GSS Families

Joanne Sun, Leqi Sun, Tzu-Ang Su, Cameron Fryer

October 19, 2020

Abstract

Home ownership is an important indicator of life quality and the economy of a country as whole. In this paper, we investigate the Canadian General Social Survey ("GSS") – Family (cycle 31) dataset, which was a probability survey administered in 2017. Consequently, we find that men are more likely to own a house than women, and that persons aged 40-65 years are more likely to own a house than those belonging to other age groups. While both of these findings provide implications for deeper social issues, it's also worth noting that the province where Canadians are least likely to own a house and most likely to own a house are Quebec, and Newfoundland, respectively. Code and data supporting this analysis is available at: https://github.com/tomsu0826/g50gssfamiliescycle31

1 Introduction

Prior studies on house ownership have discovered links between the subject matter and many other aspects of life, including marital status, employment, job satisfaction, family income, and age (Battu et al, 2008; Fisher & Gervais, 2011; Lersch & Vidal, 2014; Tumen & Zeydanli, 2014). The problem, however, is that these influencing factors were investigated rather independently and qualitatively. Thus, by utilizing data from the Canadian General Social Surveys ("GSS") on family life (cycle 31), we created a Bayesian multilevel model to quantify how age, gender, and geographical location (by province) affect an individual's likelihood of being a homeowner. Accordingly, we found that Canadians with the greatest likelihood of owning a house are male rather than female, are in the age group of 29-39 years old, and live in the province of Newfoundland and Labrador.

Second to Newfoundland in terms of its residents' probability of home ownership is the province of Alberta. Furthermore, the survey participants from Quebec are least likely to own a house, followed by those from British Columbia as second least likely. Through combining these findings with additional research, explanatory inferences are made. For example, by conducting research on the average house price in Newfoundland, insight as to why Newfoundland has the greatest probability of home ownership is gained. Other important implications drawn from the results will be interpreted in this report as well.

The paper will begin with some basic information about where the data was derived from, followed by a plot of the raw data. Thereafter, a rundown of the adopted model will be given, along with a description of the priors, regularization and other significant aspects. Then the model results will be displayed. Afterward, the results will be discussed; future work and possible weaknesses will be highlighted too.

2 Data

2.1 Data and Sampling

For this paper, the dataset 2017 Canadian General Social Surveys ("GSS") on Family (cycle 31) is used. This dataset is obtained from the University of Toronto ("U of T") Library, specifically, via the Computing in

the Humanities and Social Sciences ("CHASS") Data Centre. The dataset on GSS is provided by Statistics Canada. The dataset is downloaded digitally using the Survey Documentation and Analysis ("SDA") system in CHASS Data Centre, which is only accessible by U of T students, staff, and faculty members.

The data was collected from February 2 to November 30, 2017 via computer assisted telephone interviews ("CATI") where the respondents were interviewed in the official language of their choice. The target population was all persons who are age 15 or older in Canada but excludes residents of Yukon, Northwest Territories, and Nunavut; and full-time residents of institutions. The listing of population from which Statistics Canada can draw samples from, otherwise known as the "sampling frame" was created using a list of telephone numbers in use, both landline and cellular available to Statistics Canada from various sources. The Address Register (AR) which contains a list of all dwellings within the ten provinces was also included in creating the sampling frame. The use of AR helped Statistics Canada to group all telephone numbers together with the same address. The target number of respondents or sample size was 20,000 and in the end the actual number was 20,602.

During the sampling period, each of the ten provinces were divided into geographic areas or strata. In each stratum, a simple random sample without replacement was used, the respondents were randomly selected from each household for the telephone interview. For the people who first refused to participate either due to inconvenient call time or some other reasons were recontacted later on, along with an explanation why the survey was important and encouraged their participation. In the end, the overall response rate was 52.4%. Finally, the cost of this survey was not disclosed in the official documentation obtained from CHASS.

2.2 The Survey

The survey itself is very extensive and covers many grounds in regards to families in Canada. Aside from the basic personal information like sex and date of birth, many other concepts were introduced and asked such as family origins, conjugal history, leaving the parental home (young adults leaving home), intentions and reasons to form a union, respondent's children, fertility intentions, maternity/parental leave, organizations and decision making within the household, etc. Each concept contains multiple questions about it. One key strength about this survey is the length. Once completed, a very detailed profile of a respondent's family can be produced. However, this invariably becomes an issue, or a key weakness. The official documentation did not specify the average time a respondent took to complete the survey, but a good estimate could be within 30 minutes to an hour, and conducting such a long survey over the telephone may be too time and energy consuming. To solve this issue, every question in the survey was designed to have Don't Know (DK) and Refusal (RF) as options, so respondents can always refuse, or in a way, "skip" the question very quickly.

2.3 Data Used

Once the raw dataset was downloaded from CHASS, we used a script created by Rohan Alexander and Sam Caetano to clean up the data. See GitHub repository for the script and it's output. Even after cleaning (only a selected number of variables from the raw data were kept), the dataset is still huge, thus we have decided to not use all the variables, and only pick out ones we think are relevant. Our goal here is to find which variable has the greatest impact on homeownership, since the survey itself goes into detail on every topic, using it may not be representative of the population, hence we unfortunately cannot choose those variables as the variables we want to use. In our case, we think the more general a variable is, the better. Therefore we chose only three: age, gender, and geographical location (by province) to investigate the effects of it on homeownership. Figure 1 and 2 below shows plots of the raw data.

Figure 1: Respondents' Age Groups

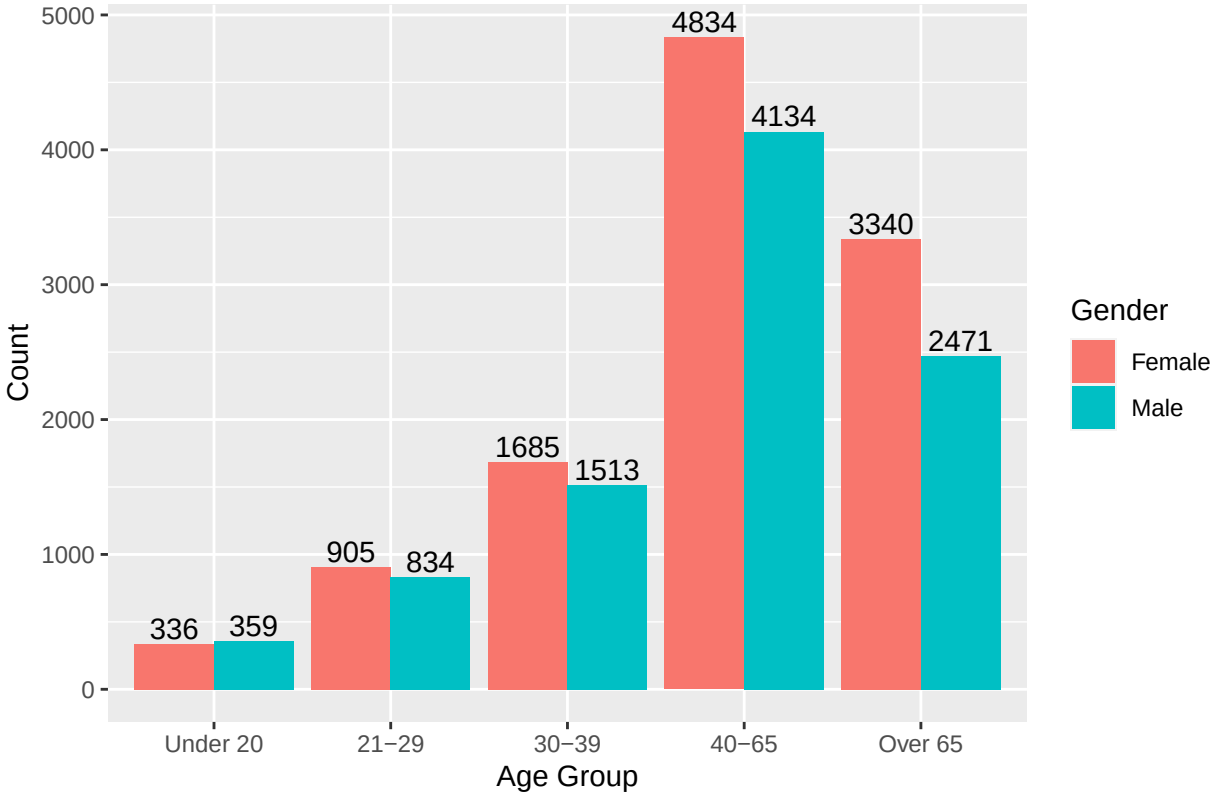
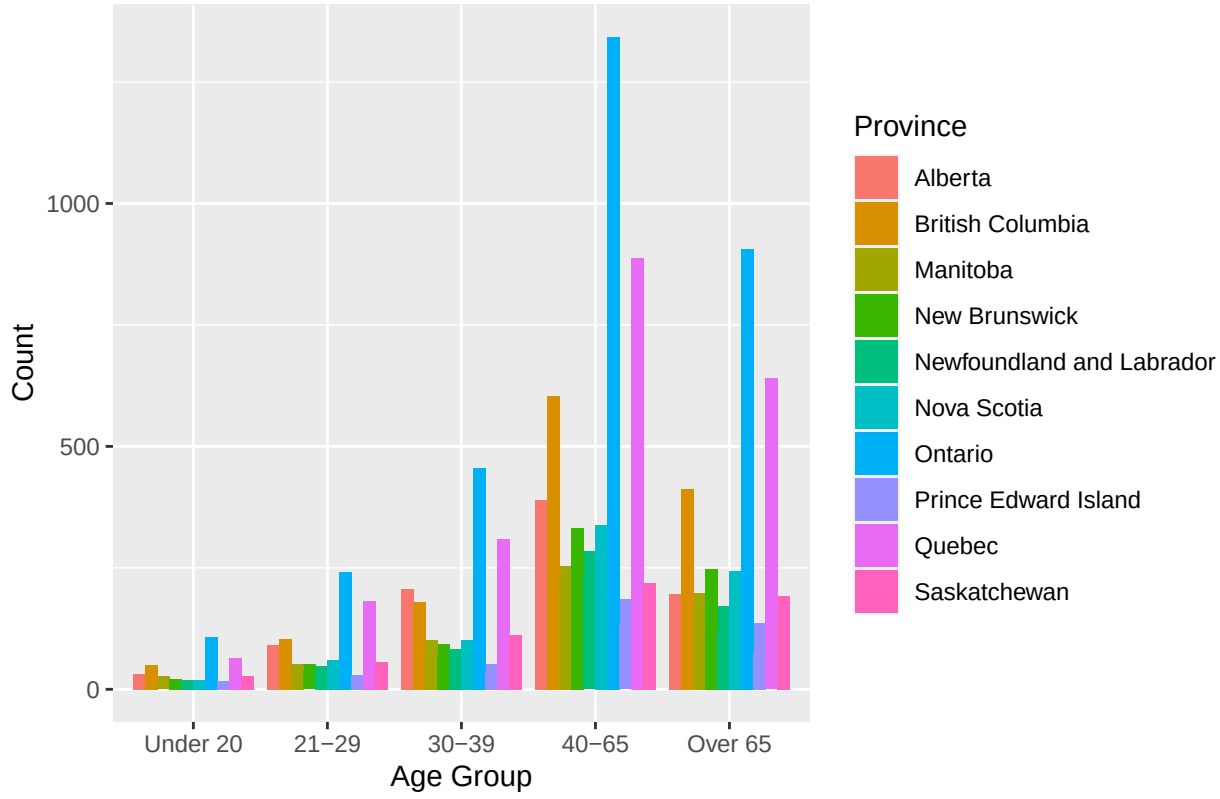


Figure 2: Respondents' Province Divided by Age Groups



3 Model

We are interested in explaining whether a person owns a house based on age, gender and province of residence. Let $y_i = 1$ if the respondent owns a house (or any kinds of dwelling). The model is as the following:

$$Pr(y_i = 1) = \text{logit}^{-1}\left(\beta_0 + \alpha^{age}_{a[i]} + \alpha^{male}_{m[i]} + \alpha^{province}_{p[i]} + \epsilon\right)$$

β_0 is the global intercept. The notation $a[i]$ refers to the age-group a to which individual i belongs. There are 5 age groups in total: under 20, 21-29, 30-39, 40-65, and over 65. The notation $m[i]$ refers whether individual i is male. Similarly, $p[i]$ refers to the province individual i resides. The priors for these variables are:

$$\beta_0 \sim N(0, 2)$$
$$\alpha^{age}_a \sim N(0, 2) \text{ for } a = 1, 2, \ldots, 6$$

The notation $a = 1, 2, \ldots, 6$ represents the age group of each individual.

$$\alpha^{male}_m \sim N(0, 2) \text{ for } m = 1, 0$$

$m = 1$ when the individual is male, $m = 0$ when the individual is female.

$$\alpha^{province}_p \sim N(0, 2) \text{ for } p = 1, 2, \ldots, 10$$

The notation $p = 1, 2, \ldots, 10$ represents the province of residency for each individual.

$$\epsilon \sim t(3, 0, 2.5)$$

The error term ϵ has been modeled as following a student t-distribution.

The priors for fixed effects are all $Normal(0, 2)$ on *logit* scale, which means that most of the samples are from about $0.018 - 0.98$ on a natural probability scale.

We chose weakly informative priors for two reasons: first of all, we do not have strong assumptions; secondly, we know that the probability must be greater than 0 and smaller than 1. By setting weakly informative priors, we hope that the data could tune the posterior distribution as efficiently as possible.

3.1 Model Checking

Let's do some model checking. Firstly, we do convergence diagnostic to make sure that the MCMC sampling size is big enough. The results of Gelman-Rubin diagnostic shows that the upper confidence limits are either 1 or very close to 1, which indicates that the sample chains are converging. It does not flag an issue.

The correct classification rate of the model is "r ccr". The Confusion Matrix of the model is shown in the table. The model tends to exaggerate the rent rate.

Before proceed to do post-stratification, one more model check is needed. A posterior predictive check is necessary. In the plot, y represents the observed data, y_{rep} refers to a randomly sample from the posterior distribution. The plot shows that the posterior density is consistent with the observed data.

The we do post-stratification based on the age, gender and province. The census data used is from the 2016 Census cycle. The plot shows that the model does a fair job.

4 Results

The model estimates and credibel intervals are shown in the table.

5 Discussions

A gendered pattern of house ownership is reflected in our model, as the probability of owning a house is greater in males compared to females across Canada. This tendency in house ownership is also consistent with the universally shared image of family housing, where it is usually the husband who provides the financial ground for a physical shelter, whereas the wife takes care of kids and daily chores. One robust explanation for this pattern lies in the gender pay gap, where women are usually paid less as compared to men with similar jobs and qualifications. The lower income presumably limits women's ability to apply and pay for mortgages without compromising life quality. What's more, marital status can also influence one's decision of buying or ability of keeping a real estate, and prior study has shown that separation is negatively associated with house ownership. Generally, ex-partners are more likely to move out of an ownership with a reduced probability of moving back into again, but the materially more well-off ex-partners are more likely to keep the current house ownership (Lersch & Vidal, 2014). Following the idea of gender pay gap, it is safe to presume that men are more often the materially well-off side in a heterosexual marriage, contributing to the higher likelihood of them owning a house seen in the model.

Based on our model, people of mature working age (40-65) were the most likely to own a house (β= 0.4814), while young adults (21-29) were the least likely (β= -0.8663). This pattern is consistent with common intuition, as older people with stable full-time jobs and more established careers are better capable of applying for and paying off mortgages. Additionally, since 40 to 65 is also around the age where people need to sustain a family on a long, martial term, these people may also be in a greater need of a permanent physical shelter. However, other research has shown that house ownership rates have actually been falling

among people aged 25-44 years due to falling marriage rates and increasing household earnings risk. This reduced incidence of marriage leads to a weaker need for stable dwelling, and the increase in the earnings risk leads to a delay in the purchase of a first house to when the household becomes wealthier and more stable to purchase a larger house that has investment value (Fisher & Gervais, 2011). Although our model reported that people aged 40-65 were the most likely to own a house, in the 30-39 group, the likelihood of owning a house was somewhat reduced (β= -0.2180), which complements for the effects seen in the older age group and makes the model consistent with literature evidence. Beside the effects of age alone, it is also worth noting that in Canada, recent increases in interest rates and required qualifications have been slowing the housing market down (Bilyk & TeNyenhuis, 2018), which may be confounding with the effects of age as it is getting harder for people with poorer qualifications and less-paid jobs to get an mortgage opportunity. Although house ownership is still the highest among working-age people in Canada, the overall rate of ownership may be downfalling, calling for governmental actions to be taken to provide more mortgage opportunities for people, possibly by means of decreasing required down payments or transaction costs.

Among all provinces, participants living in Quebec have the lowest probability of owning a house (β= -0.7343), British Columbia second lowest (β= -0.3718), and Nova Scotia third lowest (β= -0.3665). The provinces with the most likelihoods of house ownership are Alberta (β= 0, base level) and NewFoundland (β= 0.0413). Although average mortgage interest rates do not differ, the differential probabilities of house ownership seen across Canada is to be accounted for by provincial gaps in household income and house prices. According to results from Statista, among the three provinces with the least likelihood of house ownership, Nova Scotia had a much lower median annual income ($78,920) in 2018 compared to British Columbia ($87,660) and Quebec ($83,780), and the median income of these three provinces were all significantly lower than that of Alberta $101,780). The provincial gap in median annual income of households partially explains the differences seen in house ownership among these provinces. Moreover, despite the drastic gap in median income, results from Living in Canada show that the average selling price of houses in British Columbia ($736,000) ranked top in Canada, and was significantly from the second highest (Ontario, $594,000) and third highest (Alberta, $353,000), further reducing the likelihood of owning a house in that province. A reverse scenario stands true for Quebec: with an average house price similar to Alberta, it has a much lower median annual income, which has also contributed to the low likelihood of owning a house in the province. Last but not least, although the median income of Newfoundland is around the levels of British Columbia and Quebec, it is one of the cheapest provinces to buy a house in with an average house price of only $236,000, which explains why the ownership of houses in the province is the greatest in Canada.

5.1 Weaknesses and Next Steps

An obvious flaw is that no people living in the Territories were sampled in the survey. Indeed there is only less than 0.5% of the Canadian population living in the Territories. However, a comprehensive could provide valuable information about people living in remote regions. The lives in the Territories would be very different from those in the provinces as one can imagine.

Another drawback is that the gender groups only have the female and the male as options. It can not precisely reflect the demographic reality in the country. Therefore, making population predictions requires great caution.

The last one is about the analysis. Only gender, age group, and the province of residency were included in the model and post-stratification. Some variables, such as education levels, family income, and marital status, are left out of consideration because we do not have population information. A more comprehensive post-stratification matrix will benefit future studies. Besides, expert knowledge in the relevant field will help identify confounding variables and improve model performance.

6 References

6.1 References for the Report

6.2 Reference for Data Cleaning

7 Appendix

GitHub Link: https://github.com/tomsu0826/g50gssfamiliescycle31