Religion Affiliation status of Canadian and the Variables Affecting it

Hanrui Dou, Hanjing Huang, Hairuo Wang, Xuan Zhong - Group 161

October 19 2020

Abstract

In this whole project, we used the 2017 General Social Survey(GSS) on the Family to analyze the relationship between people's mental health and whether they have religious beliefs. This survey is a sample survey with cross-sectional design, It collected a large amount of data for each selected responder. In this study, we used the logistic model. We found that mental health has no influence with whether there is religious belief, while total children, people's satisfaction with their life and the importance of religion have a certain relationship with whether they have religious belief.

Introduction

Nowadays, many people have a lot of interest in religion, and some people believe that whether they have a religious belief is affected by many factors. Our group chose to study this topic because of religion. Our group is very interested in the influence of four different variables on religious beliefs. The four variables are mental health, the influence of the number of children in each family, people's satisfaction with life, and the importance of religion. We hypothesize that these four variables have an impact on whether there is a religious belief, and we guess that the two variables, mental health, and importance to religion, have the greatest impact. In this project, we used the logistic model, because we found that most of the variables in the dataset are categorical variables, and we want to see the relationship between if the respondents have a religious affiliation and other variables. In addition, readers can learn from this project, facts were found to be contrary to common things.

Data

The data of 2017 General Social Survey was collected from February 2nd to November 30th 2017, and the number of respondents was 20602 which is more than the expected number of target sample size.

The data consists of core data, which related to characteristics of families of responsers, and classification data, such as sex, education and geographic area and classficication data could classfied the different population group and analysis with core data.

The target population of 2017 General Social Survey is all of the persons 15 years of age and more in Canada without residents in following three geographic area, Yukon, Northwest Territories and Nunavut; and full-time resident of institution.

From 2013, General Social Survey had a new frame, which collect data via telephone number in use available to Statistic Canada. As well, the survey will collect data from all dwellings within the ten provinces. Thus, all of the household in ten provinces with telephone number are covered by the survey frame.

For the survey, the telephone interview would be scheduled everyday while February 2nd to November 30th 2017. If there are some respondents who refused to take the survey and the interviewer would explain the importance and encourage them to take the survey. If the respondents are not convenient to take telephone interview, the interviewer could make appointment for them or numerous call backs.

At the beginning of the survey, the survey used simple random sample without replacement since they randomly choose respondent from each family. After that, the survey divided each province into different strata which is stratified sampling. If the respondents are not eligible to take the survey, such as the respondents are not 15 years old, the interviewers would terminate the survey after inital questions. As well, the household without available household in ten province would not considered in the survey sample. As we mentioned above, the number of respondents of the survey was 20602. Thus the survey sample is 20602 household in ten provinces.

Although the survey received more number of respondents than the expected number of target sample size, there exists sampling error for the survey that was mainly formed from non-response to partial question or entire survey. The non-response would effect the results of the survey. Thus, non-response will be adjusted the weight of the respondent who responsed to the survey and that could balance effect of non-response.

Overall, the cost of General Social Survey is high. Since there are ton of interviewers with professional training to collect data with the household in ten provinces. As well, the survey continuouly held for 10 monthes. Although the cost of the survey is high, it is important and necessary to collect the data about the society and analysis the data to solve the social issues. On the other hands, the survey has new frame which could collect data via telephone that also reduced the transport cost and labour cost of the survey since the interviewers do not need to contact the respondent in person and also save a lot of time.

For the questionnaire, it covered the most characteristics of the family and the variables are very detailed. It is sufficient to analysis the relationship with variables from different angle. However, while we analysis the data, we found that some variables represent the similar meaning to the survey. For example, the number of marriage and ever marriage. The similar questions would increase the correlation between different variables in the model, and at least one of them supposed to be dropped. As well, the number of non-response like N.A for some question are too large, in our opinion, the questions might be hard for respondents to answer.

For conclusion, the target population of the data covered the most eligible persons in Canada and the sample size is sufficient, such that the data is useful to analysis the issues about family in Canada. The questionnaire is cleared and detailed from different characteristics of the family. Furthermore, non-response error could be solved by adjusting the weight of household who response to the survey that could avoid the effect of non-response error to the data.

On the other hand, the number of non-response for some question are extremly high that leads to some variables are useless and that could not analysis with other varialbes. Meanwhile, there exist a few variables with similar meaning and some of them should be dropped from the data. Because of this, some people may feel the survey is too long and some questions are unnecessary while taking it, and is will make the survey unreasonable. Furthermore, some questions in this survey such as "Satisfied time with children" are less representative and hard to quantified.

Model

Since we found most of the variables in the dataset is categorical variables, and we want to see the relationship between if the respondents have religion affiliation and other variables. Thus, we choose logistic regression, since religion affiliation is a binary variable(If we do not consider option "Don't know" and "NA") and other variables are either categorical variables or numberic variables. Furthermore, logistic regression fits the dataset more than the other regression.

If we use linear regression to model the binary variable, the resulting model might not restrict the predicted Ys within 0 (No religious affiliation) and 1(Has religious affiliation). Besides, other assumptions of linear regression such as normality of error may get violated. So instead, we model the log odds of the event , where P is the probability of the event. However, if we use linear regression, it is easier for us to understand the result than using logistic regression model since we need to do a complicated calculation to get the probability.

We would rather use feelings_life than feelings_life-groups since the numerical feelings_life provides a index, which each value represents a level of personal satisfaction of respondents life. It cannot be categorized as A,B,C,D.

We would rather use children than children-groups since the numerical children provides a index, which each value represents a child. Each increase in the number of children varies widely. It cannot be categorized as A,B,C,D since using children-groups would tend to cluster the collected data too tightly, making it impossible to accurately analyze the distribution.

Considering why we rather take the variables regilion_importance and self_rated_mental_health to be categorical than numerical, categorical variables would build a clear visualization for each situation analysis. It cannot be numerized as index.

This is a regression equation of logistic regression:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 X_{totalchildren,i} + \beta_2 X_{feelingslife,i} + \beta_3 X_{regilionimportance,i} + \beta_4 X_{selfratedmentalhealth,i}$$

pi is the probability of the respondents have religion affiliation.

beta0 is the intercept of the model which represents all the variables equal to zero, then the value of log odds.

beta1 coefficient represents change in log odds for every one child increase in total children.

beta2 coefficient represents change in log odds for every one unit increase in feeling life.

beta3 coefficient represents average change in log odds between the category which the options of religion importance = 0 and the category which option of religion importance = 1

beta4 coefficient represents average change in log odds between the category which the options of self rated mental health = 0 and the cetegory which option of mental health = 1.

Table 1: Summary of Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.325	0.492	0.660	0.509
total_children	0.215	0.015	14.287	0.000
feelings_life	0.032	0.014	2.283	0.022
as.factor(regilion_importance)Not at all important	-0.987	0.171	-5.769	0.000
as.factor(regilion_importance)Not very important	0.431	0.173	2.484	0.013
as.factor(regilion_importance)Somewhat important	1.396	0.174	8.043	0.000
as.factor(regilion_importance)Very important	1.560	0.174	8.967	0.000
as.factor(self_rated_mental_health)Excellent	0.121	0.457	0.265	0.791
as.factor(self_rated_mental_health)Fair	-0.372	0.462	-0.806	0.420
as.factor(self_rated_mental_health)Good	-0.109	0.457	-0.239	0.811
as.factor(self_rated_mental_health)Poor	-0.609	0.478	-1.273	0.203
as.factor(self_rated_mental_health)Very good	-0.049	0.457	-0.106	0.915

Table 1 shows that the coefficients of each variables in the model.

Model Diagnostic

```
##                                         GVIF Df GVIF^(1/(2*Df))
## total_children                      1.018895  1        1.009403
## feelings_life                       1.357583  1        1.165154
## as.factor(regilion_importance)      1.023371  4        1.002892
## as.factor(self_rated_mental_health) 1.365539  5        1.031645
```

Similar with the linear regression model as we learnt in STA302, we need to check if the variables in the model have multicollinearity Therefore, we check the VIF of variables of the model. As seen above, there is no variables with VIF greater than 4, thus all variables are independent and not multicollinearity.

Results

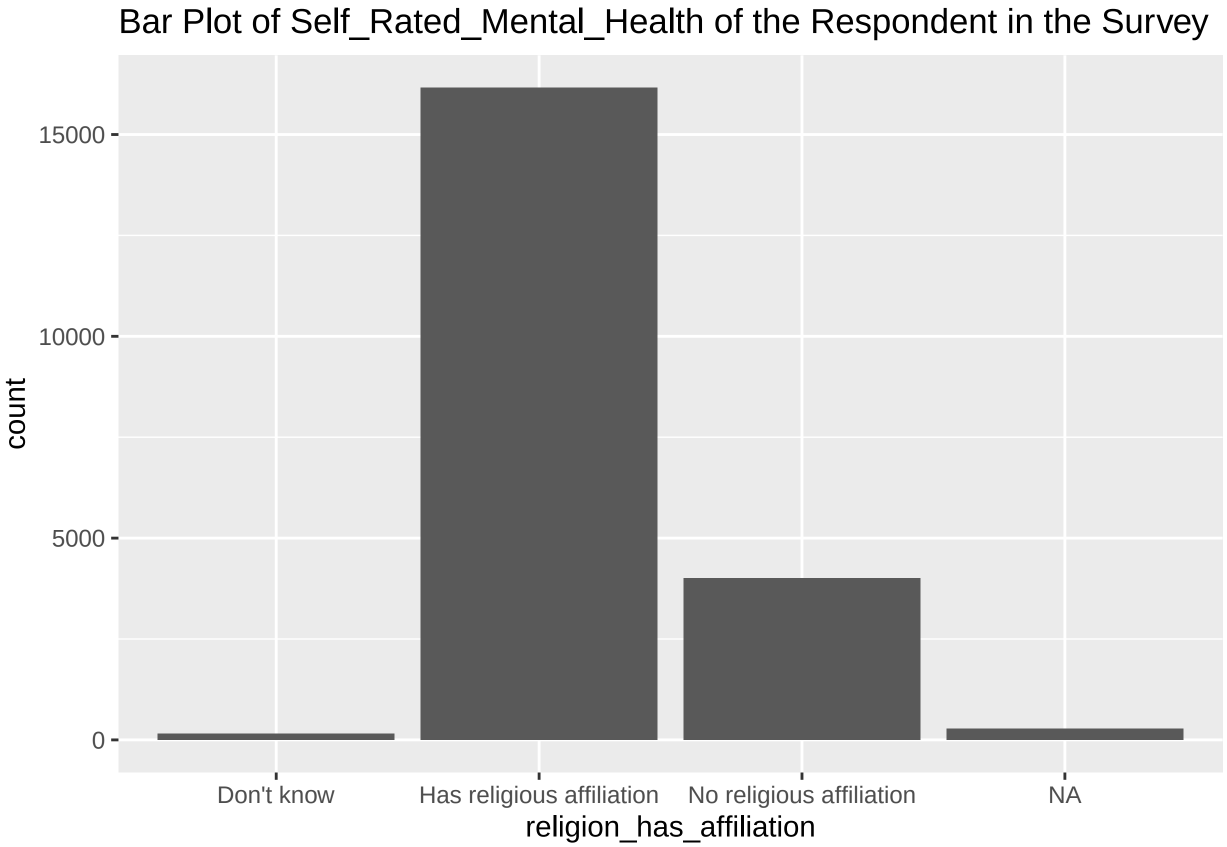

Figure 1 shows that the distribution of if the respondents have religious affiliation. Base on Figure 1, it shows that the most respondents have religious affiliation and the respondents who do not have religious affiliation take less proportion in the population. Meanwhile, N/A does not take large proportion and the non-response bias will not effect the result of this variable. We will check how the varialbes effect the respondents have religious affiliation or not.

```
## Warning: Removed 271 rows containing non-finite values (stat_boxplot).
```

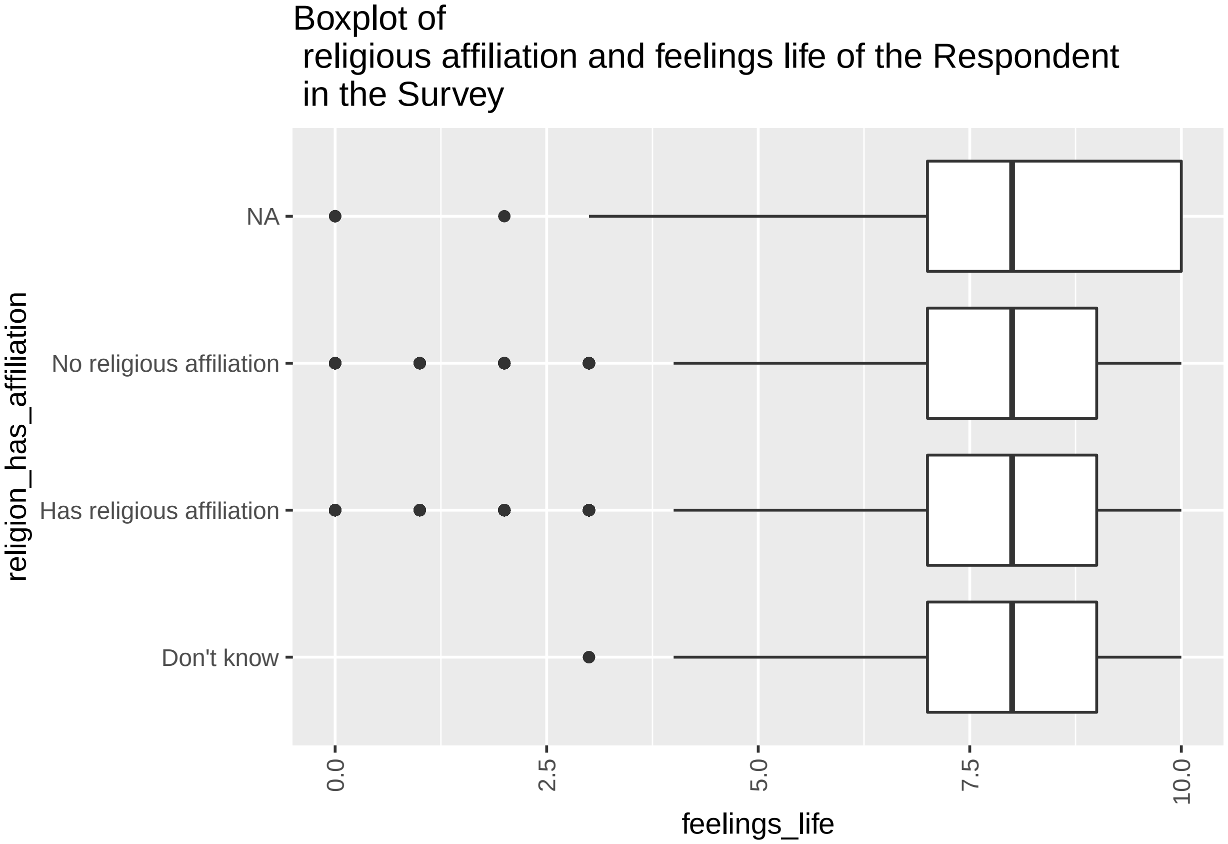

Figure 2 shows that the relationship between if the respondents have religious affiliation and feelings life of the respondents.

Base on Figure 2, this boxplot shows the relation between the religion has affiliation and feeling life. We can see that the mean of people who has religious affiliations and no religious affiliations are almost the same, which means people feeling their life good or bad are not related to whether they affiliation of religion. However, there are some outliers in this plot, it means that there are some people who have religious or have not religious both feel their life is not satisfied. Therefore, according to this plot, there are no relation between religion has affiliation and feeling life.

```
## Warning: Removed 19 rows containing non-finite values (stat_boxplot).
```

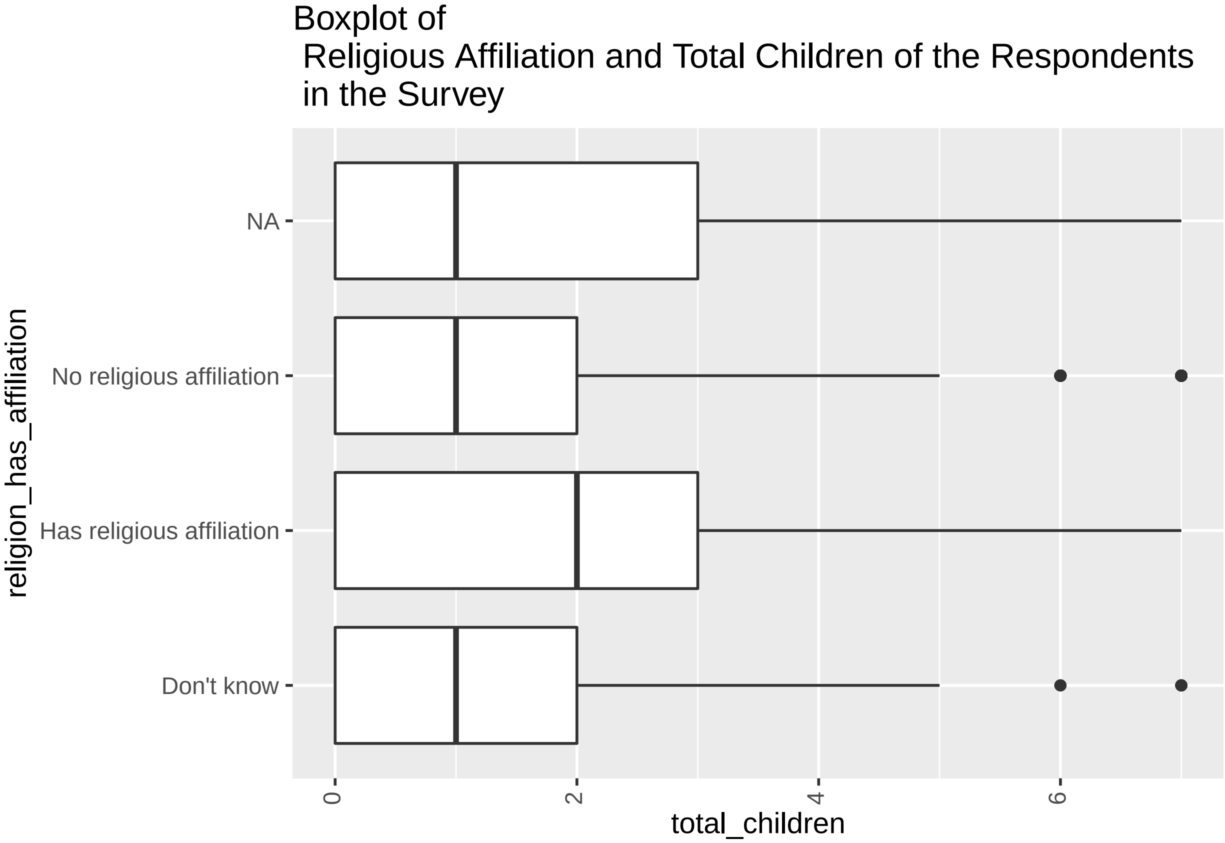

Figure 3 displays a distribution of the relationship between the the total number of children that respondents have and if they have religion affiliation.

According to Figure 3, the boxplot clearly visualizes that the distribution of the amount of children and respondents who has no religious affiliation is almost as same as respondents who don't know if they have religious affiliation. We can also observe that despite respondents who has no religious affiliation has outliers, the median amount of children for respondents has religious affiliation is higher than respondents who has no religious affiliation. It means that in average respondents who have more children would have religious affiliation.

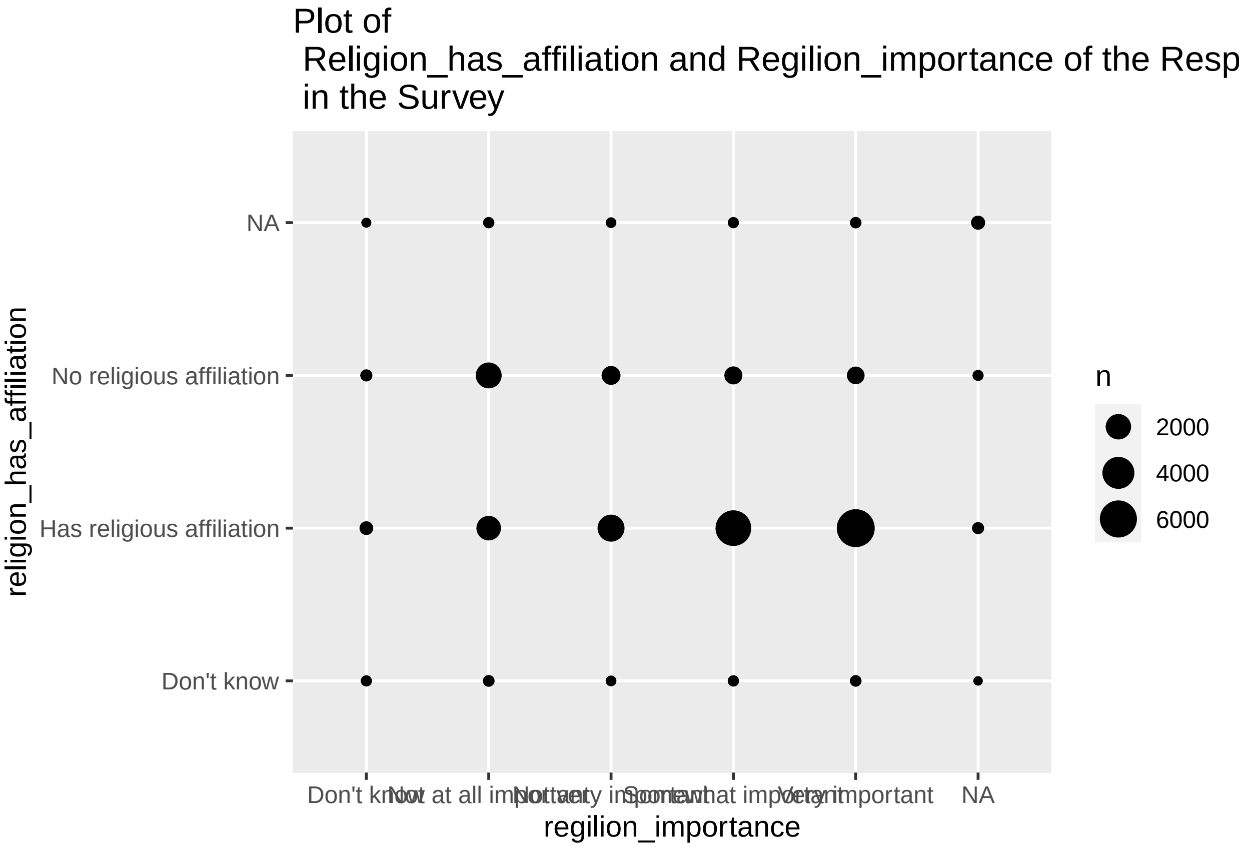

Figure 4 shows that the relationship between the religion_has_affiliation and the religion_importance.

Based on the graph, we can see that most people who has religious affiliation think religion is "Somewhat important" or "Very important", and people who have no religious affiliation are mostly think religion is "Not at all important". And it's clear in the plot that people who have religious beliefs may don't think religion is important.

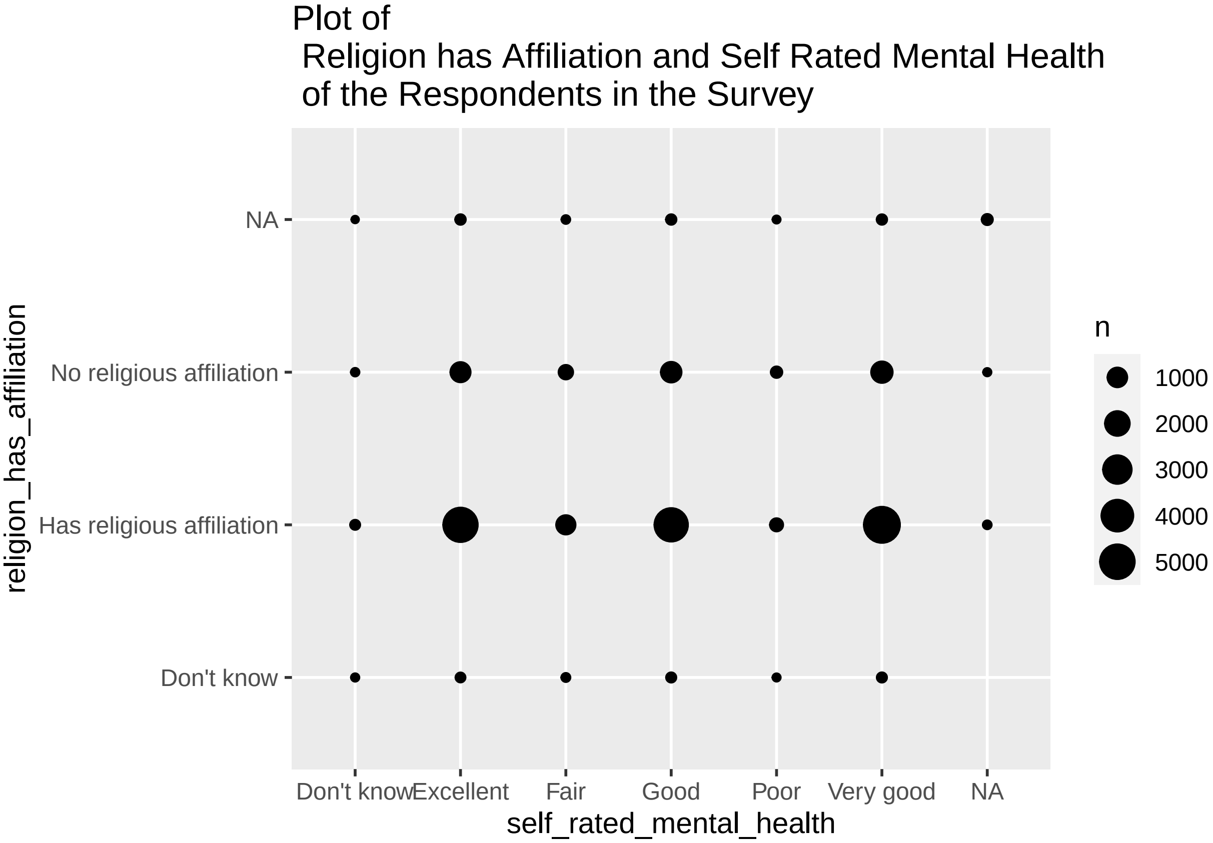

Figure 5 shows that the relationship between the self rated mental health of the respondent and if the respondent has religious affiliation.

Base on the figure 5, for respondent whose self rated mental health is positive, we could see the the number of person have religious affiliation is more than the people who do not have religious affiliation. On the other hands, for the respondent whose self rated mental health is negative, the number of person have religious affiliation is similar with the number of person do not have religious affiliation.

Discussion

Recall Table 1 for table of coefficient of the variable of model.

Table 2: Summary of Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.325	0.492	0.660	0.509
total_children	0.215	0.015	14.287	0.000
feelings_life	0.032	0.014	2.283	0.022
as.factor(regilion_importance)Not at all important	-0.987	0.171	-5.769	0.000
as.factor(regilion_importance)Not very important	0.431	0.173	2.484	0.013
as.factor(regilion_importance)Somewhat important	1.396	0.174	8.043	0.000
as.factor(regilion_importance)Very important	1.560	0.174	8.967	0.000
as.factor(self_rated_mental_health)Excellent	0.121	0.457	0.265	0.791
as.factor(self_rated_mental_health)Fair	-0.372	0.462	-0.806	0.420
as.factor(self_rated_mental_health)Good	-0.109	0.457	-0.239	0.811
as.factor(self_rated_mental_health)Poor	-0.609	0.478	-1.273	0.203
as.factor(self_rated_mental_health)Very good	-0.049	0.457	-0.106	0.915

1. For the relationship between people's satisfaction with their life and if the respondents have a religious affiliation.

From the data, we can analyze for every additional unit increase in respondents rated their people's satisfaction with their life we expect the log odd of having religious affiliation to increase by 0.031 and all the other variables are constant. It is noteworthy that the meaning of this change in this variable does n have a influence on y. As we could see in Figure 2, the variable of people's satisfaction with their life does change about whether they affiliation with religion. Therefore, The self-evaluation of people's satisfaction with their life could lead more people to have religious affliction.

2. Looking at the relationship between total amount of children that the respondents have and whether the respondents have religion affiliation, we can see that for every additional unit increase in total children we expect the log odd of having religion affiliation to increase by 0.215 and all the other variables are constant. It shows that the estimated parameter total children slightly influence respond variable religion_has_affiliation in a positive way. According to Figure 3, respondents who have more children tend to have religion affiliation compared to who have no religion affiliation. The amount of children has a small positive influence on whether the respondents have religion affiliation or not.

3. About the relationship between religion_has_affiliation and regilion_importance, based on what we got from our model, we know that:

- For every unit increase in people who rate the regilion_importance as "Not at all important", we expect the log odd of having religion affiliation to decrease by about 0.987 with all the other variables are constant.
- For every unit increase in people who rate the regilion_importance as "Not very important", we expect the log odd of having religion affiliation to increase by about 0.431 with all the other variables are constant.
- For every unit increase in people who rate the regilion_importance as "Somewhat important", we expect the log odd of having religion affiliation to increase by about 1.396 with all the other variables are constant.
- For every unit increase in people who rate the regilion_importance as "Very important", we expect the log odd of having religion affiliation to increase by about 1.560 with all the other variables are constant.
- If we compare all variables with regilion_importance, we can see that the average of the 4 estimated parameter of it are the highest among all the variables, which means it has the strongest relationship with our respond variable religion_has_affiliation. Overall interpret of regilion_importance, we can see that increase in every choice of "Not very important", "Somewhat important" and "Very important" will increase more probability of having religious affiliation. As we could see in Figure 4, the option with "Not at all important"(with smallest estimated parameter -0.987 in regilion_importance) could lead to least people who have religious affiliation.

4. For relationship between self rated mental health and if the respondents have religious affiliation.

Therefore, for every additional unit increase in respondent rated their mental health excellent we expect the log odd of having religion affiliation to increase by 0.121 and all the other variables are constant.

For every additional unit increase in respondent rated their mental health fair we expect the log odd of having religion affiliation to decrease by 0.372 and all the other variables are constant.

For every additional unit increase in respondent rated their mental health good, we expect the log odd of having religion offiliation to decrease by 0.109 and all the other variables are constant.

For every additional unit increase in respondent rated their mental health poor, we expect the log odd of having religion affiliation to decrease by 0.609 and all other variables are constant.

For every additional unit increase in respondent rated their mentral health very good, we expect the log odd of having religion affiliation to decrease by 0.049 and all other variables are constant.

Overall interpret of self rated mental health variable, only ‘Excellent’ variable has positively effects to dependence variable and the other variables have negatively effects to dependence variable. The variables of ‘Excellent’ have small effect to if respondents have religious affiliation and ‘Good’ and ‘Very Good’ variables have less negatively effects, which is similar with Figure 5, there are more number of respondents whose self-rated mental health is excellent. As well, ‘Fair’ and ‘Poor’ variables have larger negatively effect to dependent variable. As it shown in Figure 5, there are no obvious difference between number of respondent who have regious affiliation and no regious affiliation.

- Bias and what we do to adjust for that:

The response variable we choose (religion_has_affiliation) has 4 different responds:“Don’t know”, “No religious affiliation”, “Has religious affiliation” and “NA”, and we know that including “Don’t know” and “NA” may cause non-response error. In this case, we want to consider the response variable as binary, so what we do is transfer “Don’t know” to “NA”, then while we do the modeling, “NA” will be ignored and so our response variable will be treated as a binary variable with 1 represents “Has religious affiliation” and 0 represents “No religious affiliation”.

- What the world will know after reading our report:

Despite what we summarized above, there are also some fun facts found from our research:

1. Some people may think that poor mental healthy people will more likely to believe in religion since they need to rely on something like spiritual ballast. However, our research tells us that poor mental healthy people are less likely to have a religious affiliation than mentally healthy people.
2. A certain amount of people who have religious affiliation do not think religion is that important.

3.Although your feelings for life are better when you have a religious affiliation, but actually, they have a very weak relationship.

4. People with more children are more likely to have a religious affiliation.

Weaknesses

For the weaknesses of our study and data, there are two main parts. Firstly, among all the variables in the dataset, there may be other variables that are more related to our response variable, but we did not use them. The variable we choose such as feelings_life has a very weak relationship with the response variable, so there are likely to exist a better choice of the variable. Secondly, look at the box plots Figure 2 and Figure 3, there exist some outliers in our model for some variable, and this may affect our research result in a bad way.

Therefore, we need to select the variables to build data and analysis more carefully. We need to find some indexes and method, such as AIC and BIC for linear regression model, to select model and variables to build the model with more unbiased result for analysing.

Next Steps

After we get the result from dataset of general survey, then we supposed to have a specific survey and get data which focus on religious status of Canadian.

For the weaknesses I illustrate above, I think we can test all possible relating variables by putting them into our model in R before we analyze so that we can find the most correlated (with response variable) variables and analysis them.

About the survey, I think we can delete some unnecessary questions, especially questions with almost all answers being “NA”, such as reason_no_time_off_birth with 98% “NA” answer, or “main_activity” which got all answers being “NA”.

References

- R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

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

- John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.28.

- Prabhakaran Selva. 'Logistic Regression.' R-statistic, 2016-2017, http://r-statistics.co/Logistic-Regression-With-R.html

- Pruim R. 'Mathematics in R Markdown.' October 19, 2016, https://rpruim.github.io/s341/S19/from-class/MathinRmd.html

- "General Social Survey, Cycle 31: 2017: Family." Microdata Analysis and Subsetting with SDA, 2017, https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31, October 19, 2020.

- "General Social Survey Cycle 31: Families Public Use Microdata File Documentation and User's Guide." Microdata Analysis and Subsetting with SDA, April 2020. October 19, 2020.