Analysis for Influencial Factors of Canadian Divorce Rate

Yuxuan Lin

12/18/2020

Code and data supporting this analysis is available at:
https://github.com/woodsquarelyn/304final

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

In the problem set2, we looked into 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, from some paper and reference directly without considering the data set.

However, we found that in PS3, although we could infer some variables such as health are close related to the decision of divorce, the high p-value showed indicated that we could not reject the null hypothesis

In this study, we will more focus on the data set and variables by using Box-Cox Transformation and variable selection to make our model fits the data set better. 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.

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".

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

```
## # A tibble: 3 x 2
##   pop_center                                          Counts
##   <chr>                                                <int>
## 1 Larger urban population centres (CMA/CA)             15326
## 2 Not Large Urban Population Centres                     679
## 3 Rural areas and small population centres (non CMA/CA)  3823
```

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

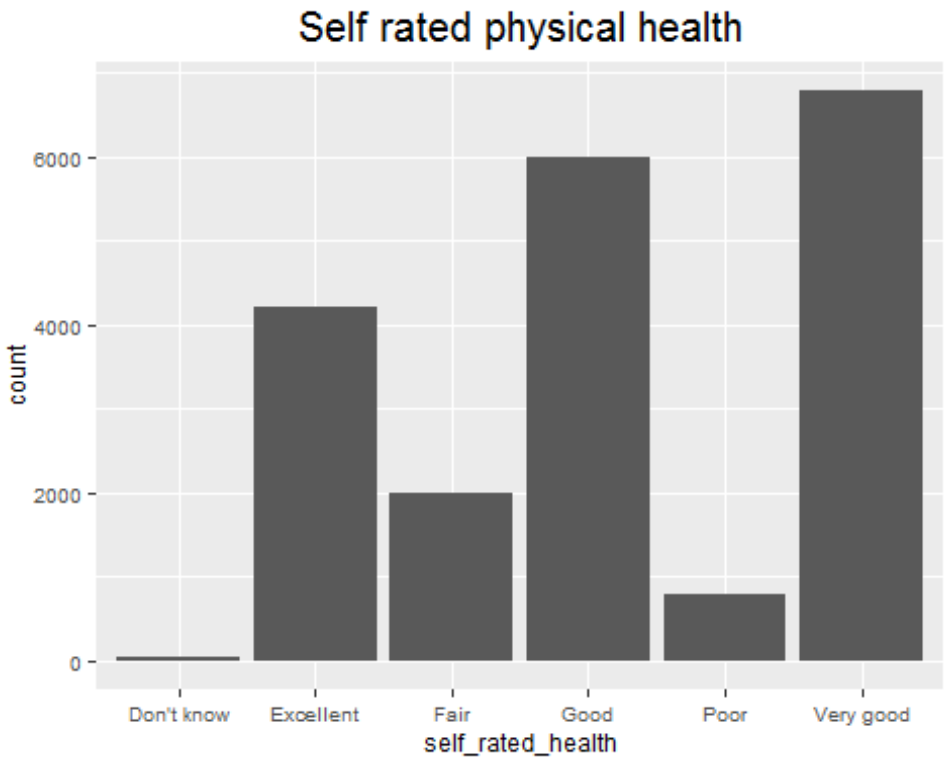
```
## # A tibble: 2 x 2
##   own_rent Counts
##   <chr>     <int>
```

```
## 1 Owned     14587
## 2 Rent       5241
```

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

```
## # A tibble: 6 x 2
##   self_rated_health Counts
##   <chr>              <int>
## 1 Don't know            47
## 2 Excellent           4214
## 3 Fair                2007
## 4 Good                5990
## 5 Poor                 786
## 6 Very good           6784
```

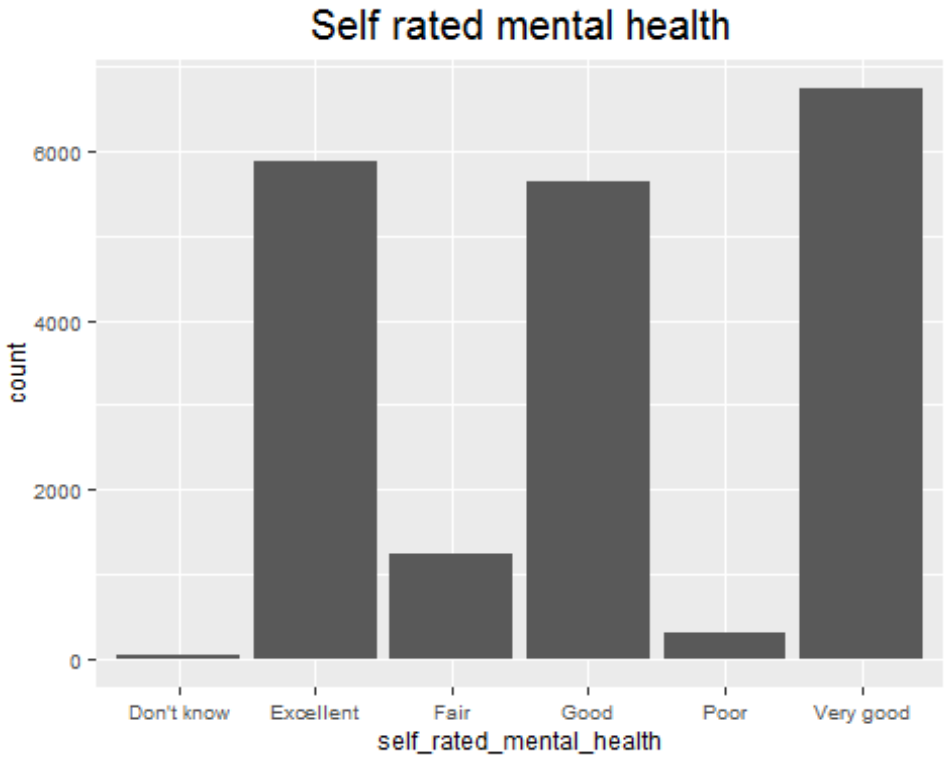


The summarized data tells us most people located the level of good physical health (including excellent, very good and good), about 16769 in total. 1985 people felt their bodies were fair enough. Conversely, 773 people were in poor physical health, 43 people did not actually know their body condition.

Self rated mental health

```
## # A tibble: 6 x 2
##   self_rated_mental_health Counts
##   <chr>                       <int>
## 1 Don't know                     37
```

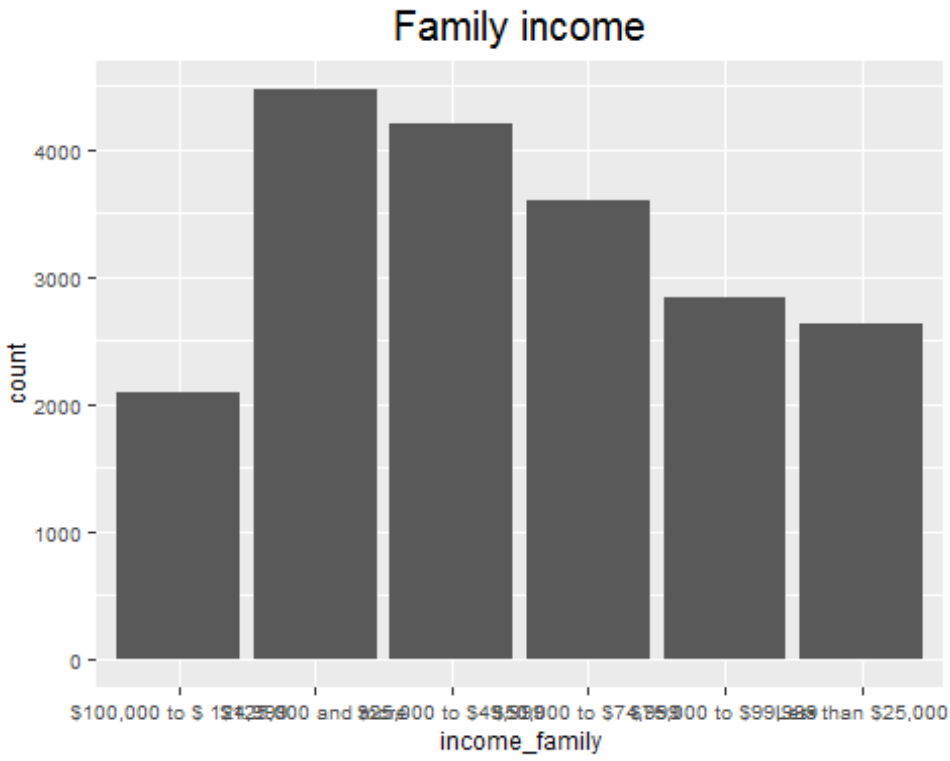
```
## 2 Excellent                  5885
## 3 Fair                       1231
## 4 Good                       5637
## 5 Poor                        306
## 6 Very good                  6732
```



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

```
## # A tibble: 6 x 2
##   income_family         Counts
##   <chr>                  <int>
## 1 $100,000 to $ 124,999   2087
## 2 $125,000 and more       4469
## 3 $25,000 to $49,999      4195
## 4 $50,000 to $74,999      3594
## 5 $75,000 to $99,999      2845
## 6 Less than $25,000       2638
```



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.

Marital status

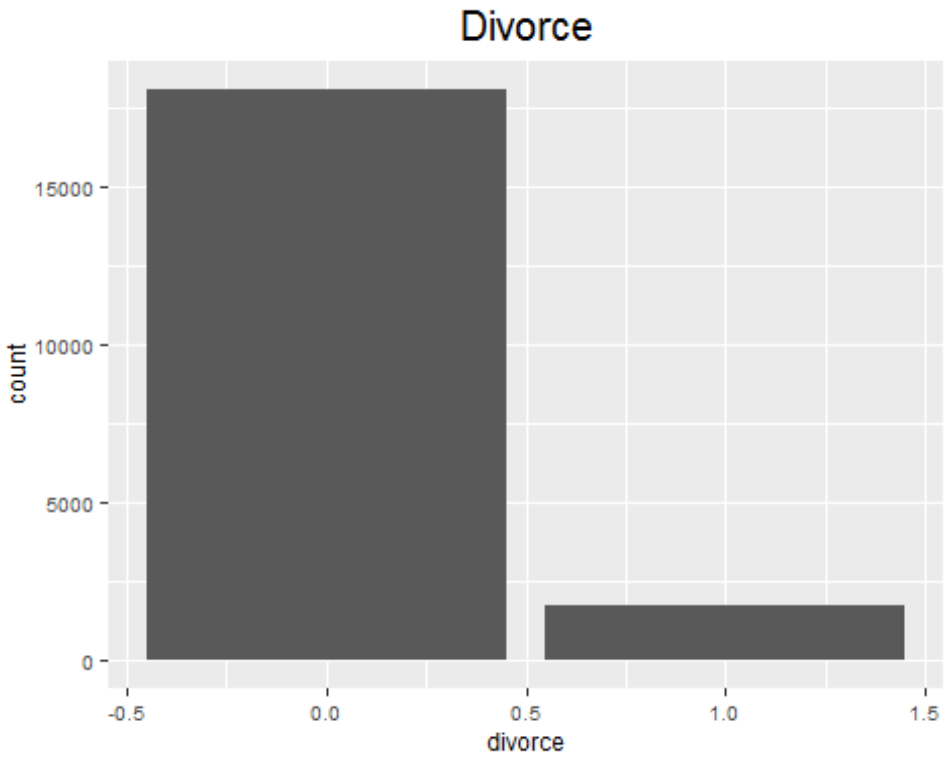
```
## # A tibble: 6 x 2
##   marital_status        Counts
##   <chr>                  <int>
## 1 Divorced                1731
## 2 Living common-law       2053
## 3 Married                 9349
## 4 Separated                626
## 5 Single, never married   4223
## 6 Widowed                 1846
```

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

```
## # A tibble: 2 x 2
##   divorce Counts
```

```
##     <dbl>  <int>
## 1       0  18097
## 2       1   1731
```



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.

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 sensitivity and specificity are inversely proportional to each other. We obtained the ROC curve by plotting the sensitivity against (1-specificity).

```
## bcnPower transformation to Multinormality
##
## Estimated power, lambda
##               Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## age              2.0827           2       1.5132       2.6523
## feelings_life    3.0000           3       2.8319       3.1681
##
## Estimated location, gamma
##               Est gamma Std Err. Wald Lower Bound Wald Upper Bound
## age            359.8983  47.5887         266.6243         453.1722
## feelings_life    9.2187   0.5842           8.0737          10.3636
##
## Likelihood ratio tests about transformation parameters
##                                LRT df pval
## LR test, lambda = (0 0) 6340.287  2    0
## LR test, lambda = (1 1) 2634.519  2    0
```

Using Box-Cox Transformation, we could find it is better to transform age to age^2, feelings to life to $feelings\_life^3$.

The original model without variable selection and transformation is:

$$log(\frac{p}{1-p})=-16.47\ -0.13feelings\_life\ -0.135self\_rated\_mental\_healthExcellent\\-0.49self\_rated\_mental\_healthFair\ -0.429self\_rated\_mental\_healthGood\\-0.67self\_rated\_mental\_healthPoor\ -0.18self\_rated\_mental\_healthVery good\\+ 12.785self\_rated\_healthExcellent\ +12.62self\_rated\_healthFair\\+12.72self\_rated\_healthGood\ +12.81self\_rated\_healthPoor\\ +0.02self\_rated\_healthVery good\ +0.025age\ +0.06total\_children\\+0.26pop\_centerNot Large Urban Population Centres \\+0.64pop_centerRural areas and small population centres (non CMA/CA)\\\\-1.806income\_family125,000 and more\ +5.831income\_family25,000 to 49,999\\+5.261income\_family50,000 to 74,999\ +2.507income\_family75,000 to 99,999\\+7.660income\_familyLess than 25,000\ +0.622own\_rentRent +0.254Prince Edward Island \\-3.853Rural areas$$

The p-value for self-rated health and self-rated mental health is very high.

The model after transformation is:

$$log\left(\frac{p}{1-p}\right)$$
$$= -16.38 - 0.008feelings\_life^3 - 0.0827self\_rated\_mental\_healthExcellent$$
$$- 0.48self\_rated\_mental\_healthFair - 0.417self\_rated\_mental\_healthGood$$
$$- 0.5529self\_rated\_mental\_healthPoor$$
$$- 0.1621self\_rated\_mental\_healthVerygood$$
$$+ 12.775self\_rated\_healthExcellent + 12.67self\_rated\_healthFair$$
$$+ 12.73self\_rated\_healthGood + 12.92self\_rated\_healthPoor$$
$$+ 12.66self\_rated\_healthVerygood + 0.00017age^2 + 0.0924total\_children$$
$$- 0.4345income\_family125{,}000andmore + 12.38income\_family25{,}000to49{,}999$$
$$+ 11.21income\_family50{,}000to74{,}999 + 0.5727income\_family75{,}000to99{,}999$$
$$+ 1.652income\_familyLessthan25{,}000 + 0.685own\_rentRent$$
$$- 0.58PrinceEdwardIsland - 0.4Ruralareas$$

After transforming, p-value for some varables are smaller but for self-rated health and self-rated mental health are still very high.

AIC removes self-rated health. After removing it, some p-value becomes smaller. After using variable selection of backward AIC, we have the model:

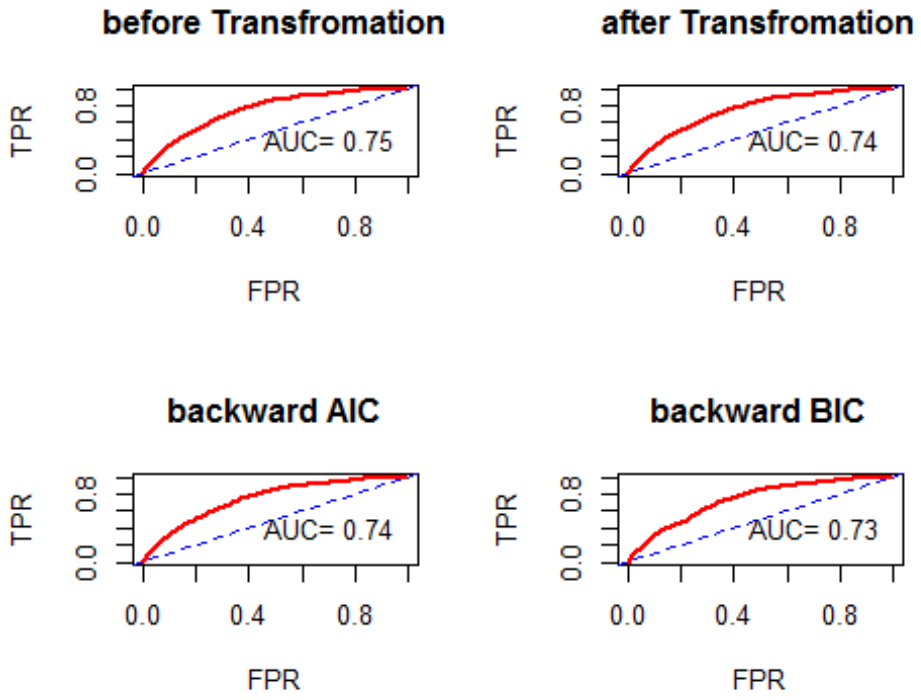
$$log\left(\frac{p}{1-p}\right)$$
$$= -3.596 - 0.008feelings\_life^3 - 0.1196self\_rated\_mental\_healthExcellent$$
$$- 0.523self\_rated\_mental\_healthFair - 0.470self\_rated\_mental\_healthGood$$
$$- 0.534self\_rated\_mental\_healthPoor$$
$$- 0.253self\_rated\_mental\_healthVerygood + 0.00017age^2$$
$$+ 0.0924total\_children - 0.4341income\_family125{,}000andmore$$
$$+ 1.235income\_family25{,}000to49{,}999 + 1.118income\_family50{,}000to74{,}999$$
$$+ 0.5715income\_family75{,}000to99{,}999 + 1.663income\_familyLessthan25{,}000$$
$$+ 0.682own\_rentRent - 0.58PrinceEdwardIsland - 0.41Ruralareas$$

BIC removes self-rated health, self-rated menial health and rent or owned. After removing them, all p-value become smaller. After using variable selection of backward BIC, we have the model:

$$log\left(\frac{\mathrm{p}}{\text{1-p}}\right)$$
$$= -3.516 - 0.0007feelings\_life^3 - 0.1196self\_rated\_mental\_healthExcellent$$
$$- 0.523self\_rated\_mental\_healthFair - 0.470self\_rated\_mental\_healthGood$$
$$- 0.534self\_rated\_mental\_healthPoor$$
$$- 0.253self\_rated\_mental\_healthVerygood + 0.00015age^2$$
$$+ 0.089total\_children - 0.4471income\_family125{,}000andmore$$
$$+ 1.326income\_family25{,}000to49{,}999 + 1.161income\_family50{,}000to74{,}999$$
$$+ 0.5956income\_family75{,}000to99{,}999 + 1.797income\_familyLessthan25{,}000$$
$$- 0.635PrinceEdwardIsland - 0.401Ruralareas$$

Results

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 0.73 or 0.74 for these 6 models. This indicates that there is 75% chance to discriminate between a person is divorced or not. The high AUC value indicates that these models have a good discrimination ability. There is no significant difference between the discrimination ability of these models. It depends on the researchers' preference to more variables or significant partial t-test to use these models.



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.

Next step, we want to focus on augment the training data set to make the model perform better on the test data set.

References

Canada's divorce is data revealing—and still murky, Paul Mayne, 2020 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

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/

Canada at a Glance 2017 Population 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

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

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

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

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

visreg Package 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/

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