A Regression model for Analyzing Factors that Affect People's Choices of Dwelling at Larger Urban Population Centres from General Social Survey data

Jiawei Wang

12/14/2020

Code and data supporting this analysis is available at:https://github.com/Amyy05/Final-Project-Jiawei-Wang.git

Abstract

The study is analyzed from the General Social Survey (GSS) data that collected in 2017 and the study aims to discover the association between the predictor variables age, total children, income of the family and the binary response variable, Canadians choose to dwell at larger urban population centres or rural areas and small population centres. A logistic regression model is built and the equation of the model is

$$logit(p) = log(p/(1-p)) = 1.618+0.00023*x_1-0.205*x_2+0.0046*x_3+0.089*x_4+0.056*x_5+0.122*x_6+0.435*x_7$$

(Here p refers to the probabilities of choosing larger urban population centres). In order to measure th fitness of the established model, likelihood ratio test and diagnostics technique are applied that shows the regression model is well fitted. My exploration about the study concluded that age, total number of children and family income are all meaningful predictors for choosing population centres as residential regions.

Keywords

Logistic Regression Model, Likelihood Ratio Test, Age, Total children, Income, (GSS) General Social Survey, P-value, Population center, Urban area, Rural area, Model Diagnostics, Cook's distance

Introduction

General Social Survey (GSS) program is founded in 1985, which gives an overview of Canadians' social life. Likewise, it helps governmental institutes to promote the well-being of Canadian society. Also, it covered various themes that the communities care about such as caregiving, families, social identities, victimization and so on. A population center determines at least 1000 size of population and the density reaches more than 400 persons per square kilometre. Also, the areas outside the population center are all defined as rural areas (Population Centre (POPCTR),2015).

I'm very interested in discovering the factors that affect whether people live in larger urban areas, which are considered as population centres. To analyze the goal of this project, a logistic regression model is established to discern if there exists a relationship between factors and whether people may choose to dwell at a larger urban population centre. Building a logistic model to study the connection between outcome and variables is a common way in Statistics analysis as well as when the outcome variable is categorical (Saishruthi S., 2018).

The report will investigate the association between the variables: age, number of total children and the family income and the outcome through 2017 GSS dataset by using logistic regression model. In methodology section, the cleaned data and model that perform population centre analysis will be described. Also, the results of analyzing whether people choose urban areas and model diagnostics are displayed in result section.

Finally, the summary, conclusion, weakness and improvements that based on results analysis will be discussed in Discussion section.

Methodology

Data

The data that used in this project is extracted from Canadian General Social Survey (GSS) in 2017, which gather data based on different topics to investigate social trends in Canadian living conditions and it works as the evidence for Canadian governments to improve well-being of society. Additionally, data is directly collected from the method of computer assisted telephone interview. The target population for the data is all non-institutional population that age is no younger than 15 years old, who live in 10 provinces all over Canada. The frame of the data gathers land-line and cellular telephone numbers from the census and a list of dwellings within ten provinces in Canada depend on Statistics Canada. The sample is the respondents of all households in 10 provinces of Canada through telephone interviews survey.

Moreover, the sampling used a stratification method. The data for the survey is all volunteering and it referenced from February to November in the year of 2017. Likewise, the respondents are from their registered address and existing telephone numbers. Non-response is filled with NA in original data, after cleaning, NAs are removed from the data. In this report, we use 5 variables, case id, population centre, age, total children of the respondent, and income of family. Case id, age and total children of the respondent are all numerical variables while population centre and income of family are both categorical variables.

The data set that used in this report is shown below. **Brief Data Table (Table 1)**

```
## # A tibble: 19,876 x 5
##    caseid pop_center                          age total_children income_family
##     <dbl> <chr>                              <dbl>          <dbl> <chr>
##  1      1 Larger urban population centres~  52.7              1 $25,000 to $49,~
##  2      2 Larger urban population centres~  51.1              5 $75,000 to $99,~
##  3      3 Rural areas and small populatio~  63.6              5 $75,000 to $99,~
##  4      4 Larger urban population centres~  80                 1 $100,000 to $ 1~
##  5      5 Larger urban population centres~  28                 0 $50,000 to $74,~
##  6      6 Larger urban population centres~  63                 2 $50,000 to $74,~
##  7      7 Larger urban population centres~  58.8               2 Less than $25,0~
##  8      8 Larger urban population centres~  80                 7 Less than $25,0~
##  9      9 Larger urban population centres~  63.8               0 Less than $25,0~
## 10     10 Rural areas and small populatio~  25.2               1 Less than $25,0~
## # ... with 19,866 more rows
```

Model

In this study, the logistic regression model is applied here to model a binary response variable, which determines whether people choose larger a urban population centre as the residential region. From the survey data, it shows that most existing variables are categorical and several of them are effective numerical variables. I know that normal simple and multiple regression model cannot use a categorical predictor, but logistic regression can use the categorical predictor. Furthermore, the data is re-arranged since the original data has a large number of variables. I selected 5 of them, including two numeric and three categorical. I 'm interested in whether people selects to dwell at larger urban population centres or rural areas and small population centres, thus, responses like "NA" and "Prince Edward Island" are removed and cleaned. As the variable income_family is categorical, the value is changed to number 1 to 6, which is ordered from the lowest income range to the highest.The cleaned data has 19876 observations.

Below is the binary response variable:

$y = 0$ if a respondent dwells at **RURAL** areas and **SMALL** population centres (Non CMA/CA)

$y = 1$ if a respondent dwells at **LARGER URBAN** population centres (CMA/CA)

The followings are 7 predictor variables:

x_1 = age = age of the respondent

x_2 = total children = the number of children that the respondent owns in total

x_3 = The income of the respondent's family is from $25,000 to $49,999, noted as rank2.

x_4 = The income of the respondent's family is from $50,000 to $74,999, noted as rank3.

x_5 = The income of the respondent's family is from $75,000 to $99,999, noted as rank4.

x_6 = The income of the respondent's family is from $100,000 to $124,999, noted as rank5.

x_7 = The income of the respondent's family is from $125,000 and more, noted as rank6.

I use **logit(p) = log(p/(1-p))**, where p is the probability of y = 1 that refers to respondent dwell at larger urban population centres.

I establish a logistic regression model, the formula is like the equation below:

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

I selected age(x_1) as different ages might have different choices of residential areas, and the number of total children(x_2) also influences the family of choices of population centres. Otherwise, it is predicted higher family income might lead the family to choose to live in larger urban population centres.

I used R Studio software to run the model, and built a logistic regression model by a function glm() in R. After fitting a model, in order to check the fitness of the model, likelihood ratio test is used to justify whether the model with seven predictor variables associated significantly to the outcome or the model with one intercept fits better. First, building a null model with one intercept, and then make the null and alternative hypothesis separately.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

$H_A : \beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq \beta_5 \neq \beta_6 \neq \beta_7 \neq 0$

Here H_0 means the null model with one intercept fits significantly better than the model with 7 predictors. Also, H_1 refers to the inverse conclusion of the significance of null model and 7-predictor model.For the two models, the difference in deviance represents the test statistics, and likelihood ratio test can be obtained by lrtest() function from R package "lmtest", which can calculate p-value. Then model diagnostics and vif() function from package "car" are applied to check the assumptions of the logistic model.

Results

Logistic Regression Output from R (Table 2)

```
## # A tibble: 8 x 5
##   term                        estimate std.error statistic  p.value
##   <chr>                          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                 1.62       0.0765    21.1    3.78e-99
## 2 age                         0.000226   0.00117    0.193  8.47e- 1
## 3 total_children             -0.205      0.0128   -16.0    1.72e-57
## 4 as.factor(income_family)2  0.00461    0.0605     0.0762 9.39e- 1
## 5 as.factor(income_family)3  0.0892     0.0633     1.41   1.58e- 1
## 6 as.factor(income_family)4  0.0561     0.0668     0.840  4.01e- 1
## 7 as.factor(income_family)5  0.122      0.0732     1.66   9.69e- 2
## 8 as.factor(income_family)6  0.435      0.0635     6.86   7.02e-12
```

In logistic regression output from R (Table 2), the established model is obtained:

$$logit(p) = 1.618 + 0.00023 * x_1 - 0.205 * x_2 + 0.0046 * x_3 + 0.089 * x_4 + 0.056 * x_5 + 0.122 * x_6 + 0.435 * x_7$$

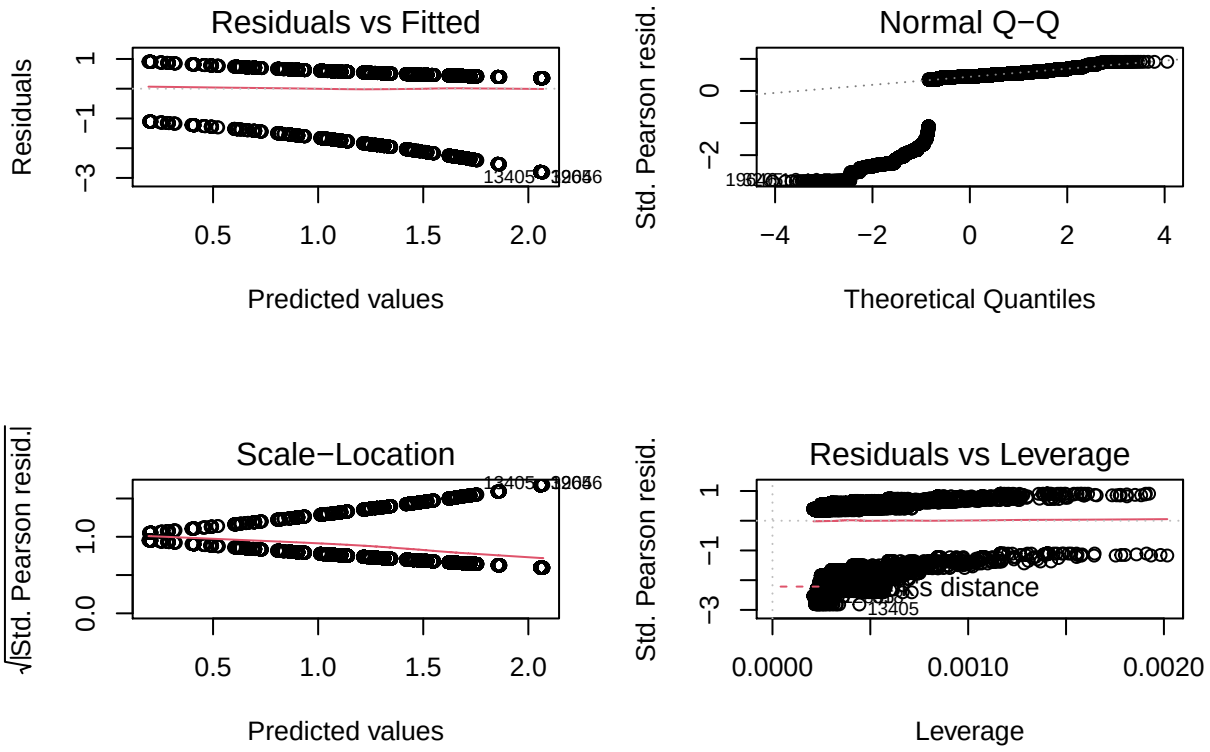
From Table 2, p-values of the predictor variables and rank 6 of are approximately $1.724e-57$ and $7.02e-12$ respectively, which are extremely smaller than 0.05 and 0.001. On the other hand, p-values of the predictor and rank 2 to 5 of are all bigger than 0.05.

Likelihood Ratio Test Output from R (Table 3)

```
## # A tibble: 2 x 5
##    X.Df LogLik    df statistic   p.value
##   <dbl>  <dbl> <dbl>     <dbl>     <dbl>
## 1     8 -9735.    NA        NA  NA
## 2     1 -9932.    -7       394.  3.74e-81
```

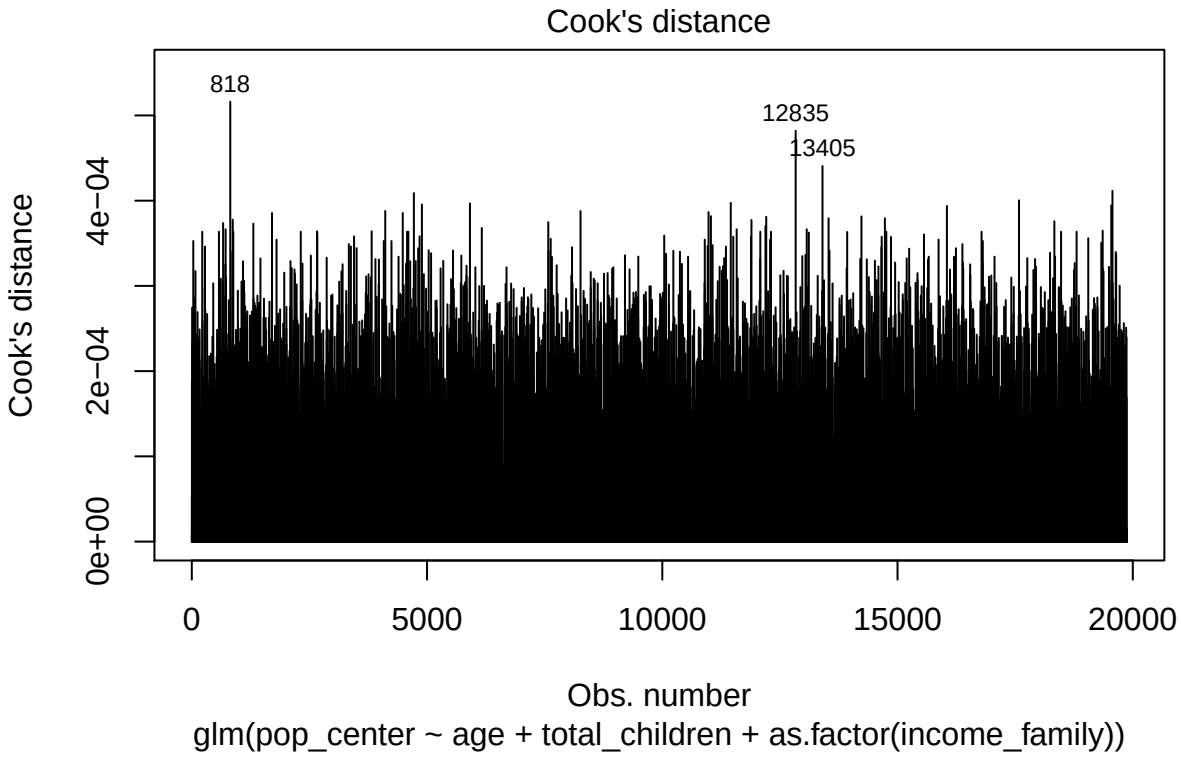
From the Likelihood ratio test output from R (Table 3), the likelihood ratio test indicates that test statistic is 394.4271 while the chi-square with 7 degrees of freedom is 394.4271 as well. Plus, the related p-value is about $3.737e-81$, which is less than 0.001.

Model Diagnostics For Logistic Regression Model Output from R (Figure 1)



The model diagnostics output (Figure 1) shows 4 plots, the plot Residual vs Fitted, it displays that there are not any non-linear patterns in the case. From the Normal Q-Q plot, it shows that residuals are not distributed perfectly on the straight dashed line and deviate severely. The plot Scale-Location wants to check if residuals are spread equally along the predictors. The graph tells that one part of residuals are followed by the red smooth line with a decreasing slope, and the residuals distribute plainly. The Residuals vs Leverage plot demonstrates there are no any observational values that outside of the Cook's distance.

Cook's Distance Output from R (Figure 2)



The Cook's distance graph (Figure 2) indicates that there exists three outlier points that have high Cook's distance scores.

Multicollinearity Check Output from R (Table 4)

```
## # A tibble: 3 x 4
##   .rownames                GVIF    Df GVIF..1..2.Df..
##   <chr>                   <dbl> <dbl>           <dbl>
## 1 age                      1.31     1            1.14
## 2 total_children           1.25     1            1.12
## 3 as.factor(income_family) 1.06     5            1.01
```

For the Multicollinearity check output (Table 4), it clarified that VIF values of predictor variables: age, total_children and income_family are about 1.31, 1.25 and 1.06, which totally not exceeds 5 and 10.

Discussion

The established model is used to explore the association between the outcome if people choose to dwell at larger urban population centres and predictor variables: ages of respondents, family income and number of total children that the respondents have. Then I apply the logistic regression model concepts to analyze the purpose and data of the study.

Data Table (Table 5)

```
## # A tibble: 6 x 6
##   caseid pop_center   age total_children income_family income_respondent
##    <dbl>      <dbl> <dbl>          <dbl> <chr>         <fct>
## 1      1          1  52.7              1 2             2
## 2      2          1  51.1              5 4             4
## 3      3          0  63.6              5 4             4
## 4      4          1  80                1 5             5
## 5      5          1  28                0 3             3
```

```
## 6      6          1  63               2 3             3
```

The data set is cleaned from Canadian General Social Survey (GSS) in 2017, and five variables: caseid, pop_center, age, total_children and income_family are selected to do the analysis. Despite the variable income_family is initially categorical, it is changed to the numerical value 1 to 6 after re-arrangement. However, the outcome variable pop_center has different values, as this report would like to study Canadians choose larger urban population centres or rural areas, I make it into binary response variables, which only has the indicator 0 or 1.

On the basis of model summary, the fitted model with coefficients is:

$$1.618 + 0.00023 * x_1 - 0.205 * x_2 + 0.0046 * x_3 + 0.089 * x_4 + 0.056 * x_5 + 0.122 * x_6 + 0.435 * x_7$$

It can be concluded with the fitted model that one unit addition in the predictors lead to the change in the log odds of the outcome (logit(p)), below are the trends: - When variable **age** of the respondent changes one unit, the average change in logit(p) is **0.00023**. - When variable **total_children** of the respondent changes one unit, the average change in logit(p) is **-0.205**. - The indicator variables **income_family** demonstrates that if family income at **rank2** (income from $25,000 to $49,999), to compare with **rank1** (Less than $25,000), and then average change in logit(p) is **0.0046**. If family income at **rank3** ($50,000 to $74,999), versus **rank1**, and the average change in logit(p) is **0.089**. Similarly, at **rank4**, **rank5** and **rank6**, the corresponding average changes in logit(p) are **0.056**, **0.122**, and **0.435** severally.

With p-values in the summary table, the predictor variables: age and income_family at **rank2**, **rank3**, **rank4** and **rank5** that the incomes of the family are below $125,000 do not have statistically significant effects on selecting larger urban population centres as residential regions. Based off this result it appears that the ages of people do not noticeably influence their choices of population centre. Otherwise, the p values for total_children and income_family at **rank6** are smaller than 0.01, thus, number of total children and high income family will choose to live in larger urban population centres.

According to the output of likelihood ratio test, the chi-square of 394.4271 with 7 degrees of freedom and the p-value for the ratio test is $3.737e-81$, which is less than 0.001. Therefore, it concluded that we have strong evidence to reject H_0 that the null model fits significantly better than the logistic model with 7 predictor variables. In other words, H_1 is accepted and the fitted model with 7 predictors fits better than the null model.

From the model diagnostics and Cook's distance outputs, since the Residuals vs Fitted and Normal Q-Q plots display, one represents the linear relationship between the predictors and binary response variable (logit(p)), which met the linearity of the logistic regression model; the other shows that the fitted model is not met normality. Moreover, the Scale-Location and Residuals vs Leverage plots state that the data do not have equal constant and the model doesn't have any cases with high Cook's distance score. Although it displays that three outliers exist, there are no influential values influencing the fitted model, which may alter the results if removing these observations. Based on VIF output, it is clearly showed that there are no problems of multicollinearity that appears the correlation among predictor variables. Consequently, all the assumptions of the logistic regression model are met through the analysis of model diagnostics.

Weakness & Next Steps

In this study, not all predictor variables in the model are effective and significant to the outcome. Meanwhile, Also, it exists the bias from the 2017 GSS data, the minority groups are not all included in the data sets. As some minority groups can also have an impact on choosing larger urban population centres or rural areas, so the survey did not include related questions about different races of people and how they affect the outcome. Furthermore, the survey questions are based on respondents' voluntary, and if they do not give the responses, "NA" might appear often on data collection and it will create errors for analyzing the data. According to my logistic model, it is necessary for me to select variables income_respondent and occupation to fit a model again.

For improving my analysis, I think that it requires some follow-up questionnaires to do further survey and questions are like "How many hours do you work in your workplace on average?" and a choice question "Do

you prefer a slow pace life or a rapid pace of life?". In addition, the GSS data needs more numerical data and the categorical data is not sufficient to fit a model. In the future analysis, building one more models is better for study the outcome, and this is a way of reducing the error for the results of the study. AIC and BIC model selection methods can be applied to select the most fitable model for the analysis.

References

Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

Brian S.Everitt & Torsten Hothorn (2017). A Handbook of Statistical Analyses Using R(First Edition). CRAN. https://cran.r-project.org/web/packages/HSAUR/

General Social Survey-Family(GSS). Statistics Canada. https://bit.ly/2T8PrNa

General Social Survey on Family (cycle 31), 2017. Statistics Canada. http://dc.chass.utoronto.ca/myaccess.html

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: AGrammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

John Fox, & etc. (2020). car: Companion to Applied Regression. R package version 3.0-10. https://cran.r-project.org/package=car

Kassambara, & U, M. (2018). Logistic Regression Assumptions and Diagnostics in R. Retrieved December 22, 2020, from http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

Lesnoff, M., Lancelot, R. (2012). aod: Analysis of Overdispersed Data. R package version 1.3.1, URL http://cran.r-project.org/package=aod

Logit Regression. UCLA: Statistical Consulting Group. https://stats.idre.ucla.edu/r/dae/logit-regression/

Population centre (POPCTR). (2015, November 27). https://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo049a-eng.cfm

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Swaminathan, S. (2019, January 18). Logistic Regression - Detailed Overview. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.