Analysis for Influencial Factors of Canadian Divorce Rate

Yuxuan Liu, Yuxuan Lin, Yangle Shang, Kexin Zhang

10/18/2020

Code and data supporting this analysis is available at: https://github.com/KexinZhang-Claire/STA304-PS2 (https://github.com/KexinZhang-Claire/STA304-PS2)

Abstract

Our study is a preliminary study focusing on the effect on the divorce rate when factoring income, age, children, education and health. By building generalized linear model, we found that it is less likely for younger couples with higher income, less children, lower education to get divorced. We care about divorce since it will negatively impacts children's mental health.

Introduction

According to some articles, divorce rates in Canada are currently in a decline pattern. At the beginning of the 21st century, the divorce rate was around 10 out of every 1000 couples, and by 2016, the proportion had dropped to 6 per 1000 couples. This may due to the fact that not as many people are getting married. Unlike the baby boomers who got married when they were young, millennials choose to marry after completing education, establishing a career and having a good financial situation.

In this analysis report, we would like to find out what factors could influence a person's decision on divorce and how divorce can be explained by taking these factors into consideration.

We raise several potential predictor variables that may have have an impact on whether people divorce, which including current age, total children number, income, education, life satisfaction level, health, rural area or urban area, etc.

The important part of this work is to determine the probability of getting divorced by building a logistic model. Therefore, we can use this model to predict the likelihood of divorce for a person. Then we would like to discuss the possible reason behind why and how these factors influence. If someone is predicted to have a high probability of getting divorced, we can try to avoid it by contacting him and providing psychological counseling.

Data

We obtained the dataset from gss2017. We downloaded the CSV data file and changed the raw data name by the labels and dictionary of gss2017.

We utilized this dataset since it is the most updated version. However, a limitation is that it has been 3 years since released, so things would change a lot. Moreover, there are 81 variables and 20602 observation in this dataset, which almost cover everything we want to know. The variables can be expressed as following main concepts: date of birth, family origins, leaving the parental home, conjugal history, intentions and reasons to form a union, respondent's children, fertility intentions, maternity/parental leave, organization and decision making within the household, arrangements and financial support after a separation/divorce, labour market new and education, health and subjective well-being, characteristics of respondent's dwelling, and characteristics of respondent of spouse/partner.

The target population for the 2017 GSS included all persons 15 years of age and older in Canada, excluding the residents of the Yukon, Northwest Territories, and Nunavut, also the full-time residents of institutions. The survey frame was created using two different components. One were the lists of landline and cellular telephone numbers in use available to Statistics Canada from various sources. Another was Address Register, which is a list of all dwellings within the ten provinces.

The sampling method used was stratified random sampling, by dividing Canada into 27 strata according to geographic location. Each record in the survey frame was assigned to a stratum within its province. A simple random sample without replacement of records was next performed in each stratum. Then the households with the corresponding phone number would be reached, and a respondent was randomly selected from each household to participate in a telephone interview.

The collection of this data was via computer assisted telephone interviews, which included a telephone agent who contacts respondents by phone and asks questions to collect information. The advantages of this collection process is that telephone interview is cost-effective. It doesn't get restricted on geographic location. However, it is harder to make connection with respondents through telephone interview. For those who refused to response the survey, up to two more times re-contacted phone call were made to explain the importance of the survey and to encourage their participation.

According to Cleek and Pearson(1993), children, financial condition, mental health, basic happiness are significantly affecting the marriage. Also, Shelby B. Scott found that education and age are also major reasons for divorce. Thus, we choose the following as our predictor variables for our research.

age: The age of the respondent in 2017.

total_children: Total number of children reported by respondent.

feelings_life: The satisfaction level towards life.

self_rated_health: The self rated physical health level reported by respondent.

self_rated_mental_health: The self rated mental health level reported by respondent.

income_family: The before tax income of the respondent received in 2016.

education: The highest certificate, diploma or degree that respondent have completed.

There are some variables that are possibly significant based on our common sense, such as: partner_sex, partner_main_activity, age_at_first_marriage, etc. However, since these variables contains large proportion of observations with NA, we didn't investigate on them.

In addition to these variables, we made some adjustments to the data.

Since the legal age for marriage in Canada is 18, we remove the data that are younger than 18. As to response variables, we changed the response variable into binomial by defining a new variable "divorce" as 1 if marital status is divorce , and 0 if marital status is other than divorce. Finally we removed all NAs in our data.

Urban city life always adds too much pressure to people's life. Meanwhile, it provides more entertainment than the rural areas. Therefore, we wanted to involve the pop_center variable into our response variable. Since we just cared about whether rural or urban area, we merged "Prince Edward Island" and "Rural areas and small population centers" into "Not Large Urban Population Centers".

Next, Shelby B. Scott found that education and age are also major reasons for divorce. So it is reasonable to separate the certificates into four groups called "high school", "college", "less than high school" and "university and above".

What's more, the information about whether the living place is rented or owned could help us determine the financial condition of the family.

Statistical Summary

The table below gives a statistical summary which relates to the numerical variables.

```
##                          max mini median  mean    SD
## age                       80   18   54.7 52.86 17.13
## total number of children   7    0    2.0  1.71  1.48
## feelings of life          10    0    8.0  8.09  1.65
```

For the categorical variables, we used the method of grouping, which calculates the number of each different group.

Pop center

pop_center <chr>	**Counts** <int>
Larger urban population centres (CMA/CA)	15139
Not Large Urban Population Centres	668
Rural areas and small population centres (non CMA/CA)	3763

3 rows

In order to see the distribution of population, we focused on the variable pop_center. It tells us most people live in larger urban population centers, namely, 15139 in total, and 668 people live where not large urban population centers. The remaining 3763 live in rural areas and small population centers.

Rent or own

own_rent <chr>	**Counts** <int>
Owned	14425
Rent	5145

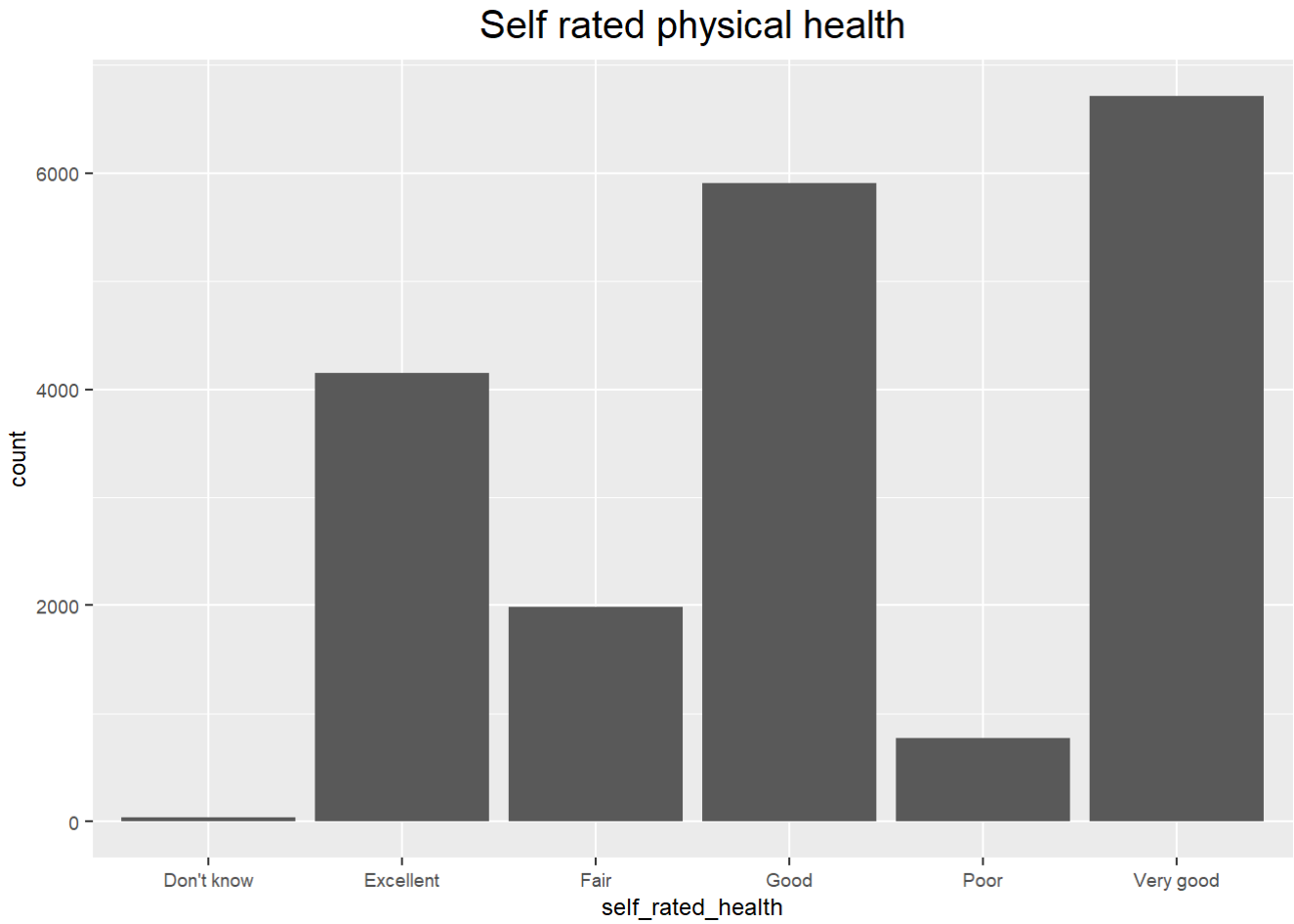
2 rows

By summarizing the information of variable own_rent. About 74% of people owned the living place, and 26% people acted as renters.

Self rated physical health

self_rated_health <chr>	**Counts** <int>
Don't know	43
Excellent	4153
Fair	1985
Good	5902
Poor	773
Very good	6714

6 rows

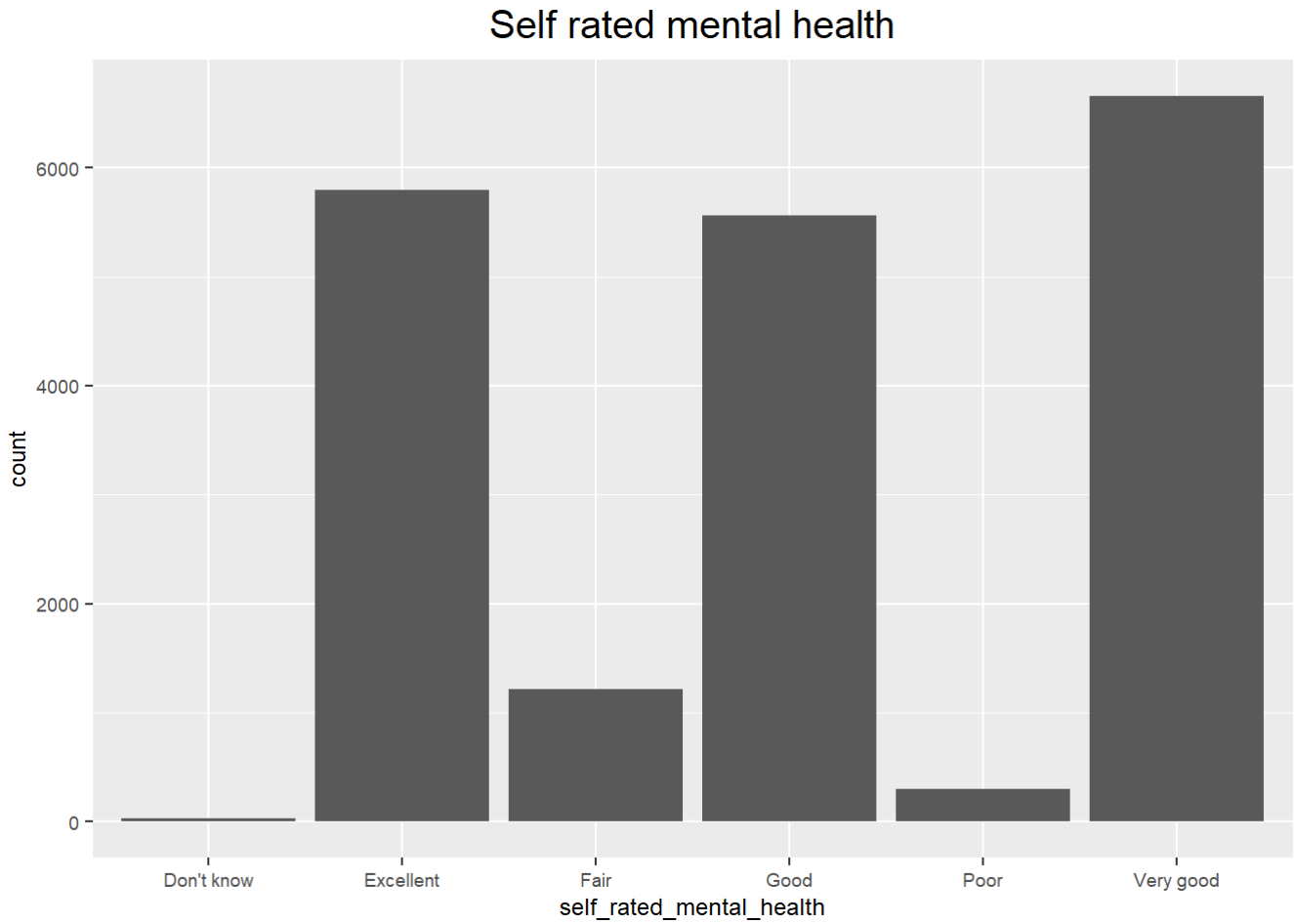


The summarized data tells us most people located the level of good physical health (including excellent, very good and good), about 16769 in total. 1985 people felt their bodies were fair enough. Conversely, 773 people were in poor physical health, 43 people did not actually know their body condition.

Self rated mental health

self_rated_mental_health <chr>	**Counts** <int>
Don't know	34
Excellent	5793
Fair	1218
Good	5563
Poor	302
Very good	6660

6 rows

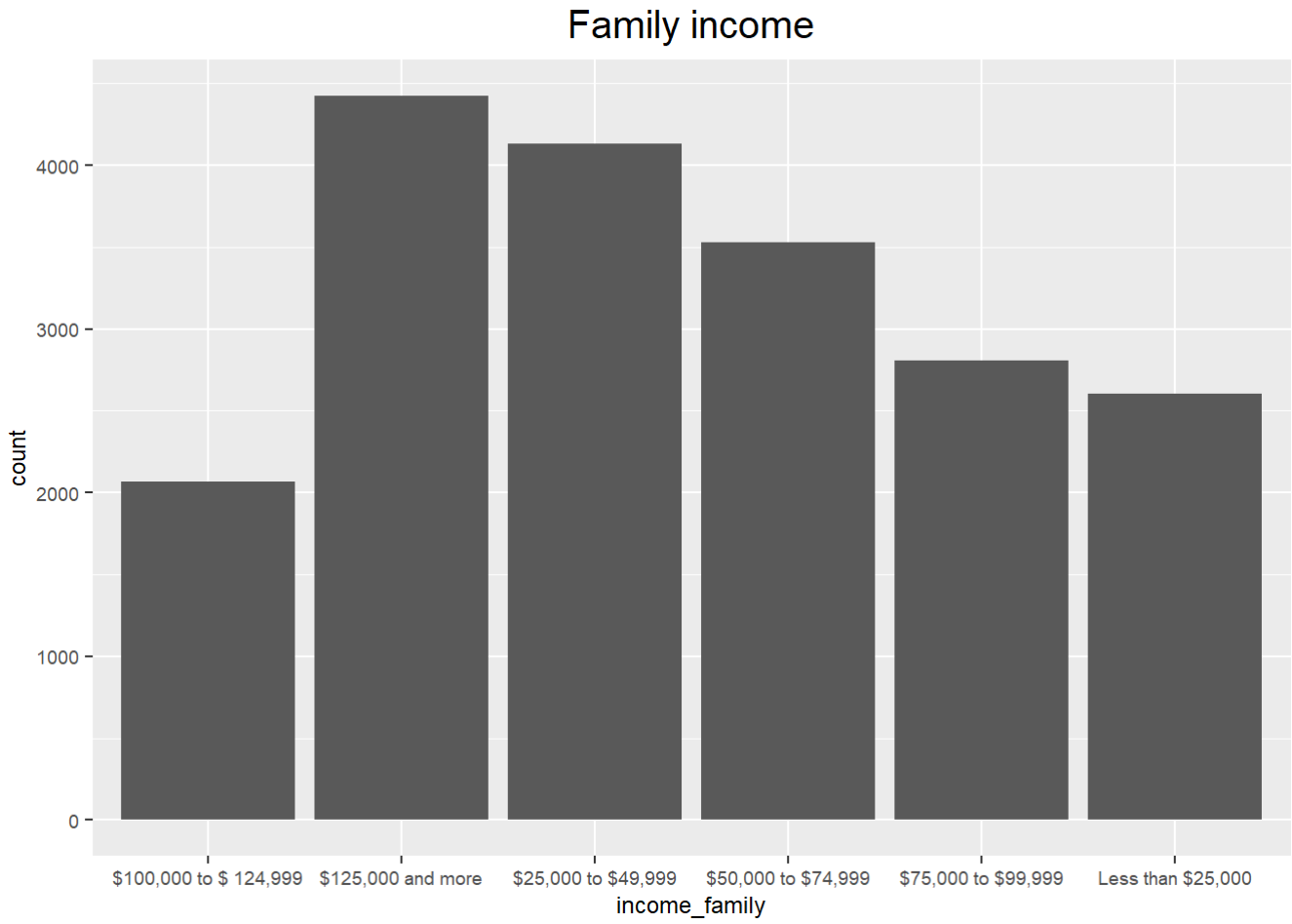


Correspondingly, self rated physical health statistics exists, it is also necessary to do a statistical summary of the data of self rated mental health. The distribution of these data is extremely similar to that of physical health. 18016 people thought they were in the level of good mental health. Instead, 302 respondents were in poor mental health and 34 people did not know their mental condition. We can also see graphically that majority of the respondents think they have positive mental health condition.

Income

income_family <chr>	**Counts** <int>
$100,000 to $ 124,999	2066
$125,000 and more	4426
$25,000 to $49,999	4135
$50,000 to $74,999	3532
$75,000 to $99,999	2808
Less than $25,000	2603

6 rows

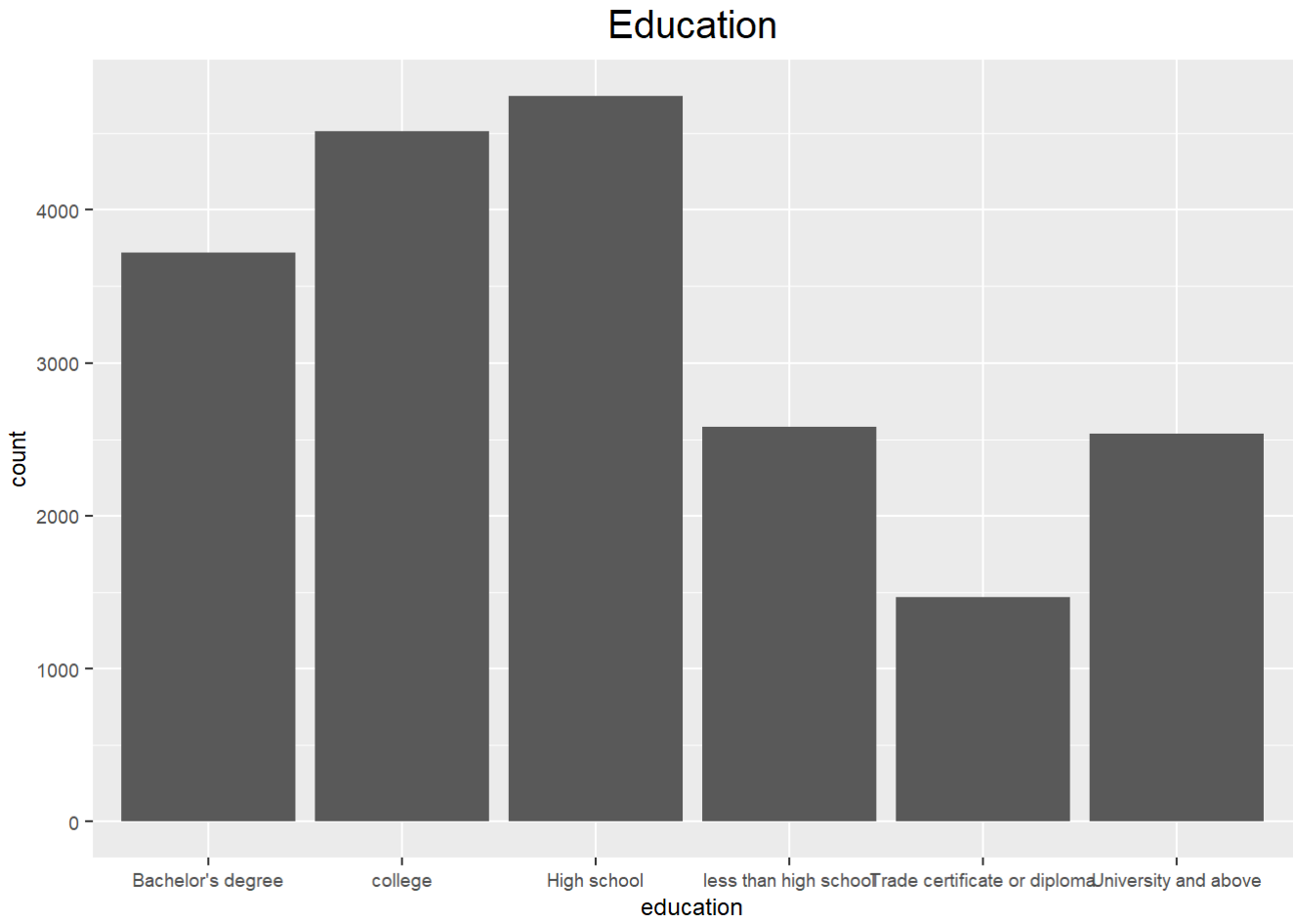


In particular, the distribution of the data of family's income is relatively on average. There are 2603 families in the lowest level of less than 25000 income. Then 4135 families distributed the next level of 25000 - 49999. In the range of 50000 - 74999, about 3532 families. As the income level is increasing, the number of families are decreasing. There are 2808 families whose income has 75000 - 99999. About 2066 families whose income is in the range of 100000 - 124999. However, the number of highest income rises, exactly 4426 families.

Education

education <chr>	Counts <int>
Bachelor's degree	3722
college	4513
High school	4745
less than high school	2584
Trade certificate or diploma	1468
University and above	2538

6 rows



This histogram tells us most people had fundamental education at high school or above. Among these people, nearly 70% of them took higher education at university or above. About 5% people act as elites who have trade certificate or diploma.

Marital status

marital_status <chr>	Counts <int>
Divorced	1708
Living common-law	2017
Married	9229
Separated	614
Single, never married	4177
Widowed	1825

6 rows

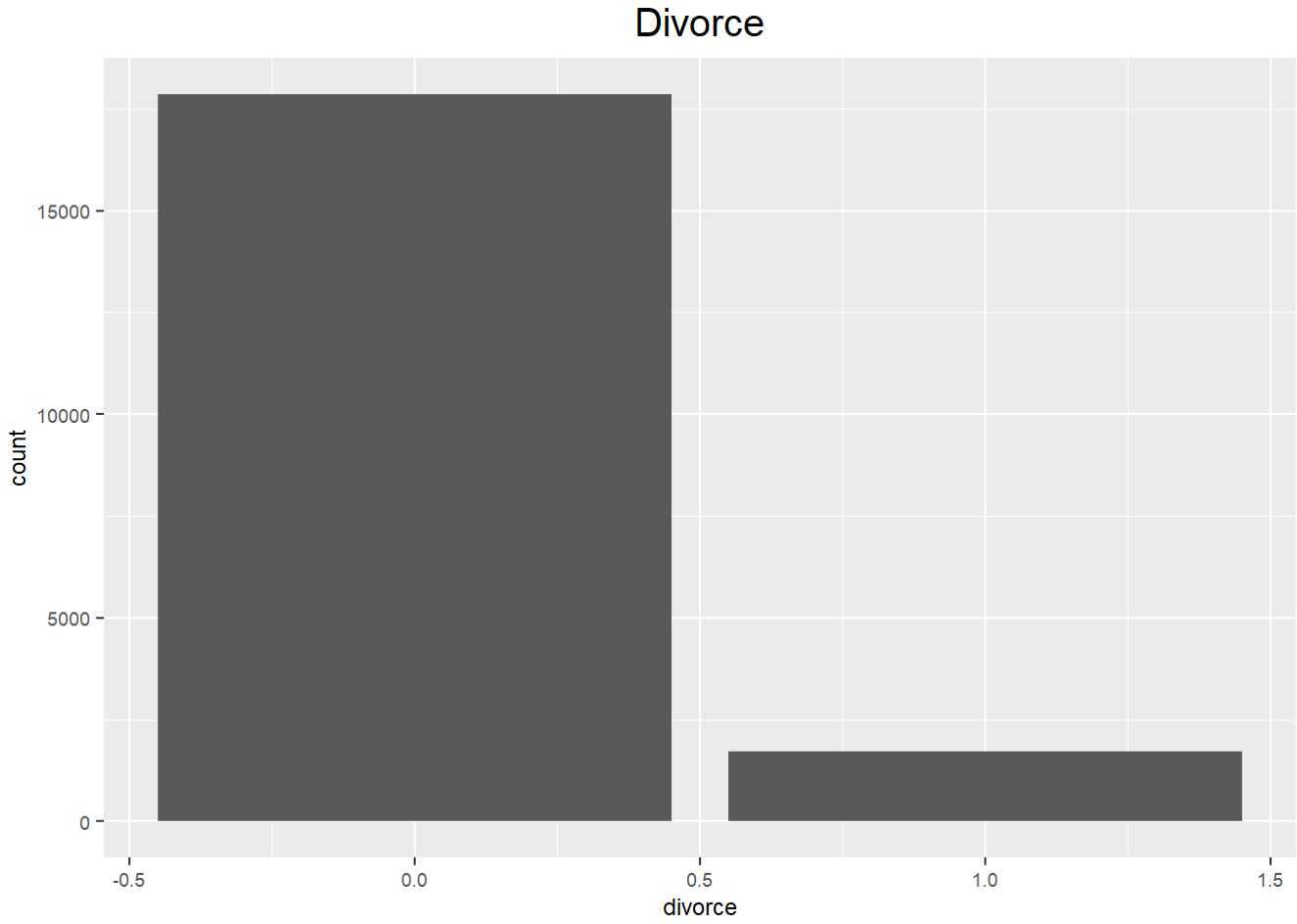
The most important data is marital status. Married people accounted for the largest proportion, 9229 out of 19570 observations. 4177 people were single and never married. In fact, there are 1708 divorced, 614 separated and 1825 widowed.

Divorce

divorce <dbl>	Counts <int>

divorce <dbl>	Counts <int>
0	17862
1	1708

2 rows



The statistical summary of variable divorce also confirmed the number of people who divorced, which about 8% of the observation.

Model

To run our model, we are going to use R on RStudio. R is a programming language for statistical computation and graphics. RStudio is an integrated development environment (IDE) for R. It supports direct code execution, and provides tools for plotting, history, debugging and workspace management.

Since the GSS data set mostly contains categorical variables and is linearly separable, logistic regression is performed to analyze the divorce of the respondents. The advantage of using logistic regression is that it is easy to implement, provides training efficiency, and is highly interpretable. In our data set, the response variable is not normally distributed, which will be well-handled with logistic regression.

We set the response variable “divorce” as binomial to fulfill the requirement for logistic regression. To fit a logistic regression model, the factor() function is applied on the categorical variables in the gss2017 data set to encode each vector as factors.

According to the user’s guide of the GSS data set, the data set was obtained from stratified sampling based on the geographic region of Canada. We divided the sample into strata by the respondent’s province. Strata are subsets of the population that have been sampled. We will assume the population size per province with

reference to the “Canada at a Glance 2017 Population” on Statistics Canada.

We created a new variable named “fpc” to set up for the finite population correction. The finite population correction is used to reduce the variance when a substantial fraction of the total population of interest has been sampled. It can be specified either as the total population size in each stratum or as the fraction of the total population that has been sampled. In the “fpc” variable, we assigned the population of each province (stratum) to the observation that corresponds to the specific province under the “province” variable.

In order to evaluate the model, we divided the data into training sets and testing sets. The receiving operating characteristic (ROC) curve will be used to perform model check. We decided to calculate the sensitivity (true positive rate) and specificity(true negative rate), noticing that ensitivity and specificity are inversely proportional to each other. We obtained the ROC curve by plotting the sensitivity against (1-specificity).

```
## Warning: package 'survey' was built under R version 4.0.2
```

We built a survey based logit model to analyze our data. The purpose of using a survey based logit model is that we can use the information from survey design to correct variance estimates. Our first step was to use the svydesign() function from the “survey” package to combine the data and important survey information. After having the survey design specified, we could use the svyglm() function to construct our model.

```
##
## Call:
## svyglm(formula = divorce ~ feelings_life + self_rated_mental_health +
##     +self_rated_health + age + total_children + pop_center +
##     education + income_family + own_rent, design = design.strs,
##     family = "binomial")
##
## Survey design:
## svydesign(id = ~1, strata = ~province, data = train, fpc = ~fpc)
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                    -7.824125
## feelings_life                                                  -0.112050
## self_rated_mental_healthExcellent                               0.765484
## self_rated_mental_healthFair                                    0.543388
## self_rated_mental_healthGood                                    0.549915
## self_rated_mental_healthPoor                                    0.592010
## self_rated_mental_healthVery good                               0.704659
## self_rated_healthExcellent                                      3.075474
## self_rated_healthFair                                           2.922693
## self_rated_healthGood                                           2.991997
## self_rated_healthPoor                                           3.044051
## self_rated_healthVery good                                      2.939686
## age                                                             0.029159
## total_children                                                  0.108213
## pop_centerNot Large Urban Population Centres                   -0.545280
## pop_centerRural areas and small population centres (non CMA/CA) -0.233421
## educationcollege                                                0.139567
## educationHigh school                                           -0.254477
## educationless than high school                                 -0.786429
## educationTrade certificate or diploma                          -0.152315
## educationUniversity and above                                   0.159979
## income_family$125,000 and more                                 -0.457676
## income_family$25,000 to $49,999                                 1.247442
## income_family$50,000 to $74,999                                 1.018754
## income_family$75,000 to $99,999                                 0.540240
## income_familyLess than $25,000                                  1.720076
## own_rentRent                                                    0.446533
##                                                                Std. Error
## (Intercept)                                                      1.398674
## feelings_life                                                    0.022271
## self_rated_mental_healthExcellent                                0.894976
## self_rated_mental_healthFair                                     0.898314
## self_rated_mental_healthGood                                     0.892483
## self_rated_mental_healthPoor                                     0.915195
## self_rated_mental_healthVery good                                0.895327
## self_rated_healthExcellent                                       1.049702
## self_rated_healthFair                                            1.050765
## self_rated_healthGood                                            1.047785
## self_rated_healthPoor                                            1.055726
## self_rated_healthVery good                                       1.047952
## age                                                              0.002174
## total_children                                                   0.023712
## pop_centerNot Large Urban Population Centres                     0.204476
## pop_centerRural areas and small population centres (non CMA/CA)  0.103860
## educationcollege                                                 0.117005
```

```
## educationHigh school                                                0.119276
## educationless than high school                                      0.143315
## educationTrade certificate or diploma                               0.166992
## educationUniversity and above                                       0.134379
## income_family$125,000 and more                                      0.223786
## income_family$25,000 to $49,999                                     0.189106
## income_family$50,000 to $74,999                                     0.191231
## income_family$75,000 to $99,999                                     0.206167
## income_familyLess than $25,000                                      0.194058
## own_rentRent                                                        0.083603
##                                                                    t value
## (Intercept)                                                         -5.594
## feelings_life                                                       -5.031
## self_rated_mental_healthExcellent                                    0.855
## self_rated_mental_healthFair                                         0.605
## self_rated_mental_healthGood                                         0.616
## self_rated_mental_healthPoor                                         0.647
## self_rated_mental_healthVery good                                    0.787
## self_rated_healthExcellent                                           2.930
## self_rated_healthFair                                                2.781
## self_rated_healthGood                                                2.856
## self_rated_healthPoor                                                2.883
## self_rated_healthVery good                                           2.805
## age                                                                 13.413
## total_children                                                       4.564
## pop_centerNot Large Urban Population Centres                        -2.667
## pop_centerRural areas and small population centres (non CMA/CA)     -2.247
## educationcollege                                                     1.193
## educationHigh school                                                -2.134
## educationless than high school                                      -5.487
## educationTrade certificate or diploma                               -0.912
## educationUniversity and above                                        1.191
## income_family$125,000 and more                                      -2.045
## income_family$25,000 to $49,999                                      6.597
## income_family$50,000 to $74,999                                      5.327
## income_family$75,000 to $99,999                                      2.620
## income_familyLess than $25,000                                       8.864
## own_rentRent                                                         5.341
##                                                                   Pr(>|t|)
## (Intercept)                                                       2.26e-08 ***
## feelings_life                                                     4.93e-07 ***
## self_rated_mental_healthExcellent                                  0.39239
## self_rated_mental_healthFair                                       0.54526
## self_rated_mental_healthGood                                       0.53780
## self_rated_mental_healthPoor                                       0.51773
## self_rated_mental_healthVery good                                  0.43127
## self_rated_healthExcellent                                         0.00340 **
## self_rated_healthFair                                              0.00542 **
## self_rated_healthGood                                              0.00430 **
## self_rated_healthPoor                                              0.00394 **
## self_rated_healthVery good                                         0.00504 **
## age                                                               < 2e-16 ***
## total_children                                                    5.07e-06 ***
## pop_centerNot Large Urban Population Centres                       0.00767 **
## pop_centerRural areas and small population centres (non CMA/CA)    0.02463 *
## educationcollege                                                   0.23295
## educationHigh school                                               0.03290 *
## educationless than high school                                    4.15e-08 ***
```

```
## educationTrade certificate or diploma                          0.36173
## educationUniversity and above                                  0.23387
## income_family$125,000 and more                                 0.04086 *
## income_family$25,000 to $49,999                               4.37e-11 ***
## income_family$50,000 to $74,999                               1.01e-07 ***
## income_family$75,000 to $99,999                                0.00879 **
## income_familyLess than $25,000                                 < 2e-16 ***
## own_rentRent                                                  9.39e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.9463098)
##
## Number of Fisher Scoring iterations: 7
```

The summary table of our model shows that there are several significant variables: Intercept, feelings_life, age, total_children, and income_family.

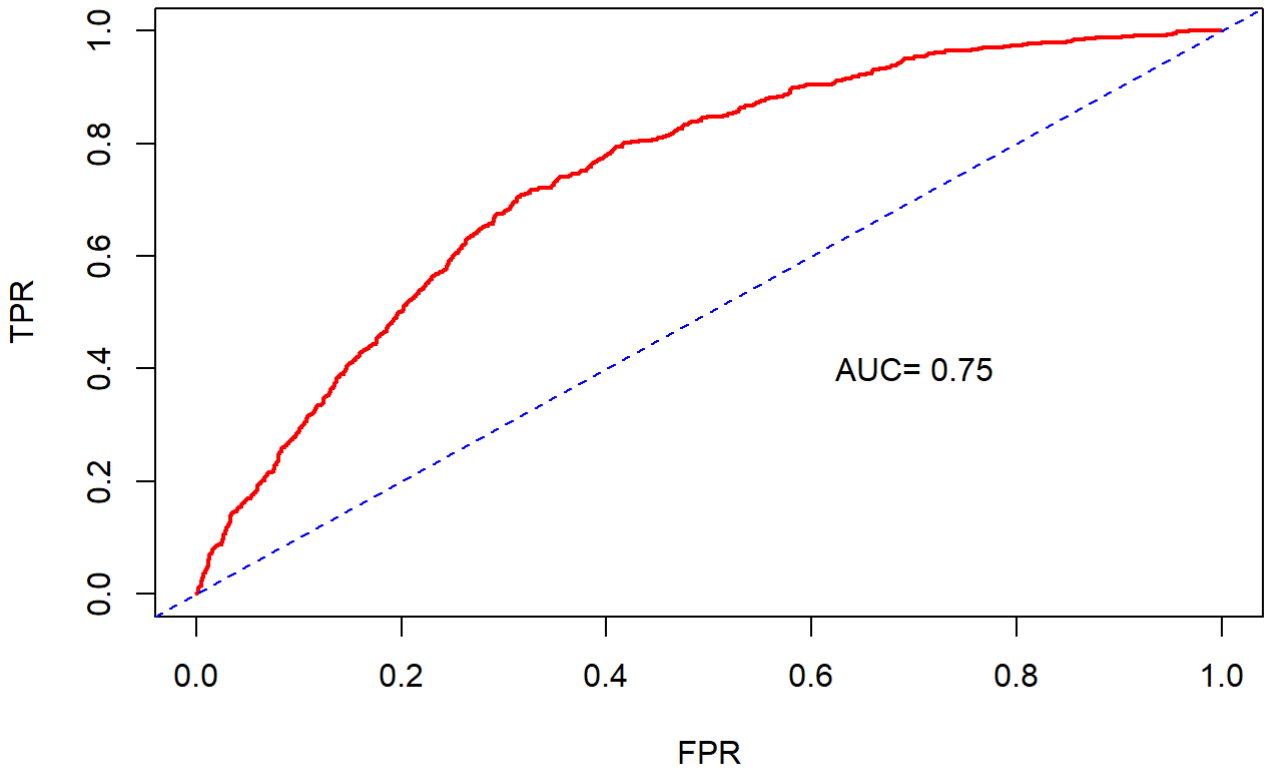
Our model is shown below, where p is the probability of getting divorced.

$$log(\frac{p}{1-p}) = -7.824 \ - 0.112 feelings\_life \ + 0.765 self\_rated\_mental\_healthExcellent$$
$$+ 0.543 self\_rated\_mental\_healthFair \ + 0.5499 self\_rated\_mental\_healthGood$$
$$+ 0.592 self\_rated\_mental\_healthPoor \ + 0.704 self\_rated\_mental\_healthVerygood$$
$$+ 3.075 self\_rated\_healthExcellent \ + 2.923 self\_rated\_healthFair$$
$$+ 2.992 self\_rated\_healthGood \ + 3.044 self\_rated\_healthPoor$$
$$+ 2.9396 self\_rated\_healthVerygood \ + 0.029 age \ + 0.108 total\_children$$
$$- 0.545 pop\_centerNotLargeUrbanPopulationCentres$$
$$- 0.233 pop_center Ruralareasandsmallpopulationcentres(nonCMA/CA)$$
$$+ 0.139 educationcollege \ - 0.254 educationHighschool \ - 0.786 educationlessthanhighschool$$
$$- 0.152 educationTradecertificateordiploma \ + 0.16 educationUniversity$$
$$- 0.458 income\_family125,000andmore \ + 1.247 income\_family25,000to49,999$$
$$+ 1.018 income\_family50,000to74,999 \ + 0.54 income\_family75,000to99,999$$
$$+ 1.72 income\_familyLessthan25,000 \ + 0.447 own\_rentRent$$

By substituting the information of the specific individual that we wanted to predict, we could get the log odds of the individual with regard to “divorce”. If we apply the exponential function, we will get the estimated odds ratio. The Odds Ratio represents the odds that an outcome will occur given particular information with the predictors, compared to the odds of the outcome occurring without this information. When the odds ratio equals 1, the exposure does not affect the odds of the outcome. When the odds ratio is less than 1, the exposure is associated with lower odds of outcome. When the odds ratio is greater than 1, the exposure is associated with higher odds of outcome.

By calculating the odds, we noticed that significant variables such as age, total_children, income 25000-49999, income 50000-74999, income less than 25000, own_rent are associated with higher odds of outcome, that is, divorce rate is related to age, the number of children, the level of income and the ownership of housing.

```
## Warning: package 'pROC' was built under R version 4.0.2
```

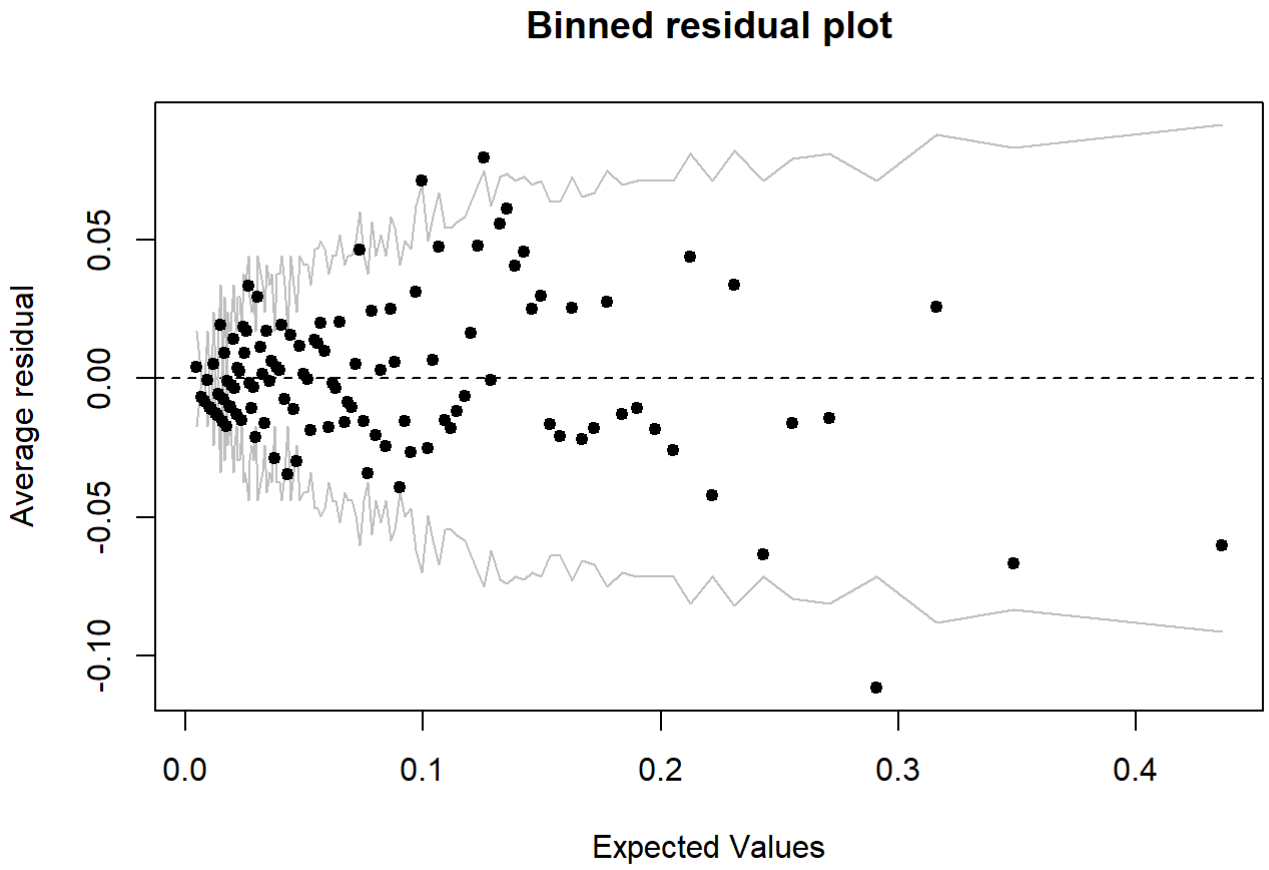


We drew a ROC curve to illustrate the diagnostic ability of our model. After drawing the ROC curve, we noticed that the area under curve (AUC) is 0.75. This indicates that there is 75% chance to discriminate between a person is divorced or not. The high AUC value indicates that our model has good discrimination ability.

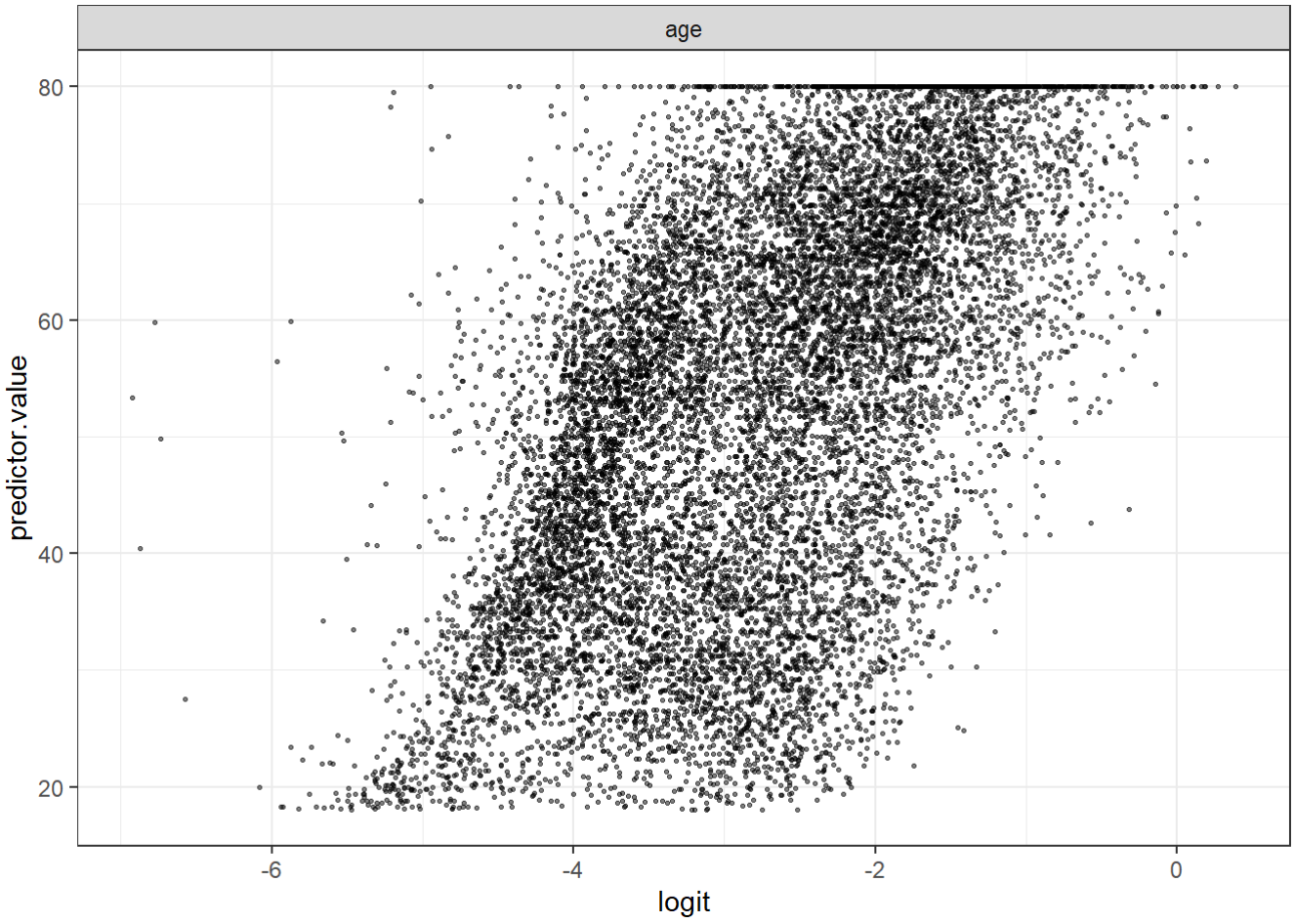
Next, we wanted to check the residuals and to find if there are some diagnostic issues.

According to Webb(2017), “In logistic regression, as with linear regression, the residuals can be defined as observed minus expected values. The data are discrete and so are the residuals. As a result, plots of raw residuals from logistic regression are generally not useful. The binned residuals plot instead, after dividing the data into categories (bins) based on their fitted values, the average residual versus the average fitted value for each bin.” Thus, we chose to use binnedplot instead of residuals plots.

```
## Warning: package 'arm' was built under R version 4.0.3
```



The grey lines represent ± 2 SE bands, which we would expect to contain about 95% of the observations. Since the majority of the fitted values fall within the SE bands, this model is reasonable.



For further model diagnostics, we have drawn the scatter plot of age variable, the scatter plot is shown above. By visually checking the plot we concluded that the scatter plot is smooth, which means that age is linearly associated with the outcome.

```
##
## Call:
## glm(formula = divorce ~ feelings_life + self_rated_mental_health +
##     self_rated_health + +age + total_children + own_rent + pop_center +
##     education + income_family, family = "binomial", data = train1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2777  -0.4705  -0.3162  -0.2034   3.0280
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                     -5.608359
## feelings_life                                                   -0.099102
## self_rated_mental_healthExcellent                                0.144160
## self_rated_mental_healthFair                                    -0.028524
## self_rated_mental_healthGood                                    -0.052663
## self_rated_mental_healthPoor                                    -0.052057
## self_rated_mental_healthVery good                                0.106504
## self_rated_healthExcellent                                       1.476167
## self_rated_healthFair                                            1.350578
## self_rated_healthGood                                            1.353912
## self_rated_healthPoor                                            1.483409
## self_rated_healthVery good                                       1.363288
## age                                                              0.028118
## total_children                                                   0.087709
## own_rentRent                                                     0.420859
## pop_centerNot Large Urban Population Centres                    -0.546340
## pop_centerRural areas and small population centres (non CMA/CA) -0.268571
## educationcollege                                                 0.141212
## educationHigh school                                            -0.222869
## educationless than high school                                  -0.747313
## educationTrade certificate or diploma                           -0.224727
## educationUniversity and above                                    0.153254
## income_family$125,000 and more                                  -0.465567
## income_family$25,000 to $49,999                                  1.240981
## income_family$50,000 to $74,999                                  1.024434
## income_family$75,000 to $99,999                                  0.518503
## income_familyLess than $25,000                                   1.713713
##                                                                 Std. Error
## (Intercept)                                                       1.238879
## feelings_life                                                     0.020248
## self_rated_mental_healthExcellent                                 0.646518
## self_rated_mental_healthFair                                      0.654346
## self_rated_mental_healthGood                                      0.645523
## self_rated_mental_healthPoor                                      0.680011
## self_rated_mental_healthVery good                                 0.645936
## self_rated_healthExcellent                                        1.031247
## self_rated_healthFair                                             1.031481
## self_rated_healthGood                                             1.029184
## self_rated_healthPoor                                             1.035168
## self_rated_healthVery good                                        1.029690
## age                                                               0.002292
## total_children                                                    0.021971
## own_rentRent                                                      0.072177
## pop_centerNot Large Urban Population Centres                      0.198524
## pop_centerRural areas and small population centres (non CMA/CA)   0.088135
```

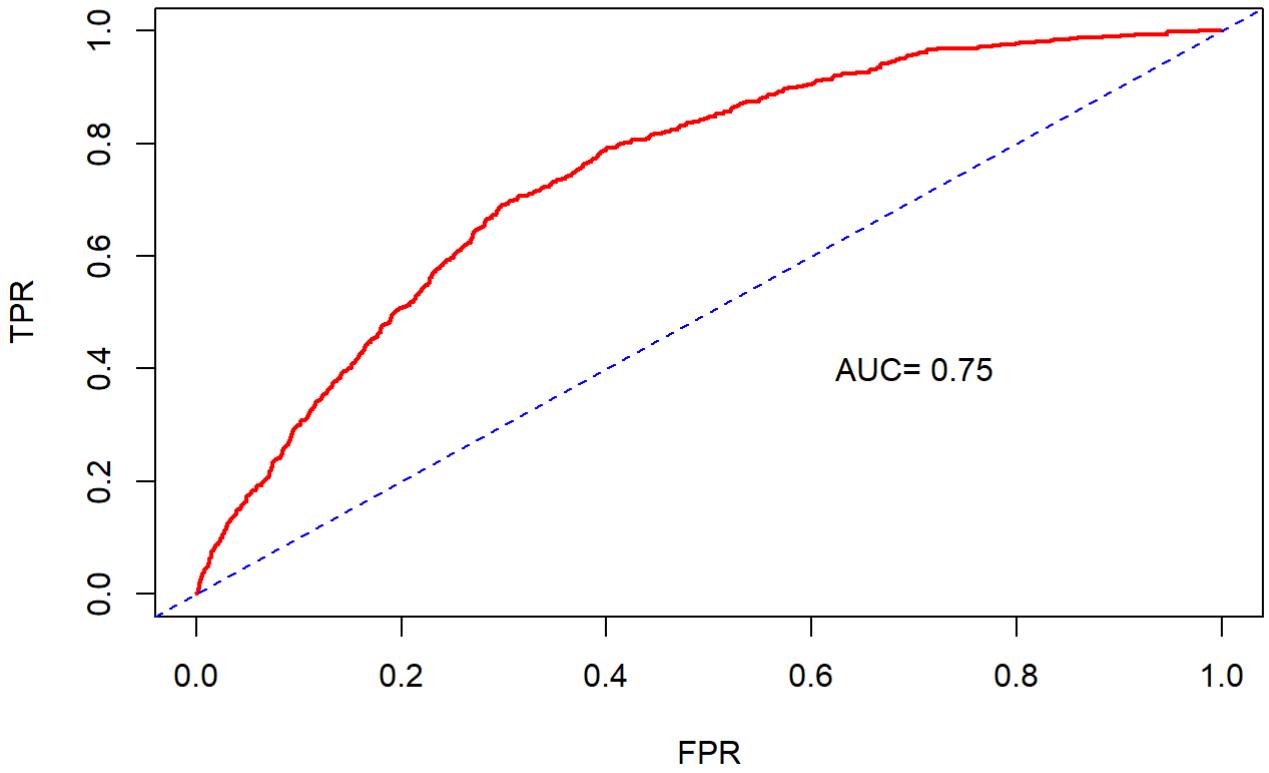
```
## educationcollege                                                0.103625
## educationHigh school                                            0.105849
## educationless than high school                                  0.127032
## educationTrade certificate or diploma                           0.148232
## educationUniversity and above                                   0.120973
## income_family$125,000 and more                                  0.196502
## income_family$25,000 to $49,999                                 0.164883
## income_family$50,000 to $74,999                                 0.166815
## income_family$75,000 to $99,999                                 0.179863
## income_familyLess than $25,000                                  0.170724
##                                                               z value
## (Intercept)                                                    -4.527
## feelings_life                                                  -4.894
## self_rated_mental_healthExcellent                               0.223
## self_rated_mental_healthFair                                   -0.044
## self_rated_mental_healthGood                                   -0.082
## self_rated_mental_healthPoor                                   -0.077
## self_rated_mental_healthVery good                               0.165
## self_rated_healthExcellent                                      1.431
## self_rated_healthFair                                           1.309
## self_rated_healthGood                                           1.316
## self_rated_healthPoor                                           1.433
## self_rated_healthVery good                                      1.324
## age                                                            12.270
## total_children                                                  3.992
## own_rentRent                                                    5.831
## pop_centerNot Large Urban Population Centres                   -2.752
## pop_centerRural areas and small population centres (non CMA/CA) -3.047
## educationcollege                                                1.363
## educationHigh school                                           -2.106
## educationless than high school                                 -5.883
## educationTrade certificate or diploma                          -1.516
## educationUniversity and above                                   1.267
## income_family$125,000 and more                                 -2.369
## income_family$25,000 to $49,999                                 7.526
## income_family$50,000 to $74,999                                 6.141
## income_family$75,000 to $99,999                                 2.883
## income_familyLess than $25,000                                 10.038
##                                                              Pr(>|z|)
## (Intercept)                                                  5.98e-06 ***
## feelings_life                                                9.86e-07 ***
## self_rated_mental_healthExcellent                             0.82355
## self_rated_mental_healthFair                                  0.96523
## self_rated_mental_healthGood                                  0.93498
## self_rated_mental_healthPoor                                  0.93898
## self_rated_mental_healthVery good                             0.86904
## self_rated_healthExcellent                                    0.15230
## self_rated_healthFair                                         0.19041
## self_rated_healthGood                                         0.18834
## self_rated_healthPoor                                         0.15185
## self_rated_healthVery good                                    0.18551
## age                                                          < 2e-16 ***
## total_children                                               6.55e-05 ***
## own_rentRent                                                 5.51e-09 ***
## pop_centerNot Large Urban Population Centres                  0.00592 **
## pop_centerRural areas and small population centres (non CMA/CA)  0.00231 **
## educationcollege                                              0.17297
## educationHigh school                                          0.03524 *
```

```
## educationless than high school                                  4.03e-09 ***
## educationTrade certificate or diploma                           0.12951
## educationUniversity and above                                   0.20521
## income_family$125,000 and more                                  0.01782 *
## income_family$25,000 to $49,999                                 5.21e-14 ***
## income_family$50,000 to $74,999                                 8.19e-10 ***
## income_family$75,000 to $99,999                                 0.00394 **
## income_familyLess than $25,000                                  < 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: 7999.4  on 13569  degrees of freedom
## Residual deviance: 7080.5  on 13543  degrees of freedom
## AIC: 7134.5
##
## Number of Fisher Scoring iterations: 6
```

We also created an alternative model that is not survey design based. We used the glm() function and set the family to be binomial. The glm() function is a function in R which fits generalized linear model. The model is shown below, where p is the probability of getting divorced.

$$
\begin{aligned}
log(\frac{p}{1-p}) = -5.608 \ - 0.099 feelings\_life \ + 0.144 self\_rated\_mental\_healthExcellent \\
- 0.029 self\_rated\_mental\_healthFair \ - 0.0523 self\_rated\_mental\_healthGood \\
- 0.052 self\_rated\_mental\_healthPoor \ + 0.107 self\_rated\_mental\_healthVerygood \\
+ 1.476 self\_rated\_healthExcellent \ + 1.351 self\_rated\_healthFair \\
+ 1.354 self\_rated\_healthGood \ + 1.483 self\_rated\_healthPoor \\
+ 1.363 self\_rated\_healthVerygood \ + 0.028 age \ + 0.028 total\_children \\
- 0.546 pop\_centerNotLargeUrbanPopulationCentres \\
- 0.269 pop_center Ruralareasandsmallpopulationcentres(nonCMA/CA) \\
+ 0.141 educationcollege \ - 0.223 educationHighschool \ - 0.747 educationlessthanhighschool \\
- 0.225 educationTradecertificateordiploma \ + 0.153 educationUniversity \\
- 0.466 income\_family125,000andmore \ + 1.241 income\_family25,000to49,999 \\
+ 1.024 income\_family50,000to74,999 \ + 0.519 income\_family75,000to99,999 \\
+ 1.714 income\_familyLessthan25,000 \ + 0.421 own\_rentRent
\end{aligned}
$$

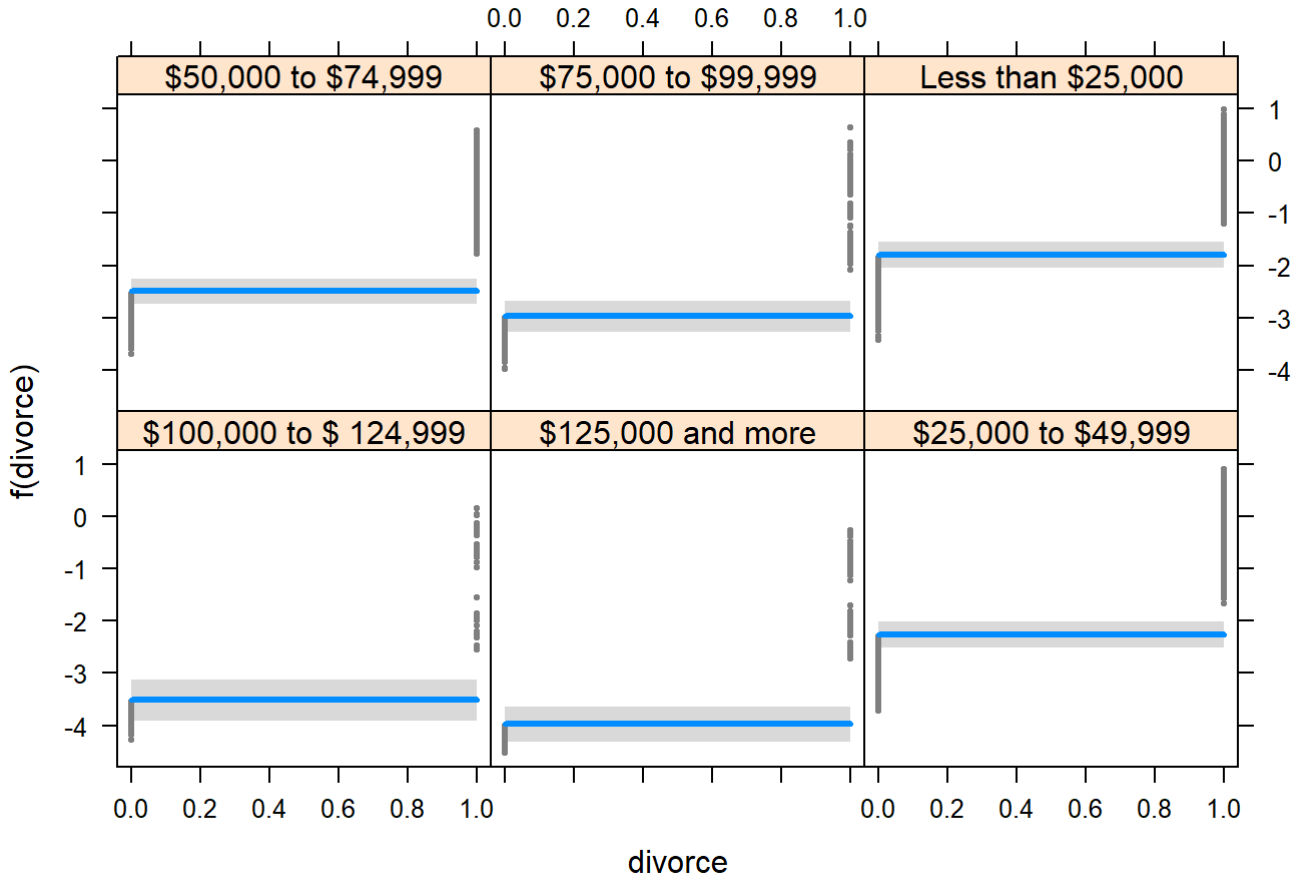
The regular model created by glm() function identifies same significant variables as the model created by svyglm() function. However, there are differences in the slope of the model, which indicates that the regular model have calculated different log odds.



We drew a receiving operating characteristic (ROC) curve to illustrate the diagnostic ability of our model. After drawing the ROC curve, we noticed that the area under the curve (AUC) is 0.75. This indicates that there is a 75% chance to discriminate between a person being divorced or not.

We noticed that the regular model has the same AUC value as the survey design based model, which means these two models have the same discrimination ability. However, since our model is based on the stratified random sampling method across Canada, it is better to use the survey design based model because it contains the information of the sampling results.

Results



In accordance with the statistical analysis result above, the realistic studies and facts also support the conclusion that the income has an impact on the possibility of divorce. It is likely that more generous cash transfers could have a stabilization effect on marriage. There would be less financial constraints and fewer arguments that lead to divorce in high-income family(Hankins and Hoekstra, 2011). The income effect from the economic theory could also explains the surprise findings partly: higher income improves living quality therefore enhances marital stability. Moreover, they will consider carefully whether get divorced since it is not simple for property division for wealthier family.

The model shows that with the age increasing, people have a greater likelihood to get divorced. It may be explained by the fact that older people may face fewer financial constraints than the younger. In addition, they will not have the concern of child support and the impact on children. From the perspective of economics, older people may attach less importance to marriage because they have more free time to pursue various substitution effects.

Our model also concludes that families with more children are more likely to get divorced. It is possible that kids increased divorce risks since they reduced the time for married couples to focus on themselves and on each other. Moreover, raising kids would add financial burden as well.

However, according to statistics, only 40 percent of divorced couples have children, compared to the 66 percent of divorced couples who do not. According to Dos & Rhoades Stanley & Markman(2009), the decline of satisfaction in a marriage with kids was nearly as steep as childless couples. We have just concluded that couples with more children are more likely to get divorced. Interestingly, this is a disputable topic. Next step, we will focus on couples with and without children to get further conclusions on the impact of children on marriage.

The ownership of housing is also linked to divorce. It is not difficult to find that a part of the people who are still providing houses after marriage, and a large group of people who have married but have not bought houses. In order to pay the mortgage, couples together to do everything to pinch pennies. In order to save money to

afford a room, they had no opportunity to relieve their emotions so that there are more quarrels. As a result, they end up in divorced.

Discussion

The strength of this dataset is that it contains plenty of observations, which indicates the sample size is quite large and our model will be more precise under the large sample size. Also, with large amount of observations, we can easily divide the data set into training and testing sets to check if our model is valid.

However, this dataset still has some weaknesses.

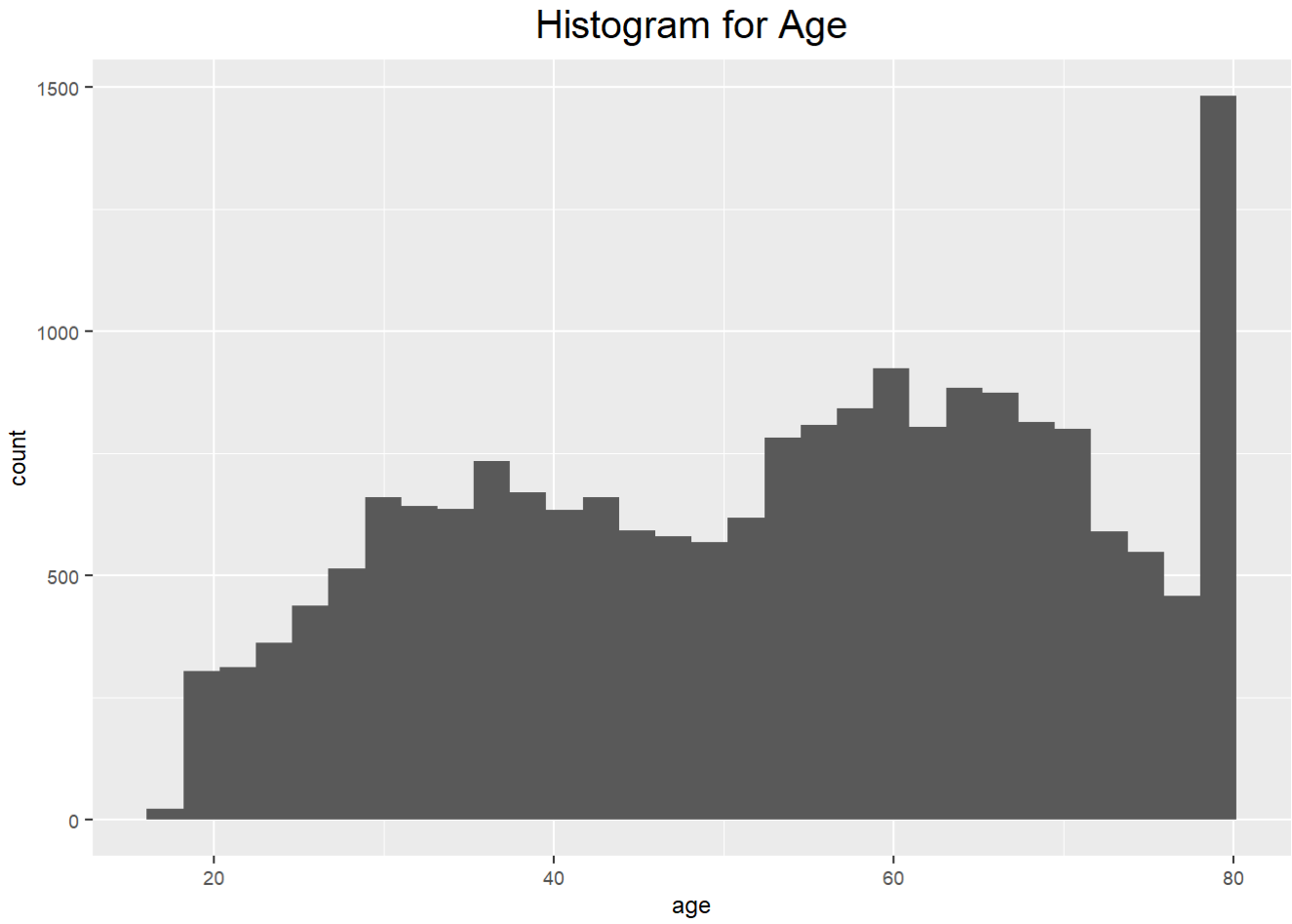
The main drawback is too many NAs, which cannot make any contributions to our analysis. After eliminating the NAs, the number of analyzable variables and observations decreases, which may influence the accuracy of our model. Another disadvantage is that the dataset has more categorical variables other than numerical variables, which limits our choice of fitting model.

Also, there is a weakness of our dataset is that it is not representative over all age ranges. The data recorded all people with 80 years and older as “80”. Our age variable indicates that the average age of the respondent is 52 years old, which is much higher than 40.8, the average age for Canadian in 2017(Erin Duffin, 2020).

Additionally, there are more potential factors that affect the age of divorce but no data has been collected such as the kind of occupation. Some busy work may cause people who do not have enough time to take care of the family, and then it is possible to trigger a divorce. What’s more, as the living environment of people is changing, the variables that affect the age of divorce are also changing. This will lead to inaccurate predictions of divorce age when people use the logistic regression model fitted by this 2017 GSS dataset to predict the age of divorce in the future.

Since the dataset does not include enough all age, we would like to look into the divorce possibility for all ages next. Since this survey was distributed via telephone, we would like to distribute our questionnaires through Internet platforms, like facebook, tweet, which would involve more young people in our respondents in the next stage.

Additionally, people of different ages are likely to hold different views to other factors like income, houses and children. Similarly, we know that children and income of the family can have an influence on the divorce rate, but we could look into whether the effect is different across income groups. Thus, we could include higher order terms or interactions between different variables.



From the above histogram, we can see that the age variable demonstrates a bimodal distribution. The bimodal distribution indicates that we possibly have two different age groups with two local maximums. A possible solution for the problem is to use Gaussian mixture model, which analyzes multivariate normal distribution.

We mainly focused on using logistic regression model. In a logistic regression, if observations are correlated, the model may overweight the significance of those observations. It is possible that logit models appear to have more predictive power than they actually do because of sampling bias.

References

Canada's divorce is data revealing—and still murky, Paul Mayne, 2020 https://phys.org/news/2020-02-canada-divorce-revealingand-murky.html (https://phys.org/news/2020-02-canada-divorce-revealingand-murky.html)

Perceived Causes of Divorce: An Analysis of Interrelationships, Margaret Guminski Cleek and T. Allan Pearson, 1993 https://www-jstor-org.myaccess.library.utoronto.ca/stable/352080?seq=3#metadata_info_tab_contents (https://www-jstor-org.myaccess.library.utoronto.ca/stable/352080?seq=3#metadata_info_tab_contents)

Reasons for Divorce and Recollections of Premarital Intervention: Implications for Improving Relationship Education, Shelby B. Scott, Galena K. Rhoades, Scott M. Stanley, Elizabeth S. Allen, and Howard J. Markman, 2014 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012696/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012696/)

Canada at a Glance 2017 Population https://www150.statcan.gc.ca/n1/pub/91-215-x/2018002/sec2-eng.htm (https://www150.statcan.gc.ca/n1/pub/91-215-x/2018002/sec2-eng.htm)

Median age of the resident population of Canada from 2000 to 2020 https://www.statista.com/statistics/444844/canada-median-age-of-resident-population (https://www.statista.com/statistics/444844/canada-median-age-of-resident-population)

Hankins, S., & Hoekstra, M. (2011). Lucky in Life, Unlucky in Love? The Effect of Random Income Shocks on Marriage and Divorce. SSRN Electronic Journal. doi: 10.2139/ssrn.1629878

Doss, B. D., Rhoades, G. K., Stanley, S. M., & Markman, H. J. (2009), The effect of the transition to parenthood on relationship quality: An 8-year prospective study. https://doi.org/10.1037/a0013969 (https://doi.org/10.1037/a0013969)

Barry Schwartz(2004), The Paradox of Choice

gss2017 family, https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/doc/gss30003.htm#csp_110c (https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/doc/gss30003.htm#csp_110c)

pROC package https://www.rdocumentation.org/packages/pROC/versions/1.16.2 (https://www.rdocumentation.org/packages/pROC/versions/1.16.2)

arm Package https://www.rdocumentation.org/packages/arm/versions/1.11-2 (https://www.rdocumentation.org/packages/arm/versions/1.11-2)

visreg Package https://www.rdocumentation.org/packages/visreg/versions/2.7.0/topics/visreg (https://www.rdocumentation.org/packages/visreg/versions/2.7.0/topics/visreg)

Logistic Regression Assumptions and Diagnostics in R http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/ (http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/)

Understanding ROC-AUC curve https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5)