A model for analyzing how housing choice can be influenced using data from General Social Survey

Michael Huang, Jiawei Wang, Huaqing Zhang, Qiushu Zhou

Oct 19, 2020

Code and data supporting this analysis is available at:https://github.com/304PS2/ProSet2-group145.git

Abstract

The study is based on the General Social Survey(GSS) data collected in 2017 and the goal of this study is to explore the relationship between the predictor variables age, total children, income of the respondent and the estimation variable, the household rented or owned by the respondent or a member of the household. We build a logistic regression model and the equation is

$$logit(p) = log(p/(1-p)) = -0.023+0.009*x_1+0.134*x_2+0.186*x_3+0.870*x_4+1.185*x_5+1.336*x_6+1.797*x_7$$

(p is the probabilities of the house owned by the respondent). After applying a likelihood ratio test and diagnostic for the regression model, it shows that the model is well fitted. Our study shows that age, number of children and income of respondent are all valuable predictors for housing decision (rent or own) estimation.

Introduction

Human beings spend roughly 1/3 of their lifetime in bed and shelter is the most basic need of human beings. Making housing choices is very important since in most cases it is the biggest purchase in the lifetime.

The goal of this study is to find out if age, number of children and income of respondent have an effect on the housing decision of owning or renting. We use a logistic regression model since we are interested in whether the respondent owns their household or rents their household which is a binary response. To check how well our model fits, we use a likelihood ratio test and diagnostic check. After analyzing our model and graphs we have found that there is a relationship between our predictors and whether or not the respondent owns their housing.

Data

The data that we choose for our study is from the General Social Survey(GSS) in 2017, which is an annual sample survey tracking the changes within the Canadian society. Also, GSS data is used as important evidence for Canadian government to improve the well-being of Canadian society. Data was collected from February 2 to November 30, 2017 and the data is directly collected via computer assisted telephone interviews.

The population of the survey is all non-institutionalized persons 15 years of age or older, living in the 10 provinces of Canada. The survey uses a frame consisting of two components, combining landline and cellular telephone numbers from the Census and the address register which is a list of dwellings within the ten provinces from Statistics Canada. The sample of this survey is all respondents who participated in a

telephone interview. Respondents are found from their registered address and telephone numbers. Non-response are not included in the sample. This is a sample survey with a cross-sectional design. This survey is very randomly distributed across Canada which means it creates less bias. Moreover, this survey obtained better coverage of households with a telephone number. There also exist 81 different variables, hence many different analyses can be done with this data. There are several drawbacks for this survey. Firstly, too many questions in this survey, which leads to many respondents not complete this survey. Also, there are some quite personal questions that respondents did not answer them precisely, which result in the bias of the information. Moreover, the 'NA' value would exist in the collected data.

In this report, we use 5 variables, caseid, own rent, age, total children of the respondent, and income of the respondent. Caseid, age and total children of the respondent are all numerical variables. In addition, own rent and income of the respondent are all categorical variables.

The dataset we use in this report is given below.

Brief Data Table (Table 1)

```
## # A tibble: 20,396 x 5
##    caseid own_rent                            age total_children income_responde~
##     <dbl> <chr>                             <dbl>          <dbl> <chr>
##  1      1 Owned by you or a member of thi~  52.7              1 $25,000 to $49,~
##  2      2 Owned by you or a member of thi~  51.1              5 Less than $25,0~
##  3      3 Owned by you or a member of thi~  63.6              5 $25,000 to $49,~
##  4      4 Owned by you or a member of thi~  80                1 $50,000 to $74,~
##  5      5 Rented, even if no cash rent is~  28                0 Less than $25,0~
##  6      6 Owned by you or a member of thi~  63                2 Less than $25,0~
##  7      7 Rented, even if no cash rent is~  58.8              2 Less than $25,0~
##  8      8 Rented, even if no cash rent is~  80                7 Less than $25,0~
##  9      9 Owned by you or a member of thi~  63.8              0 Less than $25,0~
## 10     10 Owned by you or a member of thi~  25.2              1 Less than $25,0~
## # ... with 20,386 more rows
```

Model

The selected model for this study is the Logistic Regression Model which is a model that is used when the response variable has two possible outcomes like a yes or no question. In this survey, there does not exist many numerical variables hence it is very hard to do simple linear regression. One of the intentions for this survey is to improve the well-being of Canadians, we choose housing as the response because having a nice place to live is one major factor of a person's well-being. For this report, we want to explore the relationship between the variables age, total children, income and the outcome which is whether they rent or own a household.

In order to fit a proper model and obtain relatively precise results, We made some re-arrangement on these data. We are interested in whether the person rents or owns so every other options like "Don't know" and "NA" has been filtered out. Value for income has been changed into numbers from 1 to 6. Number 1 represents the lowest income range(Less than $25000) and so on. After cleaning this data, there are 20396 observations from age 15 to 80 in the dataset we use.

As such we define the following binary response variable:

y=0 if the respondent **RENTS** the household.

y=1 if the respondent **OWNS** the household or a member of this household owned.

we use $\mathbf{logit(p) = log(p/(1\text{-}p))}$, where p is the probabilities of y = 1 which refers to the household owned by respondent or a member of this household.

We will consider the following **SEVEN** potential predictor variables:

x_1 = age = The age of the respondent.

x_2 = total children = The number of children that the respondent has in total.

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

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

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

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

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

The reason why we choose these predictor variables is that different ages(x_1) will lead to different attitudes towards household decisions, the secure feeling a owned home can provide seems to become more important as age goes up. Also the number of children (x_2) of the respondent will change the household decision of the respondent. If the respondent has more children, we predict that they are more likely to own a household. In addition, we predict that higher income of the respondent leads to higher possibilities for respondents to own a household.

We build a Logistic Regression Model:

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

In this model, we assume that once the income rank of the respondent is confirmed, then the other predictor about the income rank will automatically equal to 0. (e.g., if the income rank of the respondent is rank1, then $x_3 = x_4 = x_5 = x_6 = x_7 = 0$.)

We use glm() function to get the logistic regression model from Rstudio.

After we get the fitted model, we want to measure how well our model fits. We use a likelihood ratio test to measure whether the model with these 7 predictor variables fits significantly better than a model with just an intercept. In order to use the likelihood ratio test, we establish a null model which is a null logistic model with just an intercept. And we claim the null and alternative hypotheses.

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

- The null model fits significantly better than the model with 7 predictor variables.

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

- The model with 7 predictor variables fits significantly better than the null model.

The test statistic for this hypothesis test is the difference in deviance for the two models, one is the model with 7 predictors and another is the null model with just an intercept. We use lrtest() function from package lmtest to do the likelihood ratio test and then we get the p-value.

To test the model we built, we also try to do a model diagnostic to check the assumptions of the logistic regression model.

Results

From the logistic regression output from R (Figure 1) we get the fitted model is:

$$logit(p) = -0.023 + 0.009 * x_1 + 0.134 * x_2 + 0.186 * x_3 + 0.870 * x_4 + 1.185 * x_5 + 1.336 * x_6 + 1.797 * x_7$$

Logistic Regression Output from R (Figure 1)

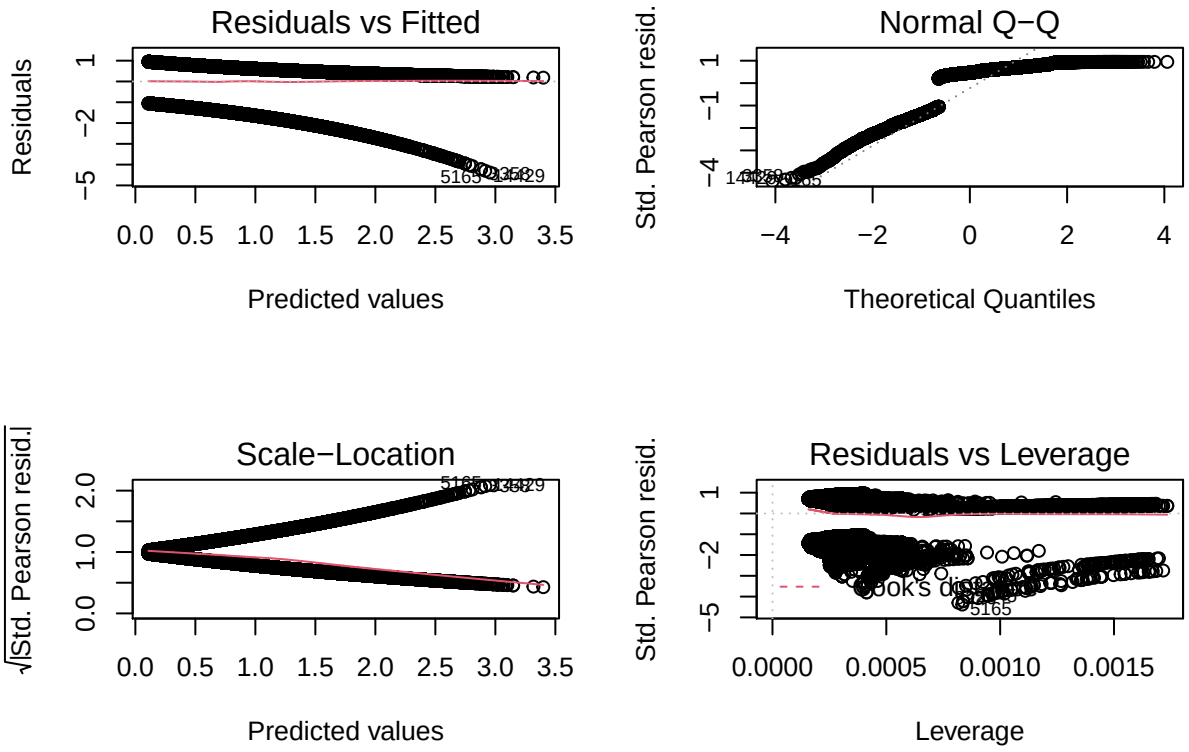
```
## # A tibble: 8 x 5
##   term                           estimate std.error statistic  p.value
##   <chr>                             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                     -0.0228    0.0514     -0.443 6.58e- 1
## 2 age                              0.00880   0.00101     8.68  3.79e-18
## 3 total_children                   0.134     0.0129     10.4   3.17e-25
## 4 as.factor(income_respondent)2    0.186     0.0385      4.83  1.37e- 6
## 5 as.factor(income_respondent)3    0.870     0.0494     17.6   1.81e-69
## 6 as.factor(income_respondent)4    1.18      0.0692     17.1   1.12e-65
## 7 as.factor(income_respondent)5    1.34      0.109      12.3   9.04e-35
## 8 as.factor(income_respondent)6    1.80      0.129      13.9   4.59e-44
```

Likelihood ratio test shows that the test statistic is 1244.5, the chi-square of 1244.5 with 7 degrees of freedom and an associated p-value less than 0.001.

Likelihood Ratio Test Output from R (Figure 2)

```
## # A tibble: 2 x 5
##    X.Df  LogLik    df statistic    p.value
##   <dbl>   <dbl> <dbl>     <dbl>      <dbl>
## 1     8 -11078.    NA        NA  NA
## 2     1 -11701.    -7     1244.  1.69e-264
```

Diagnostic for Logistic Regression Model Output from R (Figure 3)



With the appearance of the Residuals vs Fitted graph, it illustrates that there are not any distinctive patterns in the case. From the normal Q-Q graph, it shows that residuals are not lined very well on the straight

dashed line, but it follows its trend as a whole. The Scale-Location graph shows if residuals are spread equally along the predictors. It is noted that one part of residuals are followed by the red smooth line with decreasing slope, and the residuals spread evenly. The Residuals vs Leverage graph helps us find if the influential cases exist, and we observe there are no any observations outside of Cook's distance.

Discussion

Our model studies the relationship between the outcome if respondents rent or own houses and some variables i.e. age, number of total children and income of respondents. We use the logistic regression model above to analyze the data.

According to the summary of the model, the fitted model is:

$$logit(p) = -0.023 + 0.009 * x_1 + 0.134 * x_2 + 0.186 * x_3 + 0.870 * x_4 + 1.185 * x_5 + 1.336 * x_6 + 1.797 * x_7$$

From this fitted model we can conclude that the logistic regression coefficients give the change in the log odds of the outcome (logit(p)) for a one unit increase in the predictor variable.

- When **age** of the respondent changes by 1 unit, the corresponding average change in logit(p) is **0.009**.
- When the **total children** of the respondent changes by 1 unit, the corresponding average change in logit(p) is **0.134**.
- The indicator variables for **income_respondent** shows that, if the income of respondent at **rank2** ($25,000 to $49,999), versus an income of rank1 (Less than $25,000), the average change in logit(p) is **0.186**. Similarly, if the income of respondent at **rank3** ($50,000 to $74,999), versus an income of rank1 (Less than $25,000), the average change in logit(p) is **0.870**, if the income of respondent at **rank4** ($75,000 to $99,999), versus an income of rank1 (Less than $25,000), the average change in logit(p) is **1.185**, if the income of respondent at **rank5** ($100,000 to $ 124,999), versus an income of rank1 (Less than $25,000), the average change in logit(p) is **1.336**, if the income of respondent at **rank6** ($125,000 and more), versus an income of rank1 (Less than $25,000), the average change in logit(p) is **1.797**

From the output of likelihood ratio test, the chi-square of 1244.5 with 7 degrees of freedom and an associated p-value less than 0.001. It shows that we have strong evidence to reject H_0 that the null model fits significantly better than the model with 7 predictor variables. Thus we can conclude that our logistic regression model(with 7 predictors) as a whole fits significantly better than the null model(with only intercept).

Based on our model diagnostics, since the Residuals vs Fitted and Normal Q-Q graph display,, one represents the linear relationships between the predictors and outcome variable (logit(p)), which met the linearity of the logistic regression model; the other shows normality assumption of the logistic regression model. Meanwhile, the Scale-Location and Residuals vs Leverage graph concluded that the data has equal constant and the model doesn't have any cases with very high Cook's distance. Hence, there are no influential cases affecting the established model. In summary, this model met the assumptions of the logistic regression model through the analysis of model diagnostics.

Dataset Table (Table 2)

```
## # A tibble: 6 x 5
##   caseid own_rent   age total_children income_respondent
##    <dbl>    <dbl> <dbl>          <dbl> <fct>
## 1      1        1  52.7              1 2
## 2      2        1  51.1              5 1
## 3      3        1  63.6              5 2
## 4      4        1  80                1 3
## 5      5        0  28                0 1
## 6      6        1  63                2 1
```

In this report, we use 5 variables, caseid, own rent, age, total children of the respondent, and income of the respondent. The variable own rent refers to the respondent owns or rents the household, since the variable total children represents the number of total children that the respondent has. The income respondent refers to the income range of every respondent. It is initially a categorical variable, thus we change it into numbers from 1 to 6. Number 1 represents the lowest income range(Less than $25000) and so on. The outcome variable own_rent represents two ways that people choose to settle down. Therefore, we consider it as a binary response variable with two indicators 0 and 1. The dataset was collected from the General Social Survey(GSS) in 2017. There are 20396 observations in this dataset age from 15 to 80.

Based on the result of our analysis it is clear to conclude that age, number of children and income of the respondent are all important and valuable predictors for housing decision(rent or own) estimation. From our result we learned that making a housing decision actually relies on many factors. It is not a random decision that could be easily made or changed. There are some caveats, it should be very carefully used if the result is being applied to other countries. Also that there may be correlation between our chosen variable and another variable not included in this study which might affect the result. This model would be a good representation of developed countries. There wouldn't be any major difference in income and number of children for all developed countries. However some developing countries might have a large number of children in the household and the income could also defer. For any future work it might be a good idea to compare data over the course of several years to see the change.

Weaknesses

Since the study is based on questionnaires as a method to collect the data, people can choose not to respond to the questions of the survey. As can be seen from data 'gss.csv', there are many "NA" represented as "answers not available". Therefore, one of the weaknesses of our data is that it is not precise enough and there exists the missing information for the data.

Additionally, notice that we used respondent's individual income as one of the factors. However, a respondent's family income might be different from their personal income and could result in different effects on owned or rented homes. We cannot provide thoughts on which of these two kinds of income is the most eligible "income factor" for our analysis on the outcome. Whereas in fact, except the interaction with family income, there are many other factors we can consider for our study objectives such as gender, religion and education.

Next Steps

To perfect our further study of the research, it is necessary to:

- Add some interactions Adding specific analysis on interactions between each variable can be considered in Next Steps. In our current study, we focus on how significant that each factor can influence the outcome 'own/rent'; however, we did not think about whether two factors in our study would have correlations with each other and finally we might get unexpected or mistaken results.
- Add more variables such as sex, which can obtain differences between male and female.
- Do some follow-up survey: To verify our conclusion with the reality, we can set up another survey or questionnaire in order to ask participants "which factor do you think made you decide to own or rent a living place?". We can compare our analyzed results with respondent's autonomous answers to see whether our study on own/rent is rational or not.

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

Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression (Second Edition). New York: John Wiley & Sons, Inc.

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

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

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/

Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

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.

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