Likelihood of Violence at Protests in North America

Kristine Villaluna

17/12/2021

Part I - Non-technical

Problem of Interest

A political demonstration is an action by a mass group or collection of groups of people in favor of a political or other cause. Political demonstrations can also take form of people protesting against a cause of concern. These demonstrations can be nonviolent or violent, or may start as nonviolent and then turn violent depending on the circumstances. Collecting more information about what citizens want when they demonstrate against governments, how they demonstrate, and how governments respond allows us to analyze the data in order to assess the political priorities of citizenries around the world.

In this study, we are interested in examining the factors that effect the likelihood of violence (both protester and government) at protests within North America.

Data Description

The Mass Mobilization Project has compiled a data set spanning the last three decades documenting protests from around the world. The goal of the project is to understand citizen movements against governments and how those governments respond. The Principle Investigators for the project are David H. Clark (Binghamton University) and Patrick M. Regan (University of Notre Dame). The Mass Mobilization project is sponsored by the Political Instability Task Force (PITF) which is funded by the Central Intelligence Agency.

The data are gathered by searching Lexis-Nexis for four key words: protest, demonstration, riot, or mass mobilization. The search is restricted to four primary newspaper sources: New York Times, Washington Post, Christian Science Monitor, and Times of London. The aim is to identify any protest event where the protest targets the government, and where it involves at least 50 people. After the search, if the sources return more than 100 articles, then the coders proceed in searching articles for evidence of codeable protest events. If these sources do not return at least 100 events, they expand the search to include regional and other sources.

For the purpose of this analysis, the data will be subset to protests in the North American region. The total data set includes 17,145 observations from 162 countries between 1990 and March 2020. With the subset data, we will be examining 527 observations from North America.

Plots and Tables

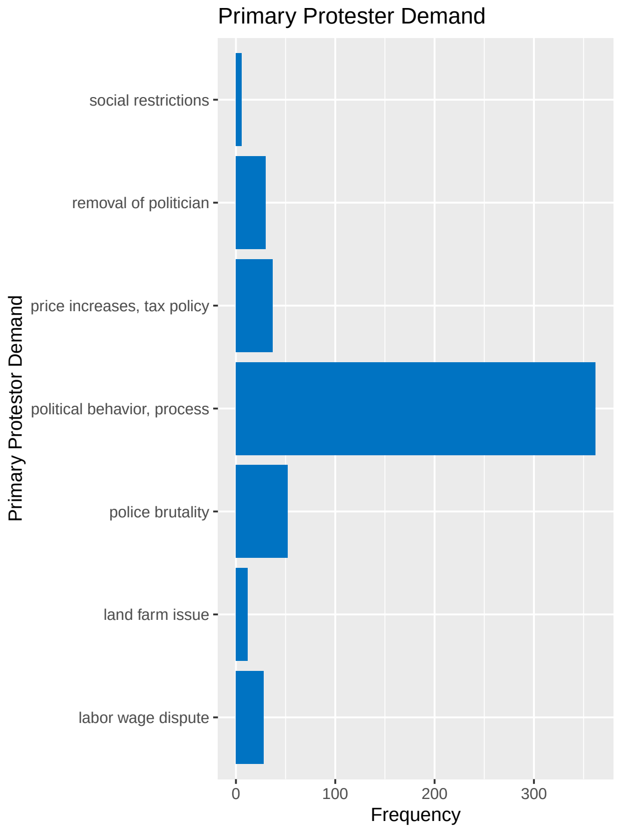
For each protest event, the project records protester demands, government responses, protest location, and protester identities. The following tables and plots are provided to get a sense of what these data look like.

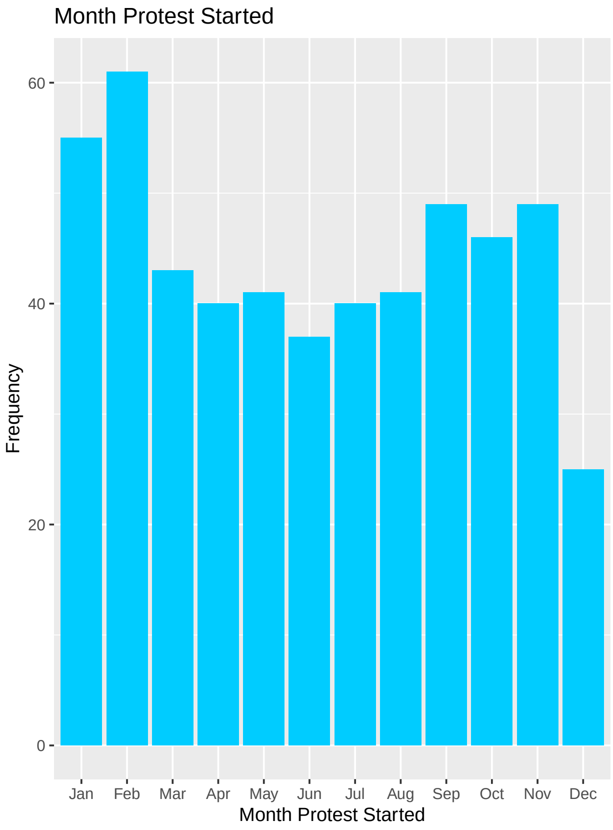
The variable that we were interested in modeling is `violence` at protests. This variable includes both protester violence, as well as state violence. Table 1 displays the occurrences of both types of violence. We can see that protester violence occurs more often. For this study, we will be looking at both protester and state violence, giving us 173 protests with violence.

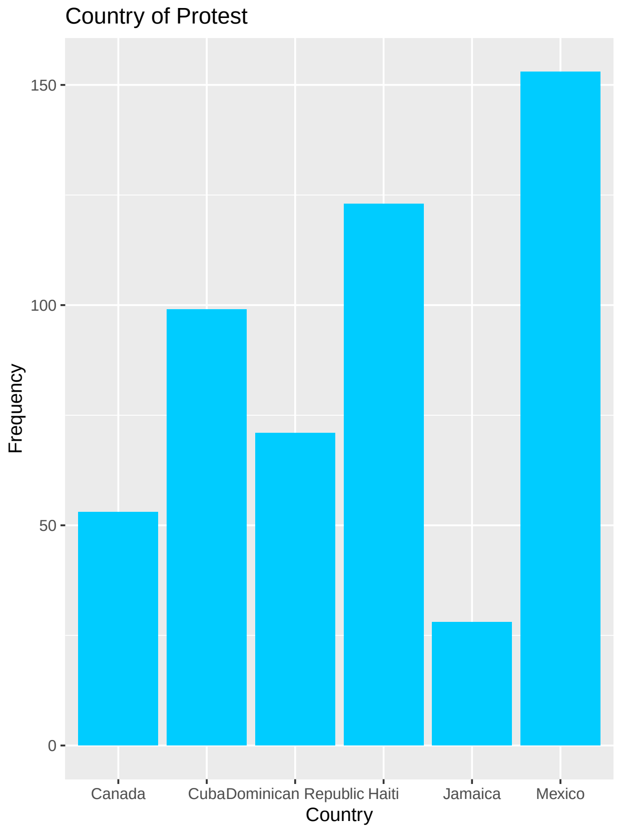
Table 1: State Violence Compared to Protester Violence

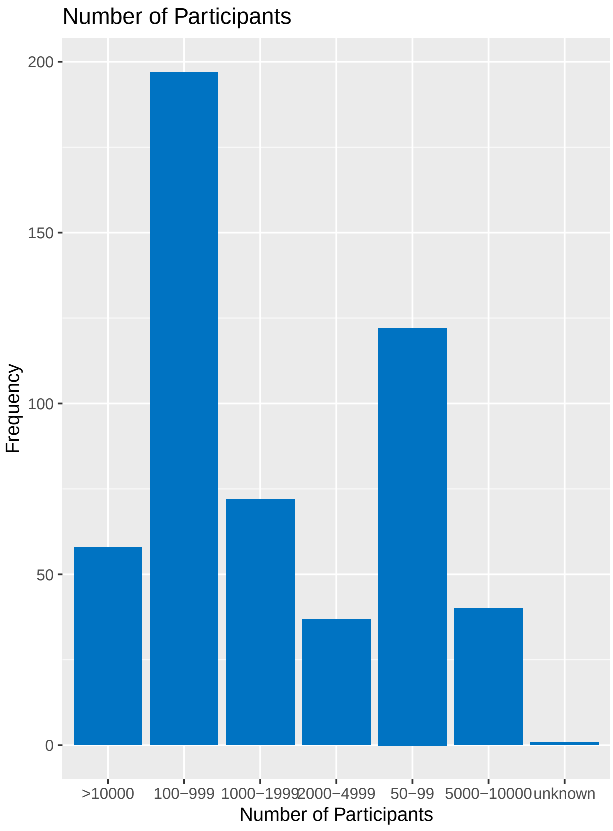
	Protesters Non-Violent	Protesters Violent
State Non-Violent	354	113
State Violent	18	42

The following plots display the primary protester demands on the left and the months that the protests started on the right. We see that the most frequent protester demand was related to political behavior or process. Each protest could have up to four demands coded in the data. A plot of the number of protest demands is available in Appendix, most protests only had one demand. Looking at the plot with the months, we can see that there seemed to be more protests at the beginning of the year (first two months), then dips in the summer time, and slightly increases again during the fall. Note that the month with the least amount of protests is December.


Primary Protester Demand


Month Protest Started


Country of Protest


Number of Participants

In the bottom two plots we see the countries of protest on the left and the number of participants on the right. We can see that the most frequent country of protest is in Mexico, followed by Haiti and then Cuba. Notably, the data set does not contain any information on protests that have occurred in the United States. Looking at the number of participants, we see that the most frequent category is 100-999 participants, followed by 50-90.

Additional plots from the exploratory data analysis is available in the Appendix. From these additional plots we can note that the most frequent years of protest were in the years 2014-2017, with the most frequent in 2016. We also see that the most frequent state response is to ignore the protest, followed by crowd dispersal and then arrests.

Summary (Non-Statistician)

In this analysis, the likelihood of violence at protests in North America was examined using a variety of different factors such as the length of the protest, number of participants, types of protester demands, and year. It was found that the most common type of primary protester demand was related to political behaviors or processes and the most frequent state response was to ignore the protest. Out of 527 protests, 173 had violence, from the either protesters and/or the government.

Appropriate statistical models were fit to the data and it was found that the most important factors that affected the likelihood of violence at a protest were country, year, number of participants, primary protester demand, length of protest (in days), number of demands, and an interaction term between country and year. To put it simply, an interaction effect between country and year exists when the effect of country on violence at a protest changes, depending on the year (and vice-versa).

Interpreting the type of model that was fit in this analysis can be done in simple terms of odds, which provide a measure of the likelihood of a particular outcome. They are typically calculated as the ratio of the number of events that produce that outcome to the number that do not. An odds ratio then, is a statistic that quantifies the strength of the association between two events.

It was found that the odds of a protest being violent increases 2.76% for a one-day increase in the length of the protest, holding all else constant. For the number of demands, there is a 101% increase in the odds of the protest being violent for a one-unit increase in the number of demands, holding all else constant. It appeared as though generally, the odds of protest violence decreased when the year increased, holding all else constant. A large limitation to this analysis is that the data set does not include any observations from the United States of America.

Part II - Technical

Models and Analysis

For this analysis, a logistic regression is used as the response variable is a binary variable, coded as 1 if there is protester and/or state violence at a protest, and 0 if there is no violence. There were a total of 527 observations from North America, with each observation representing a unit of analysis, namely, a protest.

There were a total of 8 potential covariates that were considered for the model. The variables that were considered were country, year, the month that the protest started, protest number, number of participants (grouped), the primary demand of the protesters, the length of the protest in days, and number of demands. A quick description of each variable is available below:

Year: The year the protest happened, between 1990-2014.

Month: What month the protest started.

Protest Number in Year: What number protest the protest is of that year. Ex: 3 if third protest in the year 1990.

Number of Participants: Number of participants categorized into the following groups: 50-99, 100-999, 1000-1999, 2000-4999, 5000-10000, >10000.

Primary Protester Demand: Seven categories of protester demands that describe the types of issues that motivate protest behavior.

The seven categories are: labor or wage dispute, land tenure or farm issues, police brutality or arbitrary actions, political behavior/processes, price increases or tax policy, removal of corrupt or reviled political person, or social restrictions.

Length of Protest: Length of protest in days.

Number of Demands: Derived variable based on protester demands variables (maximum 4).

Data Cleaning

After loading in the data and looking at the summary, it was found that there were no missing values for any of the numeric variables. It was found that there were missing values for the variable, `participants_category`. This was dealt with by imputing the category based on the `participants` variable. This was particularly tricky as there were some values that had to be manually reviewed, such as "hundreds", "100s", "a dozen", etc. The rest of the missing values were in the additional protester demand and state response variables (ex, `protestdemand2`, `protestdemand3`, etc.). The primary ones, `protesterdemand1` and `stateresponse1`, had no missing values.

Next, additional variables were derived to be used in the analysis, namely, length of protest, state violence, overall violence (protester and state), and number of demands. Once these variables were derived and checked for any errors, we were ready to begin examining the data.

EDA

Please see the appendix for additional plots created during the exploratory data analysis. Each potential covariate, as well as the response variable, were plotted to view the distributions and frequencies of the variables and look for any anomalies in the data. After these checks, we are ready to start modeling.

Model Selection

As previously mentioned, logistic regression was the choice of model for these data, given the binary nature of the response variable. The data was first filtered to only units that had a known participant category, and then the proper variables were coded as factors. To begin, a logistic regression model was fit with all of the potential covariates. Looking at the model fitted with all of the values, we saw that some variables were not significant at the 0.05 level, namely, `protestnumber` and `startmonth`. Making note of these variables as ones we could potentially remove, we proceeded to check this finding by using the drop1 and ANOVA functions with the likelihood ratio test (LRT). Looking at the output of both the drop1 and ANOVA functions, we found that we could remove the variable `protestnumber`.

Note that the residual deviance of this first model was 463.19 on 504 degrees of freedom. The null hypothesis of the residual goodness of fit test is that our model is correctly specified (i.e, is adequate for the data), and we do not have evidence to reject that hypothesis (p-value = 0.9). So, using the residual deviance goodness of fit test, we have strong evidence that our model fits adequately. Note that adding variables to the model will always increase our deviance, so we continued to do more model selection to see if we can remove some variables or add some interactions.

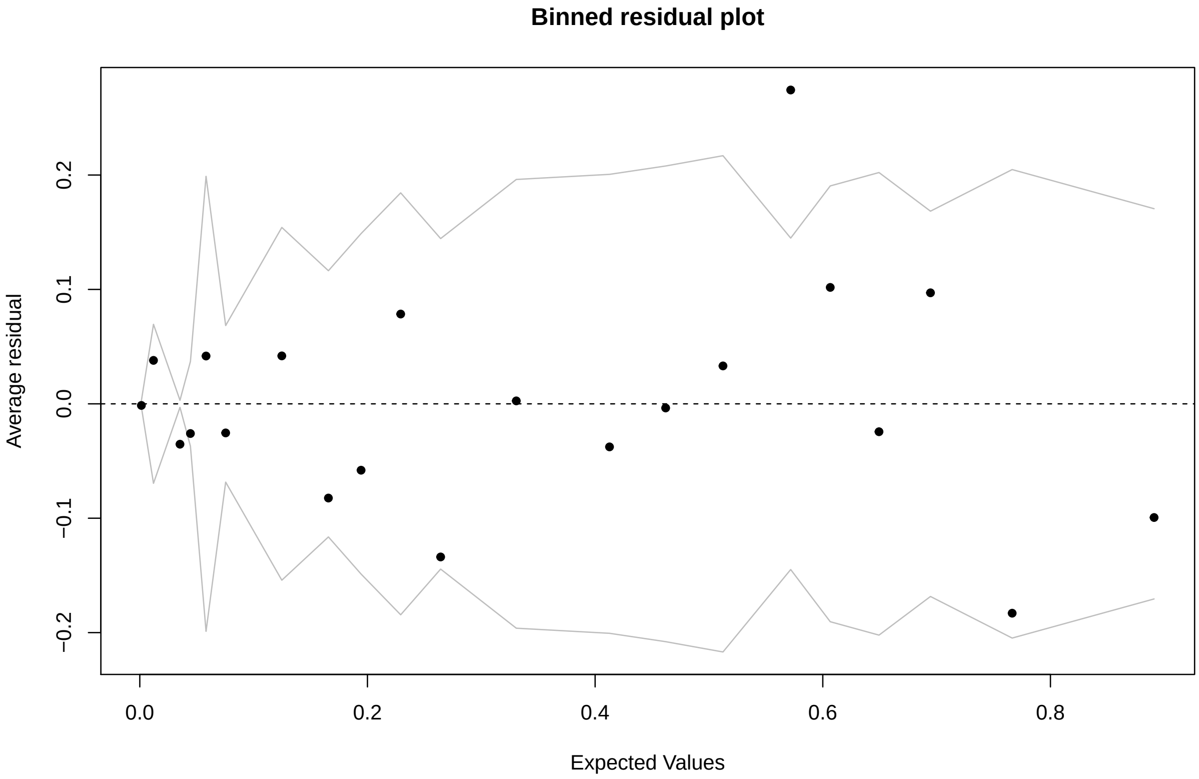
A second model was then fit without the protest number variable, and again, the drop1 and ANOVA functions were used to see if there were any variables that could be dropped form the model. This process continued until there were no more variables that were not statistically significant. After examining the second model, it was found that the variable `startmonth` could be dropped, which was expected. The third model was then fit and it was found that all the variables left in the model were significant using the LRT tests.

To check if this was a sufficient stopping point, we considered taking out one variable to see how it affected the model. After taking out `num_demands`, an ANOVA test of the two models using the LRT test was done and it was found the model with `num_demands` was significant, so we proceeded to keep it in the model. To further check that we have sufficiently selected the main effects model, we used automatic selection procedure, namely, backward elimination, to see if the variables we have chosen make sense. Backwards selection can be tricky when dealing with categorical covariates with multiple levels. In this case, we must make sure to consider the variable as a whole, and should not remove one level but not the others. In the end, the results were the same, and the final main effects model included country, year, number of participants, primary protester demand, length of protest in days, and number of demands.

After we obtained the main effects, we considered adding interaction terms into the model. To add all second-order interaction terms into the model would cause an issue due to not having enough degrees of freedom. After checking different combinations of interactions one-by-one, none appeared to be statistically significant, therefore, the main effects model will be used.

Model Checking

We then checked our final model using the residual deviance goodness of fit test. Again, the null hypothesis of the residual goodness of fit test is that our model is correctly specified (i.e, is adequate for the data), and we do not have evidence to reject that hypothesis (p-value = 0.88). So, using the residual deviance goodness of fit test, we have strong evidence that our model fits adequately.



Next, we wanted to look at a residual plot using the final model. Traditional residual plots are not helpful with logistic regression, and we can instead look at the binned residual plot. The gray lines represent +-2 SE bands, which we would expect to contain about 95% of the observations. We say that a model is reasonable if the majority of the fitted values seem to fall within the SE bands.

Based on the plot above, the model looks reasonable, but there are more outliers among the residuals than we would expect from chance alone (alpha = .05): 21 binned residuals but 2 outliers = 0.09. It looks like the model does not do well when the fitted value is around 0.57.

Summary (Statistician)

In this analysis, we set out to examine the various factors that effect the likelihood of violence at protests around North America. First a logistic regression model was fit using all of the potential covariates. Iteratively using both the `anova()` and `drop1` functions with the likelihood ratio test, we determined that the variables primary protest demand and start month could be eliminated from the model. This was confirmed using an automatic selection procedure, namely, backward stepwise selection.

Interaction terms were then added into the model one-by-one to check significance. Adding all second-order interaction terms into the model at once caused an issue due to not having enough degrees of freedom. It was found that no interaction terms were statistically significant. Finally, the residual deviance goodness of fit test was done to check the fit of the selected model which contained country, year, number of participants, primary protester demand, length of protest in days, and number of demands. Moreover, model diagnostics were checked using the binned residual plot.

The final model output can be seen in the table below. Coefficients of logistic regression may be interpreted as log-odds ratios. By exponentiating the coefficients, we can interpret the odd ratio as the odds multiplying by e^{β} for every 1-unit increase in x holding all else constant. For variables with multiple levels, the levels are coded as dummy variables, with one level as a reference category. Then, the odds ratios for a dummy variable is the factor of the odds that Y=1 within that category of X, compared to the odds that Y=1 within the reference category.

For example, the odds of violence at a protest when there are 100-999 participants are $e^{-0.727} = 0.48$ times that of the odds of violence at protests with 50-99 participants (reference group). We also found that the odds of a protest being violent increases 2.76% for a one-day increase in the length of the protest, holding all else constant. For the number of demands, there is a 101% increase in the odds of the protest being violent for a one-unit increase in the number of demands, holding all else constant.

We note that the standard error for protester demand - social restrictions is very large, 551.408 and could be indicative of underlying issues with the model such as multicollineraity or overfitting. The following section describes limitations to this project and potential next steps.

term	estimate	std.error	statistic	p.value
(Intercept)	79.487	27.100	2.933	0.003
countryCuba	-2.297	0.804	-2.857	0.004
countryDominican Republic	1.336	0.522	2.559	0.011
countryHaiti	1.914	0.465	4.119	0.000
countryJamaica	-0.094	0.597	-0.157	0.875
countryMexico	0.233	0.439	0.532	0.595
year	-0.040	0.014	-2.986	0.003
participants_category100-999	-0.727	0.351	-2.068	0.039
participants_category1000-1999	-1.129	0.429	-2.633	0.008
participants_category2000-4999	-0.575	0.489	-1.175	0.240
participants_category5000-10000	-0.772	0.485	-1.591	0.112
participants_category>10000	-2.509	0.614	-4.084	0.000
protesterdemand1land farm issue	0.390	0.837	0.466	0.641
protesterdemand1police brutality	1.351	0.589	2.293	0.022
protesterdemand1political behavior, process	-0.196	0.515	-0.381	0.703
protesterdemand1price increases, tax policy	0.763	0.629	1.213	0.225
protesterdemand1removal of politician	0.649	0.677	0.959	0.338
protesterdemand1social restrictions	-14.127	551.408	-0.026	0.980
length_days	0.027	0.010	2.652	0.008
num_demands	0.700	0.341	2.053	0.040

Limitations

As with any investigative analysis, it is important to highlight the limitations of the project, as well as potential avenues for future work of this type. The data set was limited in that there were no observations for the USA. Having this type of information would help paint a fuller picture of what protests are like in North America. Furthermore, the small amount of observations in this data set (527 observations) made this model susceptible to overfitting. This can be seen in the goodness-of-fit statistics that may seem too good to be true. It is possible that most occurrences of violence happened disproportionately more often in certain countries or years. One way to remedy this would be to use the full data set and examine different countries around the world, and consider different statistical techniques such as controlling for random and fixed effects.

Part III - Appendix

Please see submitted .Rmd and .html for additional plots and analysis.

References

Clark, David; Regan, Patrick, 2016, “Mass Mobilization Protest Data”, https://doi.org/10.7910/DVN/HTTWYL, Harvard Dataverse, V5, UNF:6:F/k8KUqKpCa5UssBbL/gzg== [fileUNF]