Analysis of the association between Canadia's mental health condition and their income and religion affiliation

A strong and significant association is found between self-rated mental health and income status as well as religion affiliation

Weijia Song (1004043689), Siqi Zheng (1005065830)

October 16, 2020

Contents

Abstract 1
Introduction 1
Data 2
Model 3
Results 4
Discussion 6
References 9
Appendix A How to Download the 2017 GSS Data 10
Appendix B GitHub Link 11

Abstract

2017 Canadian General Survey data (GSS) were used to investigate the association between Canadian's self-rated mental health condition and personal income and religion affiliation because mental health is becoming a major concern in modern Canadian society. In this report, we hypothesized that individuals with higher income and religion affiliation rated their mental health condition better and we used a logistic model to examine whether a good mental health condition (good_or_not) is significantly associated with a person's income status (below_avg) and religion affiliation (religion_or_not). The results of our model supported our hypothesis to a large extent. In particular, people with religious affiliation and people with income higher than the average income in 2017 are more likely to rate their mental condition better than people with lower income and without religion affiliation.

Introduction

This report aims to investigate how Canadian's mental health condition is associated with personal income and religious affiliation. We are interested in this topic because mental health problems have been a rising issue in recent years in Canada. According to statistics, about 20% of Canadians experience mental health issues every year; and for those who are in their 40th or older, 50% reported mental health issues. This leads us to consider what factors contribute to increased mental health issue rates in Canada. Since we have a big dataset about 2017 General Social Survey (GGS) (See Appendix about how to download the dataset), the latest available GSS dataset from Statistics Canada which would contain new insights about Canadian's mental health conditions in this era, we may be able to extract useful information from the data.

Based on this dataset, we choose the variable self_rate_mental_health as our focus of this study. First, it is very unlikely, if not impossible, to collect the results of psychological tests for a large group of Canadians due to the cost and the privacy reasons, so a variable about self-rated mental health conditions can be a fair indicator of a person's mental health condition. Second, self-rated mental health is the only variable about a person's mental health in the GSS dataset, so it is also the only option for our report.

Moreover, we choose personal income as our first factor that is used to predict a person's mental health condition because lower-income may cause stress. According to research, stresses, such as unemployment, housing, and education would lead to mental health issues. Hence we would like to investigate how an individual's personal income is related to mental health issues.

Religious affiliation is our second factor that is used to predict a person's mental health condition. This is because beliefs from religions can help deal with stress; and social support, which comes from a religious group, also promotes mental health. However, religion could also exacerbate mental health issues if people are fanatic supporters of certain religions. That is, religion can shape people's minds in opposite ways. Thus, we will explore how Canadian's religious affiliation is associated with mental health using statistical analysis in our research. Therefore, our hypothesis is that people with lower income and no religious affiliation would rate their mental health condition worse than people with higher income and religious affiliation. Hence if there is strong evidence to support our hypothesis, the report may be able to provide information to the general audience in Canada and inspirations for future researchers.

In this report, we used a logistic model to examine whether a good mental health condition (good_or_not) is significantly associated with a person's income status (below_avg) and religion affiliation (religion_or_not). All these variables are binary dummy variables created from related variables in the original GSS dataset. We also evaluated this model by comparing it with alternative models. Finally, with these results, we discussed our conclusion and suggestions for future researchers.

Data

The General Social Survey (GSS) in this report was conducted in 2017, from February 2nd to November 30th. The objectives of the survey are to collect data that can reflect social trends and to provide detailed information on emerging social issues.

There are 81 variables in this data and 20602 candidates in this dataset, classified into core content and classification variables. The core content variables, such as self-well being, are useful in scheduling social policies; while the classification variables, such as age and gender, are useful in the delineation of population. However, some of the variables are self-reported, such as self-reported mental health. This can lead to bias because there are no specific scales or criteria. Nonetheless, these variables can still be good indicators of a person's well-being.

The target population, which was all individuals covered in our study, included all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. In particular, the target population excluded residents of the Yukon, Northwest Territories, and Nunavut. We believe that the choice of geographic areas is reasonable because the data still covers the majority of the population in Canada. However, the survey results may still be biased for researchers in understanding Canadian's general attitudes towards social trends and political issues without data in certain areas.

Two survey frames (where we draw information of respondents) were constructed in this survey, one was the numbers listed in use that were available to Statistic Canada, the other one was the address registered within the ten provinces.

The sampling strategy of 2017 GSS was stratified sampling, that is, they divided the 10 provinces into strata. Simple random samples without replacement in each stratum, which means that the individuals were randomly selected from each province, and one respondent would not be selected twice. This stratified random sampling can provide greater precision than other sampling strategies because the researchers have more control over the subgroups.

The survey was conducted in a form of phone calls. Interviewers made calls from 9 am to 9:30 pm Mondays to Fridays, and also on Saturdays and Sundays. All interviewers were well trained and with former interviewing experiences. Hence we believe the strength of this form of survey is that the cost would be low and it is more time-efficient. In addition, the time they chose to make the phone call was proper, which could reduce convenience bias. However, households without telephones were excluded from the survey, which might cause a problems since those without telephones might represent a kind of social characteristics. Also, the response rate was 52.4%, the low response rate might reflect another weakness of the phone interview: some people might not pick up phone calls when saw see unfamiliar numbers.

We create variables based on the existing variables for two reasons. First, we have a genuine interest in whether or not relatively low income and religious affiliation are associated with a person's self-rated mental health, so a binary variable avoids redundant information in the original variables but still allows our audience to obtain useful information from this report. Second, studies (see reference) have demonstrated that people's mental health is associated with people's income and religious affiliation, so we want to examine whether this claim holds in Canada in 2017.

To define this dependent variable, we need to look at the variable self_rated_mental_health in the original dataset first. For this variable self_rated_mental_health, there are six possible responses (excluding NA values) Excellent, Good, Very good, Poor, Fair, Don't know, NA. We consider a person to have 'good mental health condition' if this person provides one of the three positive responses, namely "Excellent", "Very good" and "Good", under self_rated_mental_health. If this person does, we encode 1 in the new dependent variable good_or_not for this person.

Conversely, if a person provides the neutral option "Fair" or the negative option "Poor" under self_rated_mental_health, the dependent variable good_or_not will be encoded as 0 for this person. We exclude the neutral "Don't know"/NA option since it is not helpful for our analysis without knowing why these people are unable/unwilling to respond during the process of data collection.

The first independent variable is below_avg, which is a binary variable that indicates whether a person's personal income is below the average income in 2017, $59800. If a person's income falls into the category 'below $25000' or '$25,000 to $49,999' under income_respondent in the GSS data, then this person's income is considered as below average income, encoded as 1. Income that falls in other categories will all be encoded as 0 in this independent variable below-avg.

The second independent variable is religion_or_not, which is also a binary variable that indicates whether a person has religious affiliation in 2017. If a person indicates clearly that he/she "has religion affiliation" under religion_has_affiliation in the GSS dataset, then this variable will be encoded as 1 for this person. Otherwise, this variable will be encoded as 0 for this person.

Here is an overview of the data:

```
##    caseid religion_or_not below_avg good_or_not
## 1       1               1         1           1
## 2       2               0         1           1
## 3       3               1         1           1
## 4       4               1         0           1
## 5       5               1         1           1
## 6       6               1         1           1
## 7       7               1         1           0
## 8       8               1         1           1
## 9       9               1         1           1
## 10     10               1         1           0
```

Table 1. The first ten rows of the GSS dataset after cleaning

Model

Logistic regression is suitable for examining the relationship between a categorical response variable and one or more categorical or continuous predictor variables. The general formula is:

$ln[\frac{p}{1-p}] = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$

where p is the probability of the occurrence of an event, β_0 is the y-intercept of this function, β_1 is the coefficient of variable 1 and β_n is the coefficient of variable n.

The log(odds) is defined by $ln[\frac{p}{(1-p)}]$ and expresses the natural logarithm of the ratio between the probability that an event will occur, $p(Y = 1)$, to the probability that it will not occur $p(Y = 0)$. The predicted probability of an event occurring and that is defined by $p = \frac{1}{1+exp^{-z}}$, where $z = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$

We select logistic model because we are interested in whether Canadians will rate their mental health condition higher than 'Fair'. This is a dichotomy, meaning we can use logistic regression to estimate the effects. As income and religion are two important factors in a person's life, we incorporate these into our model. From the perspective of sampling and survey, the original variable income_respondent is not very informative without knowing the average income and the variable religion_has_affiliation is useful but may not be straightforward for the general audience. Therefore, we construct a new variable below_avg with information about average income and religion_or_not with only 1 and 0.

Therefore, this model has three variables, one dependent variable (response variable) and two independent variables (explanatory variables). The dependent variable is the focus of this report: whether the respondent of the survey is in good mental health condition. The two independent variables will be whether a person has religious affiliation and whether a person's income is below average. All three variables are reconstructed from the existing variables in GSS data as explained under Data section.

Then, the formula for logistic regression model will be:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 \times religion\_or\_not + \beta_2 \times below\_avg$$

where p is the probability of a person in good mental health condition, β_0 is the y-intercept of this function, β_1 is the coefficient of the variable religion_or_not and β_2 is the coefficient of the variable below_avg

Our model is also adjusted for the population. We do not have information about the population living in certain areas and institutions, but a fair estimate of the target population N would be the total population of Canada in 2017 - population of people aged 0-14 in Canada in 2017 (36708083-5877081). The sample size n 20439, however, is too small compared to the population, so this adjustment is not very significant ($\frac{n}{N} << 1$). Hence we expect that a very similar model will be fitted without such adjustment i.e. such adjustment may not be necessary.

In order to construct a logistic model with software R, we use the package "tidyverse" to clean the data and create new variables. We also use the package "survey" to fit the logistic model with population adjustment. For the clarity of this report, codes are hidden. Complete codes can be found in the associated .Rmd file.

There are two possible alternative models with a single predictor. The formula for alternative model #1 is: $log(\frac{p}{1-p}) = \beta_0 + \beta_2 \times below\_avg$ i.e. an alternative model that has only below_avg as a predictor variable and the formula for alternative model #2 is: $log(\frac{p}{1-p}) = \beta_0 + \beta_1 \times religion\_or\_not$ i.e. an alternative model that has only religion_or_not as a predictor variable. The strength and weakness can only be discussed when the results are calculated. Therefore, readers can find the detailed discussion about the strengths and weaknesses of these models in the Evaluation of this Logistic Regression Model by Likelihood Ratio Test section.

Results

Summary of the Model

```
##
## Call:
## svyglm(formula = good_or_not ~ religion_or_not + below_avg, design = gss.design,
##     family = "binomial")
##
## Survey design:
```

```
## svydesign(id = ~1, data = clean_data, fpc = fpc.srs)
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.62481    0.06576  39.913   <2e-16 ***
## religion_or_not  0.47940    0.05794   8.275   <2e-16 ***
## below_avg       -0.76702    0.06130 -12.512   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1.00267)
##
## Number of Fisher Scoring iterations: 5
```

Table 2 Summary of the logistic model (R output)

Before explaining the results of the R output, it is important to be clear about our variables. Since all three variables are binary and the two independent variables are dummy variables i.e. the possible value is 0 and 1. 1 in religion_or_not indicates the person has religion affiliation and 0 indicates the opposite. Similarly, 1 in below_avg indicates the person' has religion affiliation's income is below the average income and 0 indicates the opposite.

The R output shows the values of β_0, β_1 (coefficient of religion_or_not) and β_2 (coefficient of below_avg) and the corresponding p-values, which are very crucial in our statistical analysis. Note that the sign of coefficient shows whether the independent variable is positively associated with the dependent variable, and a small p-value shows strong evidence against the null hypothesis that the independent variable has no correlation with the dependent variable.

The intercept = 2.6248134 which corresponds to the log odds for people from higher-income group without religious affiliation. Based on this value, we can calculate that the probability that a person from higher-income group without religious affiliation is in good mental health condition is 0.9324416.

Moreover, the coefficient of the predictor religion_or_not is positive, implying that (all other things equal) a person with religious affiliation has a higher chance to rate one's mental health condition good. At the same time, the coefficient of the predictor below_avg is negative, implying that (all other things equal) a person with income that is below average has a lower chance to rate one's mental health condition good, as we would expect.

Furthermore, the p-value for these three values is very low (<2e-16), so the intercept and the regression coefficients in this model are highly significant at the 5% level. Hence we have strong evidence to reject the null hypothesis that the intercept β_0 and the coefficients β_1 and β_2 are zero.

Evaluation of this Logistic Regression Model by Likelihood Ratio Test

Alternative Model #1

Our logistic regression provides a better fit to the data if it shows improvement over a model with only one predictor. We choose the likelihood ratio test, which compares the likelihood of the data under our model in the last section against the likelihood of the data under an alternative model with only one predictor. Then we can make a null hypothesis: $H_0 : \beta_1 = 0$ i.e. an alternative model that has only below_avg as a predictor variable (because the coefficient of religion_or_not is assumed to be 0 under the null hypothesis). If the p-value of this testing is smaller than 0.05, then we have strong evidence against the null hypothesis that the coefficient of religion_or_not is 0.

The formula for alternative model #1 is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_2 \times below\_avg$$

```
## Working (Rao-Scott) LRT for religion_or_not
##  in svyglm(formula = good_or_not ~ religion_or_not + below_avg, design = gss.design,
##     family = "binomial")
## Working 2logLR =  64.64187 p= 1.006e-15
## df=1
```

Table 3 Likelihood ratio test results for alternative model #1

Indeed, the p-value is very small (1.006×10^{-15}), so we have strong evidence against the null hypothesis that this alternative model is better. Therefore, the original model with two predictors is recommended compared to the alternative model #1.

Alternative Model #2

The formula for alternative model #2 is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 \times religion\_or\_not$$

This time, the null hypothesis is $H_0 : \beta_2 = 0$ i.e. an alternative model that has only religion_or_not as a predictor variable. If the p-value of this hypothesis testing is smaller than 0.05, then we have strong evidence against the null hypothesis.

```
## Working (Rao-Scott) LRT for below_avg
##  in svyglm(formula = good_or_not ~ religion_or_not + below_avg, design = gss.design,
##     family = "binomial")
## Working 2logLR =  174.4363 p= < 2.22e-16
## df=1
```

Table 4 Likelihood ratio test results for alternative model #2

Indeed, the p-value is very small ($< 2.22 \times 10^{-16}$), so we have strong evidence against the null hypothesis that this alternative model is better. Therefore, the original model with two predictors is recommended compared to the alternative model #2. These two comparisons have shown that the model with two predictors is more appropriate in this context compared to the models with a single predictor.

In conclusion, the comparisons show that the model with two predictors is indeed very appropriate in this context compared to the models with single predictor.

Discussion

The intercept is statistically significant, and it tells us that it is highly probable (0.9324416) for a person from higher-income group without religious affiliation to rate his/her mental health condition good. That is, even though religious affiliation is positively associated with self-rated mental health conditions, wealth also contributes much to a person's mental health as this person is very likely to rate his/her mental health condition good.

Furthermore, the coefficient of the predictor religion_or_not is positive and statistically significant, showing that there is a strong positive association between the probability to rate one's mental health condition good and religious affiliation. That is, people with religious affiliation tend to rate their mental health conditions better than people without religious affiliation.

At the same time, the coefficient of the predictor below_avg is negative and statistically significant, showing that there is a strong negative association between the probability to rate one's mental health condition good and the lower-income condition. In other words, people with relatively lower income tend to rate their mental health conditions worse than people with relatively higher incomes.

To a large extent this evidence supports our hypothesis that people with lower income and no religious affiliation would rate their mental health condition worse than people with higher income and religious affiliation. If this survey is representative, we are also able to conclude that Canadians with higher income (above average) and religious affiliation tend to rate their mental health conditions better, which achieves the original goal of the study specified in the Introduction.

Weaknesses

Even though our results from the model provides evidence for our hypothesis, the limitations of our study are also apparent when we create new variables and simplify our model based on the 2017 GSS data.

The first limitation was the mismatch between the target population of the survey and the objective of our study. The GSS dataset excluded institutionalized people and people in some areas of Canada. As a result, some patterns of these institutionalized people were overlooked in this dataset. For instance, people who were in jail/in the hospital may had worse mental health conditions even if they had religious affiliation. However, we assumed that the sample was representative of all Canadians (age 15 or older) and we were unable to examine this possibility given this 2017 GSS dataset.

The design of the options for the variable self_rated_mental_health (3 positive, 2 neutral, 1 negative) in the original GSS dataset could be misleading for respondents as well. This was because people might choose to respond in a less negative way given that most options were non-negative. Consequently, this bias might still exist when we created variable good_or_not based on this existing variable.

Moreover, the original 2017 GSS is an observational study, so our report can only identify association instead of causal relationship. It is difficult to conduct experiments for this topic, so it is extremely hard to establish causal relationship with GSS data.

Also, we assumed the self-related mental health could indicate a person's actual mental health condition to some extent. Nevertheless, people might conform to social practice when they responded to questions about their mental health state, especially people with high income and high social status. Therefore, there might be conformity bias in the variable self_rated_mental_health.

Furthermore, the methods we used to create new variables could have some limitations. For example, it might not be appropriate to use average income as a universal standard to separate two income groups because the same amount of income had different purchasing power in different areas. Moreover, constructing a new binary variable might overlook the difference between each category under the original variable in GSS data. As a result, we were taking a risk to oversimplify the model.

Another risk of oversimplification was that NA values/"Don't know" were simply discarded in our study because the number of such responses was limited. However, these NA values might have some patterns and characteristics of a certain group of people.

In summary, the design of the 2017 GSS, the way to simplify a logistic model and the methods to create new variables deserve further investigation and examination. Thus, we suggest some possible improvements in the next session.

Next Steps

To better address the first limitation in the previous session, future researchers should conduct a survey for people who are underrepresented, including people from Yukon, Northwest Territories, and Nunavut so that the results are more representative.

Another feasible approach to deal with the limitation of the design is to create a balanced set of options for each variable in the survey. For example, there can be a balanced number of options (2 positive, 2 neutral and 2 negative) for the variable self_rated_mental_health instead of options in favor of positive mental health conditions. This way of design can also reduce the convenience bias in the survey.

Third, researchers who are interested in the mental health conditions of individuals can conduct follow-up surveys about the mental conditions of the same group of people in the upcoming years to test whether the results of this model are consistent. Also, researchers on this issue should study the reasons for the missing values and study what roles these missing values play in this survey. If future researchers are interested in the difference in the association between mental health condition and income among provinces, they can use other criteria to separate two income groups. For example, they can calculate an average income for each region based on its price level (CPI index).

If researchers have access to the detailed information about what religion each individual belongs to in 2020, then they can fit a better model with up-to-date data in 2020. A more comprehensive and up-to-date model may be more practically significant given the current COIVD-19 pandemic.

References

1. Canadian Mental Health Association (2007, November). Poverty and Mental Illness. Retrieved October 17, 2020, from https://ontario.cmha.ca/documents/poverty-and-mental-illness/
2. The Centre for Addiction and Mental Health (2020). Mental Illness and Addiction: Facts and Statistics. (n.d.). Retrieved October 17, 2020, from https://www.camh.ca/en/driving-change/the-crisis-is-real/mental-health-statistics
3. Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
4. Mathew Analytics (2015, August 18). Evaluating Logistic Regression Models. Retrieved October 18, 2020, from https://www.r-bloggers.com/2015/08/evaluating-logistic-regression-models/
5. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
6. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
7. Statistics Canada (2020). General social survey on Family (cycle 31), 2017. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31
8. Statistics Canada (2018, April 30). Annual Demographic Estimates: Canada, Provinces and Territories, 2017. Retrieved October 11, 2020, from https://www150.statcan.gc.ca/n1/pub/91-215-x/91-215-x2017000-eng.htm
9. T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.

Appendix A How to Download the 2017 GSS Data

The 2017 GSS data is obtained from the U of T library. That data is available to U of T students, but it needs to be put into a tidy format before it can be analyzed. The codes for data cleaning can be customized based on the needs of researchers.

The main issue of this dataset is that the data are released with codes for variables, whereas, we want the variable. e.g. sex is 1 or 2, but we want sex is female or male. So we create a dictionary type dataset that has the variable names and their possible values. In that we embed some R code that will do a replacement. We then apply that dataset to the raw dataset. Finally we do all the usual cleaning to the dataset.

A detailed procedure is provided by the professors: Go to: http://www.chass.utoronto.ca/ 2. Data centre –> UofT users or http://dc.chass.utoronto.ca/myaccess.html 3. Click SDA @ CHASS, should redirect to sign in. Sign in. 4. Continue in English 5. Crtl F GSS, click 6. Click "Data" on the one you want. We used 2017, but you may want a different wave. In particular the General Social Survey on social identity (cycle 27) 7. Click download 8. Select CSV data file, data definitions for STATA (gross, but stick with it for now). 9. Can select all variables by clicking button next to green colored "All". Then continue. 10. Create the files, download and save.

For the original codes for data cleaning, you may want to contact `rohan.alexander@utoronto.ca`

Appendix B GitHub Link

The link to the associated GitHub repo: https://github.com/siqi-zheng/STA304-PROBLEM-SET-2