Estimating the probability of Workplace Harassment and Violence in Canada using Multilevel Regression and Post-Stratification

Dec.20th 2020
Author: Yueyang Ji
Keywords: logistic regression; post-stratification; forecast
Github link: https://github.com/vegaqwj/STA304/tree/main/final

Abstract:

Firstly, we perform exploratory data analysis on the dataset. Then, we clean and prepare the data for modeling. Next, we use logistic regression to model the history proportion of the participants reported to have encountered the Workplace Harassment and Violence. We then perform the post-stratification analysis based on this model to predict the future proportion.

1. Introduction

Everyone should be able to work in a safe and healthy workplace. However, the below are some related statistics provided by Statistics Canada[1], and we can see that no matter what your gender is, you may encounter harassment and violence in the workplace with a high probability.

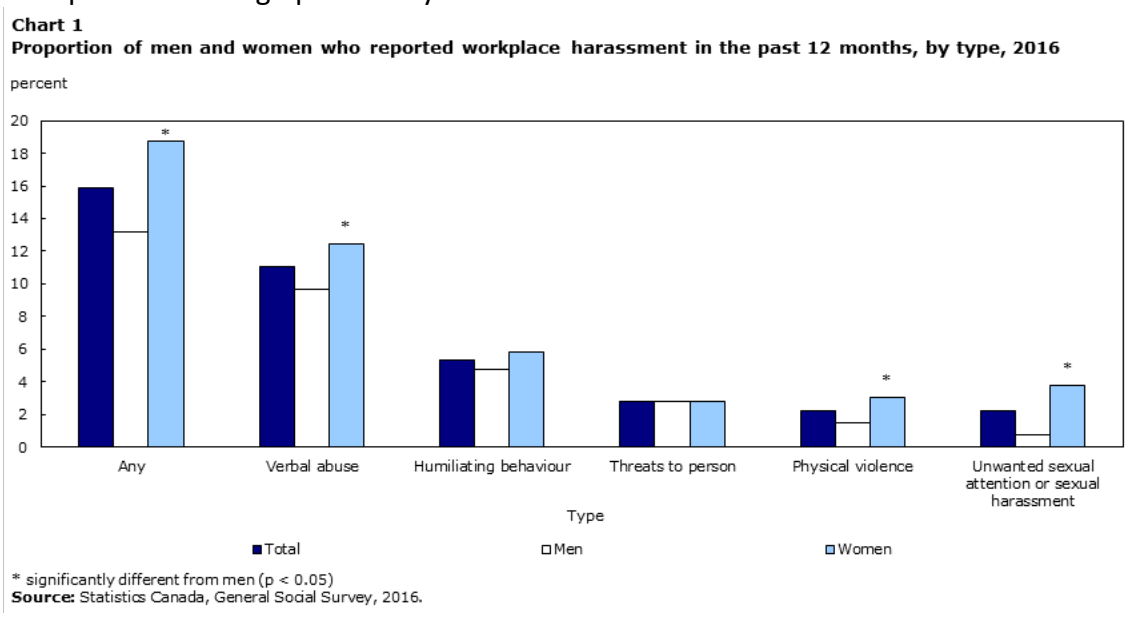

Chart 1

However, many still don't believe workplace harassment and violence exist in our daily life. As we can see from the below[2], 97% CEOs do not think workplace harassment is an issue.

[1] General Social Survey (GSS), Statistics Canada, 2016

[2] InfoGraphic: Workplace Harassment – The Unseen Issue? https://www.safetyvantage.com/workplace_violence-unseen_issue/

Chart 2 From https://www.safetyvantage.com/workplace_violence-unseen_issue/

To further address the importance of this issue, we will take a look at the Workplace Harassment and Violence Survey, 2017, and use post-stratification with General Social Survey data to estimate the probability of a Canadian encountering workplace harassment and violence.

2. Methodology

a. Data

The survey dataset we are going to use is Workplace Harassment and Violence Survey, 2017 from Government of Canada Open data[3]. The original dataset is quite messy. Therefore, I choose to use Excel to clean the dataset, and select the data I want. I only select those participants who have completed the survey along with their response I want: whether there exists harassment in his/her workplace, gender, age, education, income and employment status. I will do some exploratory data analysis before moving to the modeling stage.

The below is the gender component in this dataset. It seems like female participate in this survey more than male. This is understandable. Since female tend to encounter more harassment than male, female will pay more attention to Workplace Harassment and Violence problems than any other gender.

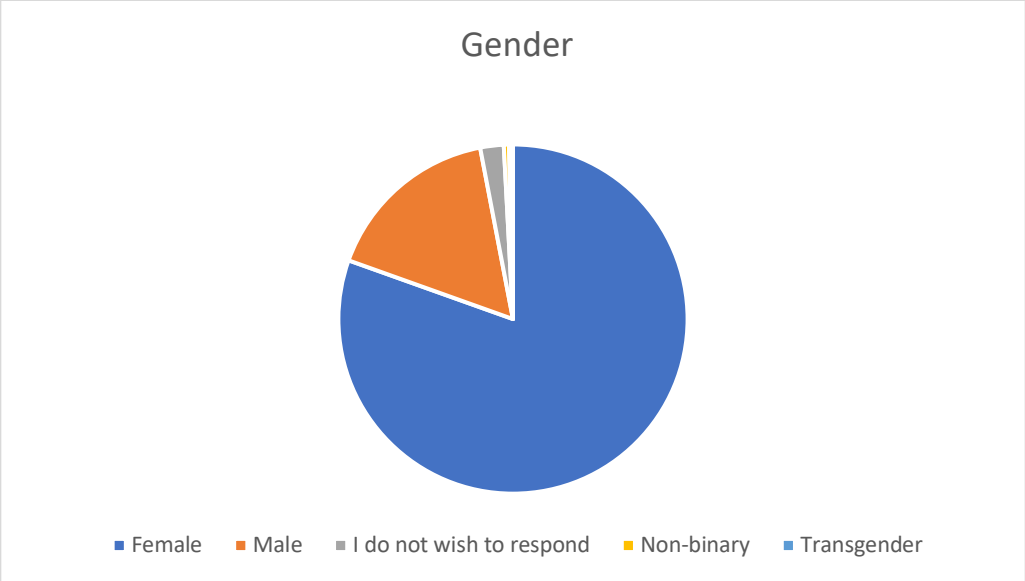

Chart3

[3] Workplace Harassment and Violence Survey, 2017, Government of Canada Open data, https://open.canada.ca/data/en/dataset/e32fed6f-9d5a-4823-93b8-5a07292935e1

The below is the pie chart of whether the participant encounter the Workplace Harassment and Violence. And we can see from the graph, most of the participants reported to have encountered the Workplace Harassment and Violence.

Chart 4

b. Model

I will be using a logistic regression model to model the proportion of the participants reported to have encountered the Workplace Harassment and Violence. I will be using age, which is recorded as a numeric variable, gender, education, income and employment status, which are recorded as categorical variables, to model the probability of encountering the Workplace Harassment and Violence. The logistic regression model I am using is:

$$\log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0 + \beta_1\ \mathrm{x}_{\mathrm{age}} + \beta_2\ \mathrm{x}_{\mathrm{gender}} + \beta_3\ \mathrm{x}_{\mathrm{education}} + \beta_4\ \mathrm{x}_{\mathrm{income}} + \beta_5\ \mathrm{x}_{\mathrm{employmentStatus}} + \epsilon$$

Where p represents the proportion of the participants reported to have encountered the Workplace Harassment and Violence. β_0 represents the intercept of the model. Additionally, $\beta_1,\ \beta_2, \beta_3, \beta_4, \beta_5$ represent the coefficients of each corresponding variable. So, for everyone one unit increase in age, we expect a e^{β_1}increase in the odds ($\frac{p}{1-p}$).

Similarly, for the categorical variables, since we use dummy coding for the category variables, one unit change (0 to 1 or 1 to 0) will increase the odds by exponential of the corresponding coefficient.

Then, in order to estimate the proportion of the participants reported to have encountered the Workplace Harassment and Violence, I need to perform a post-stratification analysis using census data[1]. Post-stratification involves adjusting the sampling weights so that they sum to the population sizes within each post-stratum. Here I create cells based off different ages, genders, education, income and employment status. Using the model described in the previous sub-section I will estimate the proportion of voters in each age bin. I will then weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size. In the end, I should get an estimated prediction of the percentage of people encountering the Workplace Harassment and Violence with decreased bias.

3. Results

From the output of our MRP model, I estimate that the proportion of the participants reported to have encountered the Workplace Harassment and Violence to be 0.6787578. I used a logistic model with age, gender, education, income and employment status as variables to perform a post-stratification analysis with 1977 subgroups. The mathematical theory behind this post-stratification analysis is demonstrated as the equation below:

$$\widehat{y^{PS}} = \frac{\sum N_j \widehat{y}_j}{\sum N_j}$$

where $\widehat{y^{PS}}$ is the proportion of the participants reported to have encountered the Workplace Harassment and Violence.

If we look at the logistics regression model, we could find something interesting. The p-values for some variables, such as age and income, are relatively large. This indicates the Workplace Harassment and Violence may not very related to your age and income. Also, this may indicate that the regression model can be improved.

4. Discussion

Firstly, from statistics point of view, this model definitely can definitely be improved. Few non-significant predictors may be removed. More predictors can be added.

Back to the main topic, with a prediction of 68 precent(0.6787578) of Canadians may encounter the Workplace Harassment and Violence, we have the confidence to say that this issue is very serious, and everyone should do something about it. The government needs to set up several regulations regarding this issue, the company should comply

with the regulation and provide necessary training to the employees, the individuals should be aware of the situation and equip relative knowledge.

References:

1. General Social Survey (GSS), Statistics Canada, 2016
2. InfoGraphic: Workplace Harassment – The Unseen Issue? https://www.safetyvantage.com/workplace_violence-unseen_issue/
3. Workplace Harassment and Violence Survey, 2017, Government of Canada Open data, https://open.canada.ca/data/en/dataset/e32fed6f-9d5a-4823-93b8-5a07292935e1