Family Income Analysis Using Logistic Model

Abstract I

Family income is the most important factor reflecting the living standard of the people. However, income gap issue has become a serious topic for a nation. If the income gap is too large, it will lead to social inequality. In this report, we will discuss how the family income level is affected by age, gender, education level, province, working time and marital status. A dataset of GSS in 2017 was obtained and we have picked out some of the variables interested in. Then, binary logistic model is applied to classify which family have high income. The results reveal that family income levels are strongly correlated with whether they have completed bachelor's education, whether they live in Newfoundland and Labrador province and whether they got married and how many hours they worked per week.

Introduction II

Our goal is to evaluate how the interested factors impact the classification of family income level. A dataset is obtained from 2017 GSS and we further narrow down our scope to the family income and some interested attributes. In this report, we first set a threshold to separate the high-income family from the collected income data, which is 100,000 dollars. Then, a binary logistic model is constructed to fit. Binary logistic model has a dependent variable with two possible values ("High", "Regular" in our report), and a series of predictors included in the model to estimate the coefficients. The predictors that will be included in our following model are: a) age b) gender c) education level d) province e) working hours f) marital status

Through fitting the model, we can get the estimations of the coefficients and some statistical results to explain how these attributes affect the income level. In the end, we will also make a discussion and advise next step. All the code can be found here:https://github.com/suuuusieyeah/Family-income-analysis/blob/main/ps3%20remastered.Rmd

Data III

```
data<-read.csv("gss(5).csv")
##str(data)
```

We obtained the dataset from 2017 GSS, the dataset contains variables of different features of random selected families. The target population is 20602, the sampling method is good and investigate a number of 81 variables. However the dataset was not clean, it contains a lot missing value of each variables, another drawback is the data was obtained within Canada, so it could be a little narrotive in some degree.

Model IV

```
data$income_family.type<-as.factor(ifelse(data$income_family>100000,"High","Regular"))
data$income_family.type<- relevel(data$income_family.type, ref = "Regular")
N=100000
```

```r
n=length(data$income_family)
fpc.srs = rep(N, n)
example.design.srs <- svydesign(id=~1, data=data, fpc=fpc.srs)
svyglm.srs.logit <- svyglm(income_family.type ~ age+sex+education+province+average_hours_worked+marital_
summary(svyglm.srs.logit)
```

```
##
## Call:
## svyglm(formula = income_family.type ~ age + sex + education +
##     province + average_hours_worked + marital_status, design = example.design.srs,
##     family = "binomial")
##
## Survey design:
## svydesign(id = ~1, data = data, fpc = fpc.srs)
##
## Coefficients:
##                                                                    Estimate
## (Intercept)                                                      -0.8860603
## age                                                               0.0004991
## sexMale                                                           0.0586967
## educationCollege, CEGEP or other non-university certificate or di...  0.2040072
## educationHigh school diploma or a high school equivalency certificate  0.4918145
## educationLess than high school diploma or its equivalent          0.4557813
## educationTrade certificate or diploma                             0.4441319
## educationUniversity certificate or diploma below the bachelor's level -0.0486854
## educationUniversity certificate, diploma or degree above the bach...  -0.0119664
## provinceBritish Columbia                                          0.4634191
## provinceManitoba                                                  0.1102086
## provinceNew Brunswick                                             0.3699340
## provinceNewfoundland and Labrador                                 0.1486627
## provinceNova Scotia                                               0.1895264
## provinceOntario                                                   0.2367836
## provincePrince Edward Island                                      0.3006137
## provinceQuebec                                                    0.3999789
## provinceSaskatchewan                                              0.3854202
## average_hours_worked0.1 to 29.9 hours                            -0.8202073
## average_hours_worked30.0 to 40.0 hours                           -1.7951072
## average_hours_worked40.1 to 50.0 hours                           -1.7816753
## average_hours_worked50.1 hours and more                          -1.3129225
## average_hours_workedDon't know                                   -0.8279909
## marital_statusLiving common-law                                  -1.5591224
## marital_statusMarried                                            -2.3055691
## marital_statusSeparated                                           0.2703518
## marital_statusSingle, never married                               0.2219051
## marital_statusWidowed                                            -0.1128063
##                                                                  Std. Error
## (Intercept)                                                       0.5540671
## age                                                               0.0023995
## sexMale                                                           0.0645367
## educationCollege, CEGEP or other non-university certificate or di...  0.0988856
## educationHigh school diploma or a high school equivalency certificate  0.0954232
## educationLess than high school diploma or its equivalent          0.1195074
## educationTrade certificate or diploma                             0.1317654
## educationUniversity certificate or diploma below the bachelor's level  0.2028959
```

```
## educationUniversity certificate, diploma or degree above the bach...   0.1360246
## provinceBritish Columbia                                               0.1408646
## provinceManitoba                                                       0.1756160
## provinceNew Brunswick                                                  0.1664524
## provinceNewfoundland and Labrador                                      0.1882918
## provinceNova Scotia                                                    0.1653608
## provinceOntario                                                        0.1283590
## provincePrince Edward Island                                           0.2026773
## provinceQuebec                                                         0.1325923
## provinceSaskatchewan                                                   0.1679395
## average_hours_worked0.1 to 29.9 hours                                  0.5254221
## average_hours_worked30.0 to 40.0 hours                                 0.5231078
## average_hours_worked40.1 to 50.0 hours                                 0.5311569
## average_hours_worked50.1 hours and more                                0.5337570
## average_hours_workedDon't know                                         0.5411712
## marital_statusLiving common-law                                        0.1496786
## marital_statusMarried                                                  0.1218302
## marital_statusSeparated                                                0.1411592
## marital_statusSingle, never married                                    0.1049061
## marital_statusWidowed                                                  0.1735617
##                                                                       t value
## (Intercept)                                                            -1.599
## age                                                                     0.208
## sexMale                                                                 0.910
## educationCollege, CEGEP or other non-university certificate or di...    2.063
## educationHigh school diploma or a high school equivalency certificate   5.154
## educationLess than high school diploma or its equivalent                3.814
## educationTrade certificate or diploma                                   3.371
## educationUniversity certificate or diploma below the bachelor's level  -0.240
## educationUniversity certificate, diploma or degree above the bach...   -0.088
## provinceBritish Columbia                                                3.290
## provinceManitoba                                                        0.628
## provinceNew Brunswick                                                   2.222
## provinceNewfoundland and Labrador                                       0.790
## provinceNova Scotia                                                     1.146
## provinceOntario                                                         1.845
## provincePrince Edward Island                                            1.483
## provinceQuebec                                                          3.017
## provinceSaskatchewan                                                    2.295
## average_hours_worked0.1 to 29.9 hours                                  -1.561
## average_hours_worked30.0 to 40.0 hours                                 -3.432
## average_hours_worked40.1 to 50.0 hours                                 -3.354
## average_hours_worked50.1 hours and more                                -2.460
## average_hours_workedDon't know                                         -1.530
## marital_statusLiving common-law                                       -10.416
## marital_statusMarried                                                 -18.924
## marital_statusSeparated                                                 1.915
## marital_statusSingle, never married                                     2.115
## marital_statusWidowed                                                  -0.650
##                                                                      Pr(>|t|)
## (Intercept)                                                          0.109802
## age                                                                  0.835240
## sexMale                                                              0.363098
## educationCollege, CEGEP or other non-university certificate or di...  0.039126
```

```
## educationHigh school diploma or a high school equivalency certificate 2.59e-07
## educationLess than high school diploma or its equivalent              0.000137
## educationTrade certificate or diploma                                 0.000752
## educationUniversity certificate or diploma below the bachelor's level 0.810371
## educationUniversity certificate, diploma or degree above the bach...  0.929900
## provinceBritish Columbia                                              0.001005
## provinceManitoba                                                      0.530307
## provinceNew Brunswick                                                 0.026269
## provinceNewfoundland and Labrador                                     0.429814
## provinceNova Scotia                                                   0.251759
## provinceOntario                                                       0.065104
## provincePrince Edward Island                                          0.138042
## provinceQuebec                                                        0.002561
## provinceSaskatchewan                                                  0.021749
## average_hours_worked0.1 to 29.9 hours                                 0.118537
## average_hours_worked30.0 to 40.0 hours                                0.000602
## average_hours_worked40.1 to 50.0 hours                                0.000798
## average_hours_worked50.1 hours and more                               0.013915
## average_hours_workedDon't know                                        0.126041
## marital_statusLiving common-law                                        < 2e-16
## marital_statusMarried                                                  < 2e-16
## marital_statusSeparated                                               0.055485
## marital_statusSingle, never married                                   0.034425
## marital_statusWidowed                                                 0.515736
##
## (Intercept)
## age
## sexMale
## 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...
## 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                               *
## average_hours_workedDon't know
## marital_statusLiving common-law                                       ***
## marital_statusMarried                                                 ***
## marital_statusSeparated                                               .
## marital_statusSingle, never married                                   *
## marital_statusWidowed
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.974077)
##
## Number of Fisher Scoring iterations: 6
```

1.We then continue to analysis by building a logistic regression model to predict the 'Family Income' using variables 'age', 'sex', 'education', 'province', 'average working hours' and 'marital status'. A logistic model is better to investigate the categorical data which can give a good inference about the importance of each feature. The reason we choose a logistic model is because we believe the variables we focused on most are collected as categorical data such as income level, education and marital status etc while only a few numerical variables like number of marriages are less important for our investigation. Using the data selected we build a logistic model with the following formula: $log(p/1-p) = \beta_0 + \beta_1 age + \beta_2 * sexMale + \beta_3 * educationCollege + ... + \beta_k * marital_statuswidowed$ when sex is male, sexMale =1 otherwise sexMale =0. By checking the p-value, we notice that people who educated at high school or high degree, average working hours more than 30 hours and married people have larger p-value, which means those variables are more significant.

```
vif(svyglm.srs.logit)
```

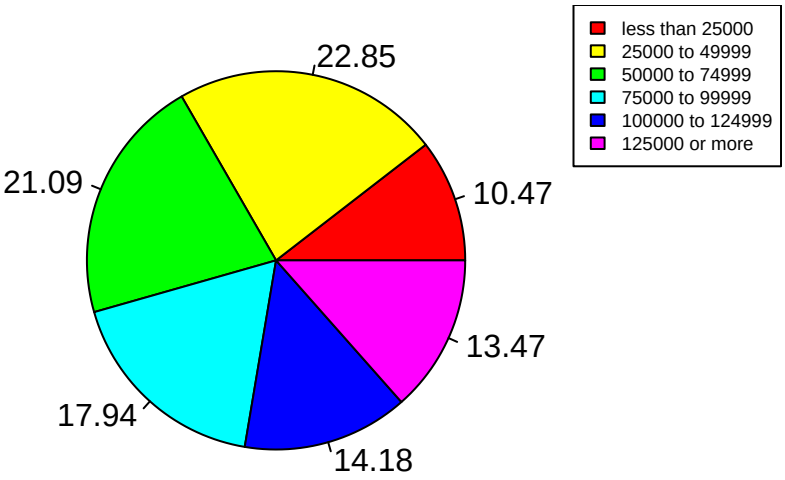
```
##                           GVIF Df GVIF^(1/(2*Df))
## age                   1.775572  1        1.332506
## sex                   1.160037  1        1.077050
## education             1.432503  6        1.030405
## province              1.283471  9        1.013961
## average_hours_worked 1.323713  5        1.028441
## marital_status        1.978995  5        1.070642
```

```
summary(svyglm.srs.logit)$adj.r.squared
```

```
## NULL
```

2.We checked the multicollinearity of the model by compute its VIF. We find that none of the value is greater than 5, which means there is no multicollinearity existing.

```
income<-data$income_family
income<-filter(data,income_family=='income')
x1<-as.numeric(table(data$income_family))
piepercent<-(paste(round(100*x1/sum(x1),2)))
pie(x1, labels = piepercent, main = , col = rainbow(length(x1)))
legend("topright", c('less than 25000','25000 to 49999', '50000 to 74999', '75000 to 99999', '100000 to
```



3.From the pie chart above, we can see almost half of income data is below 75,000 dollars. Only 9% have

income less than 25,000 dollars. 25,000-49,999 and 50,000-74,999 have almost same percentage. For the high income groups, the proportions of 75,000-99,999, 100,000-124,999 and 125,000 or more groups are 17.15%, 16.77% and 13.85% respectively. We can realize that the Canadian family income level in 2017 is quite even.

Result V

All the independent variable used are categorical. From the summary of the logistic model, the p-values of *educationUniversity certificate or diploma below the bachelor's level*, *provinceNewfoundland and Labrador*, *marital_statusMarried*, *average_hours_worked30.0 to 40.0 hours* and *intercept* is less than 0.05, which means these factors are significant in our model(significant level 0.05) and should be included in the model. The two factors of whether they obtained bachelor's diploma and whether they live in Newfoundland have large coefficients are both larger than 15. In other words, the log odds will less 15 if the family member obtained bachelor's diploma or if the family live in Newfoundland and Labrador when other conditions remain unchanged. Stronger impact of these two factors on the respond than that of the other two.

For the further step, we can consider to remove the "bad" variables and keep the good ones, then fit anothor model to get better estimated coefficients. Not only that, but we can add some potential variables to mine more information.

References VI

https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm