 1 we can observe that there are relatively more people in lowest three income brackets compared to highest three, approximately 85% of immigrants without a Bachelor's degree are in the lower half of income brackets. Moreover, 61% of these individuals make less than $50,000 per year, while only 2.6% earn more than $125,000. In Figure 2, we observe a more distribution closer to uniform. Although, there are still more people in the lower half of income, approximately 39% is located in the higher half of the brackets. In addition, only 42% of the individuals with Bachelor's degree earn less than $50,000 a year compared to 61% of the individual's without a Bachelor's degree. We also observe that, over 10% of individuals with Bachelor's degree earn more than $125,000 a year compared to 2.1% from Figure 1.

Overall, we see that there might be significant effect of having received a Bachelor's degree on which income bracket are in. However, our main goal is to analyze how income bracket of a Canadian citizen by naturalization determine whether they have received a Bachelor's degree or a higher degree of education (e.g. Phd). To determine this effect we have modeled three different logistic regressions. Results for the first model is:

```
##
## Call:
## glm(formula = bachelors_degree ~ ., family = "binomial", data = immigrants)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8685  -1.0230  -0.7882   1.2307   1.7995
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -0.216342   0.354005  -0.611  0.54112
## age                                 -0.013355   0.003926  -3.402  0.00067
## sexMale                             -0.062810   0.104586  -0.601  0.54813
## average_hours_worked0 hour          14.154658 368.836267   0.038  0.96939
## average_hours_worked0.1 to 29.9 hours  0.160404   0.242968   0.660  0.50913
## average_hours_worked30.0 to 40.0 hours  0.159805   0.213596   0.748  0.45436
## average_hours_worked40.1 to 50.0 hours -0.330578   0.255875  -1.292  0.19638
## worked_last_weekYes                    0.246265   0.142629   1.727  0.08424
## income_respondent$100,000 to $ 124,999  1.341290   0.278991   4.808 1.53e-06
## income_respondent$125,000 and more      1.830924   0.300023   6.103 1.04e-09
## income_respondent$25,000 to $49,999     0.000176   0.164496   0.001  0.99915
## income_respondent$50,000 to $74,999     0.034121   0.181186   0.188  0.85063
## income_respondent$75,000 to $99,999     1.121812   0.208739   5.374 7.69e-08
## income_family$100,000 to $ 124,999      0.261505   0.264854   0.987  0.32347
## income_family$125,000 and more          0.217711   0.251068   0.867  0.38587
## income_family$25,000 to $49,999        -0.098894   0.254336  -0.389  0.69740
## income_family$50,000 to $74,999         0.155238   0.251510   0.617  0.53709
## income_family$75,000 to $99,999         0.094761   0.256427   0.370  0.71172
##
## (Intercept)
## age                                  ***
## sexMale
## average_hours_worked0 hour
## average_hours_worked0.1 to 29.9 hours
## average_hours_worked30.0 to 40.0 hours
## average_hours_worked40.1 to 50.0 hours
## worked_last_weekYes                   .
## income_respondent$100,000 to $ 124,999 ***
```

```
## income_respondent$125,000 and more     ***
## income_respondent$25,000 to $49,999
## income_respondent$50,000 to $74,999
## income_respondent$75,000 to $99,999    ***
## income_family$100,000 to $ 124,999
## income_family$125,000 and more
## income_family$25,000 to $49,999
## income_family$50,000 to $74,999
## income_family$75,000 to $99,999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2417.7  on 1747  degrees of freedom
## Residual deviance: 2240.4  on 1730  degrees of freedom
## AIC: 2276.4
##
## Number of Fisher Scoring iterations: 12
```

Here we observe that age and belonging to the top three income brackets compared to the lowest income bracket have statistically significant effects on determining the probability of the individual having a Bachelor's degree. However, this is not the main take away from this model. When we consider all our variables in the data set, some economical variables fail to predict the probability of an individual having received a Bachelor's degree. We have previously predicted that working less hours may help predict the probability, but this was not the case. All levels in average hours worked have a high standard error and p-value, especially for X-AVERAGE HOURS WORKED = 0 hours. The standard error associated with the variable is 368.8 which is drastically high. This may be due to various reasons, but it suggest that people who work 0 hours have various different characteristic that must be analyzed in order to understand its influence on probability of the individual having received a Bachelor's degree, or maybe average hours worked have no explanatory value for the probability that we are interested in.

Second important take away from the first model is that, although some personal income levels have statistically significant effect on the probability, none of the family income brackets have that significant effect. They all have high p-values, hence, we cannot say that individual's family income determines the likelihood of that individual having received a Bachelor's degree. Although, this is understandable, we also expected that there might be significant relationship between family's income and the probability. We expected that, like the personal income, higher income brackets for the family may increase the changes of the individual belonging to the group with Bachelor's degree, however, this was not the case. Working last week was significant under the 10% interval, however, we saw that relationship is lost when we drop the other insignificant variables. Due to these reasons, we have removed variables average hours worked, worked last week and income of the family which yielded in the second model.

```
##
## Call:
## glm(formula = bachelors_degree ~ age + income_respondent, family = "binomial",
##     data = immigrants)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8651  -1.0160  -0.8593   1.2621   1.5974
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         0.19486    0.19481   1.000 0.317198
```

```
## age                                    -0.01483    0.00385  -3.852 0.000117 ***
## income_respondent$100,000 to $ 124,999  1.50357    0.24741   6.077 1.22e-09 ***
## income_respondent$125,000 and more      1.88365    0.25813   7.297 2.94e-13 ***
## income_respondent$25,000 to $49,999     0.01683    0.14096   0.119 0.904984
## income_respondent$50,000 to $74,999     0.17220    0.15024   1.146 0.251718
## income_respondent$75,000 to $99,999     1.22454    0.17634   6.944 3.80e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2417.7  on 1747  degrees of freedom
## Residual deviance: 2260.8  on 1741  degrees of freedom
## AIC: 2274.8
##
## Number of Fisher Scoring iterations: 4
```

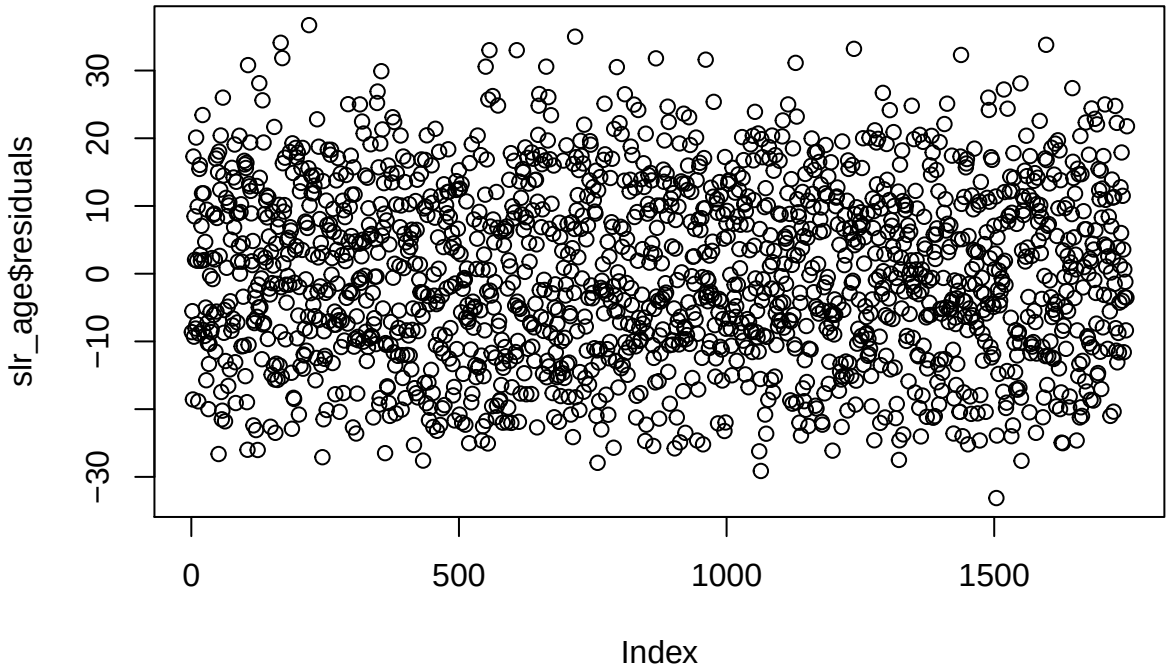
After removing said variables, we observe the logarithmic regression results above. We observe that our intercept is still insignificant, however, we expect that it may be due the variable age. Age here, like the previous regression, is statistically significant with a p-value of 0.000117. Although, it is a low enough p-value, we may need to make further consideration. We must consider that finding a statistically significant intercept would help us get the probability of having received Bachelor's degree, if the respondent belongs to the lowest income group, less than $25,000. We know that there are a total of 6 explanatory variables in this regression: one of them is age and other five are the personal income categories. If a person does not belong to any of those five categories (i.e. if the individual belongs to the less than $25,000 income bracket), those five variables would be 0 so our model becomes:

log(p-hat/(1-p-hat)) = B-hat0 + B-hat1(0) + B-hat2(0) + B-hat3(0) + B-hat4(0) + B-hat5(0) + B-hat11(X11) = 0.19486 - 0.01483(X11)

Hence, when a person belongs to the lowest income bracket, most explanatory power falls in age, which may be causing intercept to be insignificant. Further on, when we use logistic regression on our desired probability by only using X11 (i.e. age variable), we did not observe as statistically significant p-value as the one from the second model. The p-value obtained from the second model (0.000117) falls to 0.0262, although, it is still statistically significant in 5% level. Here we also need to consider implications of removing age from our regression. Age here may become a confounder if removed since we expect that someone who is 30 years of age to be more probable to have a Bachelor's degree compared to a 15 year old. Moreover, age may also effect the income bracket, by the same argument, someone who is 30 years old might have higher chance of being in a higher income bracket compared to a 15 year old.

```
##
## Call:
## lm(formula = age ~ income_respondent, data = immigrants)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.127  -9.927  -0.627  10.073  36.702
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            43.2976     0.6726  64.373  < 2e-16 ***
## income_respondent$100,000 to $ 124,999  8.6347     1.4591   5.918 3.92e-09 ***
## income_respondent$125,000 and more      9.4474     1.4117   6.692 2.96e-11 ***
## income_respondent$25,000 to $49,999     6.1295     0.8776   6.984 4.07e-12 ***
## income_respondent$50,000 to $74,999     4.8714     0.9493   5.131 3.20e-07 ***
## income_respondent$75,000 to $99,999     6.1940     1.0821   5.724 1.22e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.08 on 1742 degrees of freedom
## Multiple R-squared:  0.0449, Adjusted R-squared:  0.04216
## F-statistic: 16.38 on 5 and 1742 DF,  p-value: 7.983e-16
```




When we do a linear regression on age with personal income brackets, we see that all income categories and the intercept have really small p-values making all of them statistically significant in all confidence intervals. We also observe from the plot of residuals that errors seem uncorrelated and fairly symmetric. Taking these into consideration, we may want to keep age variable in our regression to make our estimations more accurate because it has a significant relationship with having a Bachelor's degree and it may become a confounder if it is taken away from the regression. However, we cannot certainly say that age is a confounder by only considering the data above. Overall, age has a quite small and negative effect size of -0.01483 and a statistically significant effect on probability of an individual having received a Bachelor's degree.

For the third model, we have regressed only the personal income brackets variable on log probability of receiving a Bachelor's degree which yields in:

```
##
## Call:
## glm(formula = bachelors_degree ~ income_respondent, family = "binomial",
##     data = immigrants)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7501  -0.9968  -0.9672   1.3268   1.4033
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -0.44087    0.10538  -4.184 2.87e-05 ***
## income_respondent$100,000 to $ 124,999  1.36403    0.24349   5.602 2.12e-08 ***
## income_respondent$125,000 and more      1.72872    0.25351   6.819 9.15e-12 ***
## income_respondent$25,000 to $49,999    -0.07591    0.13801  -0.550    0.582
```

```
## income_respondent$50,000 to $74,999     0.09629    0.14805   0.650    0.515
## income_respondent$75,000 to $99,999     1.12144    0.17301   6.482 9.06e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2417.7  on 1747  degrees of freedom
## Residual deviance: 2275.8  on 1742  degrees of freedom
## AIC: 2287.8
##
## Number of Fisher Scoring iterations: 4
```

Here we observe that the intercept, X3: X-PERSONAL INCOME = $75,000 to $99,999, X4: X-PERSONAL INCOME = $100,000 to $124,999, and X5: X-PERSONAL INCOME = $125,000 and more are statistically significant in all confidence interval. Furthermore, we have large effect sizes for all significant income categories(B-hat3, B-hat4, B-hat5 $> 1$). Now, we can more confidently talk about effect sizes. Intercept suggest that if an individual earns less than $25,000, the log probability of receiving a Bachelor's degree is -0.44087 which is approximately 39%. If an individual earns between $75,000 to $99,999, the log probability of that person having a Bachelor's degree is (-0.44087 + 1.12144) 0.68057, and the actual probability of having a Bachelor's degree is approximately 66%. We observe that as we go higher in the income brackets, the effect size enlargers in a positive direction. This suggests that as we go higher in income bracket, the probability of having a Bachelor's degree increases. We have predicted this at the start of the study, it is sensible that the probability is higher in the higher income categories even though X1: X-PERSONAL INCOME = $25,000 to $49,999 and X2: X-PERSONAL INCOME = $50,000 to $74,999 are not statistically significant. Seen from Figure 1 & 2, most people regardless of having Bachelor's degree or not are in these income brackets. Thus, we may need additional controls or numeric personal income data for these variables to also be significant.

## Discussion

Models 1, 2 and 3 all can be useful in terms of understanding relationship between the economical variables and log probability of having a Bachelor's degree. However, out of these 3 we will consider only 2 and 3 in this discussion because the main goal of this study is to determine the effect of the income bracket of an individual on the probability of them having a Bachelor's degree. Model 1 is not going to be assessed in this section because most parameters fail to have significant effects on the response variable. Out of these economic variables, we have seen that some personal income categories and age have significant effects on the response variable. Another, reason why we have neglected the Model 1 is the abundance of categorical variables, besides age all variables are categorical which may cause us to fail to obtain enough variation in our predictors which can be used to predict the variation in our response variable. Hence, we have created models 2 and 3.

Model 2: We assume that this model is different in terms of control compared to the third model. The removed variables in the previous model, is unlikely to serve as control, since they are unlikely to affect both the probability of receiving a Bachelor's degree and belonging into a certain income bracket. Model 2, is valuable because it shows that even with strong predictors of probability of receiving a Bachelor's degree (i.e. personal income bracket) age still has a statistically significant effect. Although, in the Results section we have considered that age might be a confounder, we do not certainly know that it does. Model 2 would be really valuable if we had income, and other economical variables as a continuous or at least numeric. Considering income as brackets really limits the variation in the predictors, hence, limits the amount of variation that can utilized to predict the variation in the response variable. Thus, if we were to have other economical variables (e.g. consumption, tax payments), income as a continuous or a numeric variable instead of categorical, and if we were sure that age is a confounder when taken out of the regression then using this model would be more beneficial and accurate.

Model 3: This model is the desired regression for our study. We were aiming to determine the effect of the income bracket of a Canadian citizen by naturalization on that individual's probability of having received a

Bachelor's degree. Here our response variable is the log probability of having a Bachelor's degree and we have 5 different income variables as our explanatory variables and the lowest income bracket as our reference category. As said for Model 2, if we were to have a numeric value as income, this regression may have been even more accurate.

From the results we have, we conclude that we have statistically significant results for 4 out of 6 categories of income: the lowest (Less than $25,000), fourth ($75,000 to $99,999), second highest ($100,000 to $ 124,999) and the highest ($125,000 and more) income brackets. For the lowest income bracket we have log probability of -0.44087 which suggest that p-hat equals to approximately 39%. This suggest that if a Canadian Citizen by naturalization earns less than $25,000, there is 39% chance that individual has a Bachelor's degree and 61% chance that he does not. For the fourth income category, we can say that the log probability is 0.68057 and the actual probability of that indivudal having a Bachelor's degree is approximately 66%. For the fifth and the second highest income category, if an individual earns between $100,000 to $ 124,999 we expect the log probability to be 0.92316 and the actual probability of having a Bachelor's degre to be approximitely 72%. Lastly, if a Canadian by naturalization earns more than $125,000, we expect that individual's log probability of having a Bachelor's degree to be 1.28785 and the actual probability to be approximately 78%. We see that when we compare the lowest and highest income brackets, the probability of an individual having a Bachelor's degree approximately doubles. This does not entail that someone who receives a Bachelor's degree is more likely to be in higher brackets but it means that if a person belongs to the highest income bracket it is almost two times more likely that person has a Bachelor's degree.

## Weaknesses

There are three main weaknesses to this study: one is too many categorical variables as our variables of interest, second is our data set is not designed specifically for this type of data analysis and third is unable to produce logistic regression graph with a categorical variable in R. Firstly, numeric and/or continuous variables might have been more helpful in analyzing the relationship between income and probability of whether a person has a Bachelor's degree. Second, the data used is GSS 2017 Family data, this data is not designed for analysation of economical variables. Rather the focus was family, maritial and household variables. As said in the introduction there are only 4 variables that are related to income, economics and working habits of individuals. More variables would have helped us with more control and understanding, other than income bracket, how economical vairables prediict the probability of an individual to receive a Bachelor's degree. Third, weakness is related to my preficency with R. The models we have picked are logistic regression and we are yet unable to plot these graphs in R, unlike SLR models.

## Next Steps

Firstly, the variable age should be modeled with income and receiving a Bachelor's degree on a larger sample so that its effects are determined. Moreover, we can understand if it confounds income brackets and probability of having a Bachelor's degree, if it confounds then Model 2, and if it does not Model 3 should be used. In addition, we enlarge our sample and solely focus on Canadian's by naturalization, unlike GSS 2017 which also includes people who are citizens by birth or not citizens. We can also consider, naturalized Americans for this sample so that we can enlarge our population to the immigrants that became citizens in North America, or even naturalized citizens in general. Second, we should try to obtain numeric data and re-do the study. As stressed from the start of the study, categorical variables limit our approach to studying these variables and data presented in categorical variables can easyly be recorded as numeric, if a person volunteers to give out his information. Moreover, this numeric data is likely already obtained by Canadian Goverment because people report their income to the government for tax and using that data instead of categorical income brackets may explain data more accurately than this study already did.

repo link: https://ilkes-sta304-website.netlify.app/1/01/01/determining-the-effect-of-the-income-bracket-of-a-canadian-citizen-by-naturalization-on-that-individuals-probability-of-having-recieved-a-bachelors-degree/

## References

"Obtaining Canadian Citizenship." Statistics Canada: Canada's National Statistical Agency / Statistique Canada : Organisme Statistique National Du Canada, 25 July 2018, www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-010-x/99-010-x2011003_1-eng.cfm.

Government of Canada, Statistics Canada. "Immigration and Ethnocultural Diversity." Government of Canada, Statistics Canada, 19 Oct. 2020, www150.statcan.gc.ca/n1/en/subjects/immigration_and_ethnocultural_diversity.

Paquette, Jerry. "Educational Attainment and Employment Income: Incentives and Disincentives for Staying in School." CANADIAN JOURNAL OF EDUCATION, vol. 2, no. 24, 1999, pp. 151–168.