Modeling Social Unrest in Canada

Matthew Neba

April 1, 2024

1 Introduction

The goal of this project is to choose an appropriate generalized linear model(GLM) to capture the relationship between some independent variables and a single dependent variable. The independent variables are: year, month, Canadian provinces and population. The dependent variable is the number of protests to have been recorded. Once the GLM has been chosen, regression would be performed to obtain the parameters of the model.

Once the parameters of the model have been determined, bootstrap methods will be used to construct a confidence interval for the model parameters. This will help determine the significance of the model parameters.

Finally, the model parameters would then be used to simulate Monte Carlo protests for the year of 2025.

2 Model Selection

The number of protests that can occur is a count. Therefore, an appropriate model would be able to model count data. Some natural model choices would be a Poisson, negative binomial or a zero inflated model. Further analysis was performed on the data to determine a model to use. During analysis of the data, It was discovered that approximately 15 % of the data points were zero. This suggested that some zero-inflated model would likely be a good choice. However, over dispersion of count data can possibly to occur. Since a Poisson distribution assumes:

$$E[x] = \text{Var}[x],$$

The negative binomial model would be an excellent choice since it can account for this over dispersion of the data. Here is the relationship between the expected value and variance for the negative binomial distribution:

$$E[x] = \mu, \quad \text{Var}[x] = \mu + \alpha\mu^2$$

where α represents the over dispersion parameter in the nb2 parametrization of the negative binomial distribution. Since this parameter can be 0, it also accounts for the case where there is no over dispersion present in the data.

Optimally, A zero inflated negative binomial model would have been used. This would handle both zero-inflated data and over dispersed data. However, for simplicity a negative binomial distribution was used instead.

3 GLM Regression

To perform negative binomial regression, Several steps where taken. Firstly, the data had to be processed so that regression could be properly done in python. The year variable contained values: Year $\in \{2022, 2023\}$. This was converted to be Year $\in \{1, 2\}$. This conversion also served the additional purpose of allowing Monte Carlo simulations for the year 2025 by incrementing the year value. Similarly, the month variable was converted to be Month $\in \{0, 1, 2, ..., 11\}$. Since there is no innate ordering of the provinces, the provinces were instead converted to dummy variables. These variables could take the value of $\{0, 1\}$ depending on whether or not the protest occurred in the related province.

A common problem that arises in Regression when using dummy variables is the dummy variable trap. The dummy variable trap occurs when a dummy variable is created for each category of a categorical variable. This can lead to multicollinearity. Since the model assumes the independent variables are indeed independent, this can lead to regression errors in the model. This trap was avoided by excluding one province from being turned into a dummy variable.

Negative binomial regression was performed using the statsmodel.api library from python. This library performs regression by using the iterative re weighted least squares method to maximize the likelihood of the data for some set of parameters. The logarithmic function is also used to link the mean: μ of the negative binomial distribution with the parameters of the GLM.

An issue with GLM regression with the negative binomial model in python using the statsmodel.api is that the library does not determine the over dispersion parameter for the negative binomial distribution when it performs regression. Therefore, the over dispersion parameter had to be obtained in with some other algorithm.

A simple line search algorithm for the over dispersion parameter α from 0.05 to 20 was performed. This was accomplished by performing regression on the data using the various α's. The α that obtained the greatest log-likelihood was then chosen as the over dispersion parameter.

4 Bootstrap Methods

After the parameters for the GLM with the optimal alpha was determined through regression, The next goal was to derive the necessary formulas to perform parametric bootstrapping. The negative binomial model is parameterized by :

$$Protests \sim \text{Negative Binomial}(r, p)$$

Firstly, since the log function was used as the link during regression, here is the relationship between the model parameters, E[Protests] and Var[Protests]:

$$E[Protests] = \mu = \exp\left(\beta_0 + \beta_1(\text{year}) + \beta_2(\text{month}) + \beta_3(\text{prov}) + \beta_4(\text{pop})\right)$$

$$Var[Protests] = \sigma^2 = \mu + \alpha\mu^2$$

The variance equation originates from the fact that the statsmodels.api library uses nb2 parametrization for the negative binomial regression model. Having obtained the mean and variance from regression, this values can be corresponded to the mean and variance of the negative binomial distribution:

$$\mu = \frac{r \cdot (1-p)}{p}$$

$$\sigma^2 = \frac{r \cdot (1-p)}{p^2}$$

solving for p and r gives:

$$p = \frac{\mu}{\mu + \alpha\mu^2}$$

$$r = \frac{p \cdot \mu}{1-p}$$

After p and r were obtained, 5000 parametric bootstrap samples were then created for all the combinations of the independent variables found in the original dataset. This was done through the numpy module in python which allows sampling from the negative binomial distribution.

After these bootstrap samples were obtained, GLM regression was then performed on each bootstrap sample to obtain the model parameters. This was repeated for each bootstrap sample to obtain 5000 bootstrapped model parameters. The top and bottom 2.5% of these bootstrapped model parameters were then cut off to produce a 95% bootstrapped confidence interval for the model parameters.

Significance of the model parameters was determined by observing whether 0 was contained in the 95% CI for each model parameter. If 0 was contained within the interval, this suggested that the independent variable linked to the specific model parameter may be insignificant in determining the E[protests] and thereby, in determining the number of protests.

Recall:

$$E[Protests] = \mu = \exp\left(\beta_0 + \beta_1(\text{year}) + \beta_2(\text{month}) + \beta_3(\text{prov}) + \beta_4(\text{pop})\right)$$

, therefore, a one unit increase in the i'th independent variable will increase the E[protests] by a factor of:

$$e^{\beta_i}$$

where β_i is the corresponding model parameter for the independent variable.

5 Monte Carlo Simulation

After the model parameters were obtained, Monte Carlo simulation for the year 2025 could be done.

Firstly, since the population for the year 2025 is unknown, the population for the year 2022 was used for each combination of month and province. Then, since the year independent variable was modeled as an integer to perform regression, the year variable was incremented from 1 to 4 to represent a 3 year jump from 2022 to 2025.

After the proper combinations of month,province and population were calculated, these independent variable combinations were used to then calculate the p and r parameters for the negative binomial distribution corresponding to that particular combination, analogous to the process in bootstrap methods but using the model parameters obtained during GLM regression on the original dataset. Since:

$$Protests \sim \text{Negative Binomial}(r, p)$$

, these p and r parameters were used to create 5000 Monte Carlo samples for the combinations of the independent variables. Once these samples were created, the top and bottom 2.5% of the the samples were removed to create a 95 % CI for the year of 2025.

6 Results

6.1 Model Results

The results of the negative binomial regression model are summarized in figure 1.

	coef	std err	z	P>|z|	[0.025	0.975]
year	-0.0119	0.080	-0.148	0.882	-0.170	0.146
month	-0.0207	0.012	-1.755	0.079	-0.044	0.002
pop	6.132e-07	4.24e-08	14.470	0.000	5.3e-07	6.96e-07
prov_British Columbia	0.0861	0.184	0.467	0.640	-0.275	0.447
prov_Manitoba	1.5626	0.163	9.584	0.000	1.243	1.882
prov_New Brunswick	1.4662	0.181	8.112	0.000	1.112	1.820
prov_Newfoundland and Labrador	1.3060	0.194	6.746	0.000	0.927	1.685
prov_Northwest Territories	-0.4568	0.330	-1.384	0.166	-1.104	0.190
prov_Nova Scotia	1.4607	0.174	8.372	0.000	1.119	1.803
prov_Nunavut	-0.0731	0.291	-0.251	0.801	-0.643	0.497
prov_Ontario	-5.2924	0.552	-9.581	0.000	-6.375	-4.210
prov_Prince Edward Island	0.6670	0.232	2.872	0.004	0.212	1.122
prov_Quebec	-1.9844	0.294	-6.758	0.000	-2.560	-1.409
prov_Saskatchewan	1.1026	0.177	6.222	0.000	0.755	1.450
prov_Yukon	0.7636	0.234	3.268	0.001	0.306	1.222
prov_Alberta	-0.2551	0.609	-0.419	0.675	-1.448	0.938

Figure 1: Model Results

Figure 1 illustrates the model parameters and some of their properties obtained by GLM regression in python

6.2 Bootstrapped CI

A 95% confidence interval for the model parameters is displayed in figure 2.

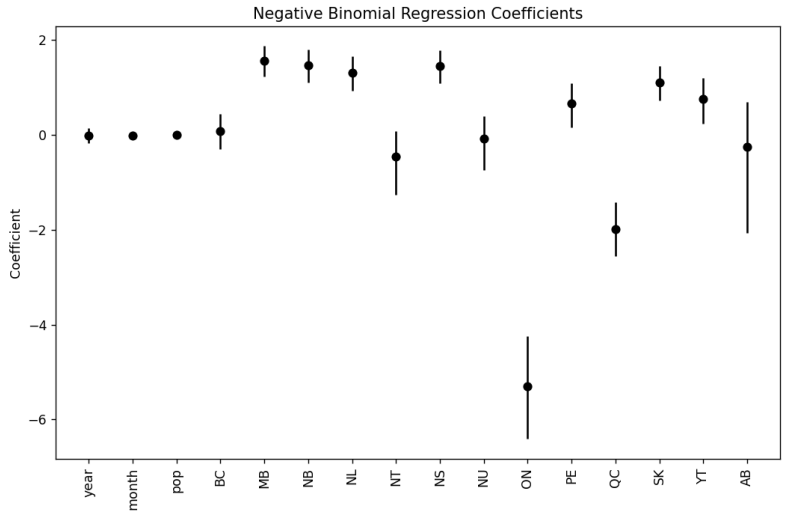

Figure 2: 95% CI for Model Parameters

The length of the vertical lines for each model parameter represents the width of the confidence interval. If the vertical lines crosses 0, this suggests 0 is contained in the CI for the model parameter and therefore, the model parameter is insignificant.

Here are independent variables corresponding to the model parameters that were found to be significant:

Population

Province $\in$ {MB, NB, NL, NS, ON, PE, QC, SK, YT}

6.3 Monte Carlo Simulation for 2025

Figure 3 shows the 95% confidence interval for the expected number of protests in 2025. The simulation took into account the population and other variables to predict potential protest numbers for the future.

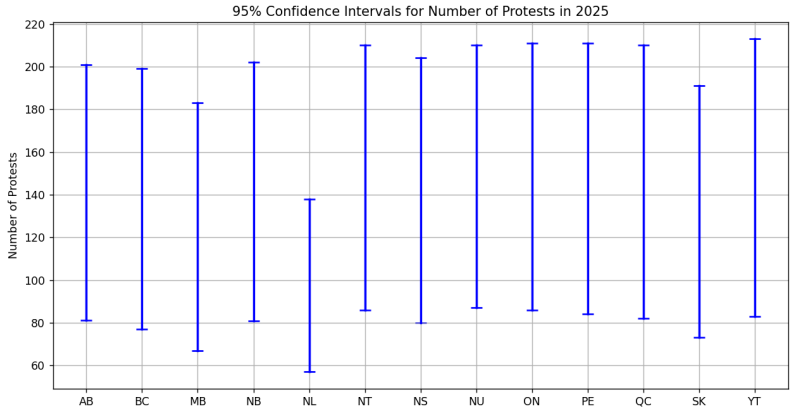

Figure 3: Monte Carlo Simulation for 2025

7 Conclusion

In conclusion, this study aimed to model social unrest in Canada using a negative binomial regression approach. Here are the main findings and their significance:

7.1 Model Selection

The negative binomial model was chosen due to its ability to handle count data with over dispersion, which was observed in the dataset. While a zero-inflated model could have been used for additional complexity, the negative binomial model provided a robust choice.

7.2 GLM Regression

The negative binomial regression was performed, with the model parameters obtained through iterative reweighed least squares. An additional step was taken to find the optimal over dispersion parameter, ensuring the model's accuracy.

7.3 Bootstrap Methods

Bootstrap methods were employed to determine the most significant model parameters and their impact on protests. By creating 5000 parametric bootstrap samples, confidence intervals for the model parameters were obtained. Significant variables included population and specific provinces.

7.4 Monte Carlo Simulation

The model parameters were then used to simulate Monte Carlo protests for the year 2025. By considering the population and other variables, a 95% confidence interval for the expected number of protests in 2025 was generated.

7.5 Results

The model results, as shown in Figure 1, provided insight into the relationships between independent variables and the expected number of protests. Significant variables such as population and specific provinces were identified.

The bootstrapped confidence intervals, displayed in Figure 2, highlighted the significant variables, aiding in understanding their impact on social unrest.

The Monte Carlo simulation for 2025, illustrated in Figure 3, gives a range of expected protests based on the model parameters and population considerations.

Overall, this study attempts to provide a framework for modeling and understanding social unrest, potentially offering insights into the number of protests of each province in Canada for the year 2025