Summary of Newspaper Closure Dataset

Simon J. KIss

19/05/2021

R Markdown

This is how the data-set is structured right now.

status	province	municipality	CSD code	year	turnout	number of candidates	margin_percent	na	n	acclaimed	treatment
0	Alberta	Calgary	4806016	2004	19.80	7	0.74	0	NA	NA	0
0	Alberta	Calgary	4806016	2007	32.90	9	0.44	0	NA	NA	0
0	Alberta	Calgary	4806016	2010	53.39	15	0.08	0	NA	NA	0
0	Alberta	Calgary	4806016	2013	39.43	9	0.52	0	NA	NA	0
-1	Alberta	Calgary	4806016	2017	58.10	10	0.08	0	NA	NA	1
0	Alberta	Canmore	4815023	2007	24.56	1	NA	1	1	1	0
0	Alberta	Canmore	4815023	2010	42.03	2	0.06	0	NA	NA	0
-1	Alberta	Canmore	4815023	2013	37.64	3	0.57	0	NA	NA	1
-1	Alberta	Canmore	4815023	2017	40.90	2	0.29	0	NA	NA	1
0	Alberta	Edmonton	4811061	2004	41.80	8	0.08	0	NA	NA	0
0	Alberta	Edmonton	4811061	2007	27.20	9	0.40	0	NA	NA	0
0	Alberta	Edmonton	4811061	2010	33.40	7	0.26	0	NA	NA	0
0	Alberta	Edmonton	4811061	2013	34.50	6	0.43	0	NA	NA	0
0	Alberta	Edmonton	4811061	2017	31.50	13	0.66	0	NA	NA	0
0	Alberta	Grande Prairie	4819012	2007	29.20	2	2.42	0	NA	NA	0
0	Alberta	Grande Prairie	4819012	2010	24.00	5	0.15	0	NA	NA	0
0	Alberta	Grande Prairie	4819012	2013	20.95	2	1.17	0	NA	NA	0
0	Alberta	Grande Prairie	4819012	2017	24.00	4	0.41	0	NA	NA	0
0	Alberta	Jasper	4815033	2007	37.22	2	0.27	0	NA	NA	0
-1	Alberta	Jasper	4815033	2010	39.94	2	0.77	0	NA	NA	1

Right now we have data on 117 municipalities. They come from a basically a combined public data-set that has tracked newspaper closures and a data-set of historic election results. There's a big range of municipalities in here, from large metropolises to rural townships, so I'm going to guess we'll want to include some kind of control for population size.

WE have three dependent variables:

1. Margin of Victory (Percent)
2. Turnout (percent)
3. Number of mayoral candidates.

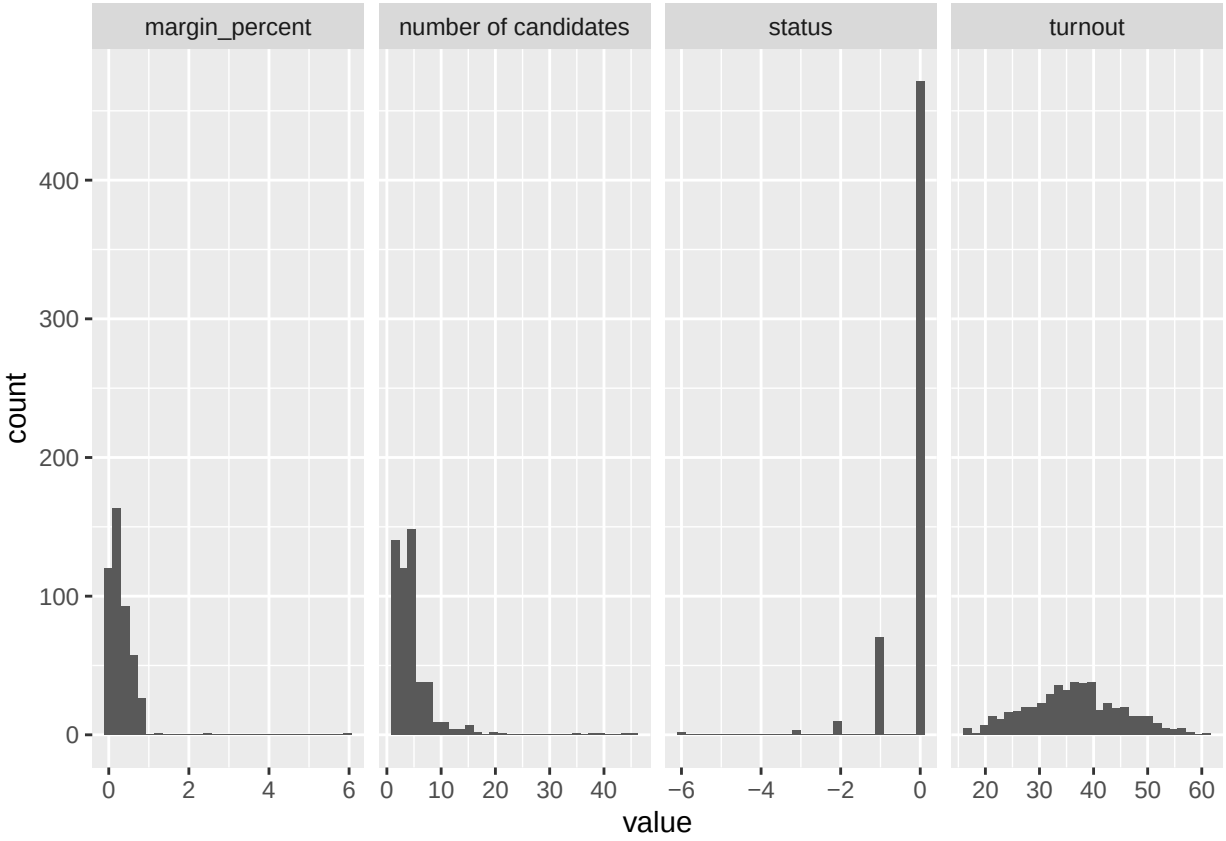
We have mixed data on each. For example, we have 461 election-year combinations where we have all three

dependent variables, 481 where we have two dependent variables and 534 where we have one dependent variables.

In terms of newspaper closures, we have set up the data where each municipality enters the data-set with a value of 0, and then receives a -1 each time we have a record of a newspaper closure before that election-year. Once it has suffered a newspaper closure, that score stays with it through the data-set, so the treatment is I guess hypothesized to have lasting effects. If another newspaper is closed, then the score goes down by one again. So with this, we can quickly define a dichotomous variable 0 or 1 whether a municipality in a year has had a newspaper closure, or we can try to model the effects linearly.

The treatment is distributed as follows:

The treatment and dependent variables are distributed as follows. So there are some weird outliers here. For example, one municipality has 6 newspaper closures recorded; one municipality had like 40 candidates for mayor But those can be dealt with, although that will then reduce sample size a bit.



So, I am wondering three things now. First of all, is our sample size roughly big enough to pick up an effect? Second, what do I need to do next to structure the data-set? Do I need to calculate the change in turnout, nuumber of candidates and margin percent for each election year i.e. there will be missing values for the first row for each municipality? Third, how do actually regress this in R?

Lastly, we intend to go further and get some demographic data to add to this data-set. Specifically we are thinking getting change in population size for each census sub-division, change in share of degree holders, maybe change in unemployment rate and maybe chane in income. Does all that sound good?