Panel Data Exercise

This assignment consists of two exercises. First, you will work with the same Twitter takedown data from last week and investigate Russian trolls' tweeting patterns in response to different political events. Second, you will replicate the findings in Berman, Shapiro & Felter (2011) about the "Hearts and Minds" effect in Iraq. Please be sure to set your working directory to your folder on Dropbox.

Data

For Exercise 1, you will use the same dataset on Russian troll tweets from last week's exercise.

For Exercise 2, we are providing you ESOC's Iraq Civil War Dataset v2 that is used in Berman, Shapiro & Felter (2011), which you can also download from here. The codebook for all the datasets is `"030810_HAM_codebook.doc"`. You should consult this to identify the correct variables for the second exercise.

Guidelines

1 Exercise #1: Russian Troll Tweets

Since 2016, many journalistic accounts have reported that Russia aimed to stoke polarization on issues related to police violence and race relations. For this exercise, we will test these claims quantitatively and see whether IRA tweeting patterns really are affected by BLM-related events in the US. Specifically, we'll look at the following events: the death of Freddie Grey on August 19, 2015, Sandra Bland on July 13, 2015, and the Alon Sterling shooting on July 5, 2016.

(a) Create a panel of the following variables at the day level (English language tweets only):

- Number of tweets
- Average engagement metrics (replies, likes, quotes, and retweets)
- Number of tweets mentioning either "BLM" or "Black Lives Matter"

(b) How many observations do you have?

(c) Run the following event study:

$$Y_t = B_0 + \beta_1 T + \beta_2 X_t + \beta_3 X_t T_t + \epsilon \tag{1}$$

where Y_t is the outcome variable at time t, X_t is an indicator that equals 0 during the pre-treatment period and 1 post-treatment, and $X_t T_t$ is an interaction term. Set the pre/post window as 30 days. Your code should loop over each of the 3 events as well as each of the 6 variables in your tweets dataset (tweet count, four engagement metrics, tweets about BLM). Try not to use a pre-made package to run this model, but rather set up each variable yourself.

(d) What is the interpretation of each coefficient?

(e) Report your results. What do they indicate about IRA trolls' tweeting patterns in response to racial violence events in the US?

2 Exercise #2: Can Hearts and Minds Be Bought?

In Berman, Shapiro and Felter (2011), the authors develop an information-based model of insurgency where violence is modeled as a context between violent rebels, a government, and civilians whose main role is deciding whether or not to share information about insurgents with government forces. Governments seek to minimize violence and they do so using some combination of counter-insurgency efforts and service provision. The authors argue that service provision and economic aid has a violence-reducing effect because it increases incentives for civilians to share information about insurgents, which in turn determines the effectiveness of counterinsurgent actions. They test their theory using data on attacks against Coalition and Iraqi forces from SIGACT reports and reconstruction spending directed toward local public goods provision under three programs (which they refer to as CERP spending). For this exercise, you will replicate the main findings from their paper which uses the following estimating equation:

$$v_{it} = \alpha_i + \beta g_{it} + \gamma' z_{it} + \epsilon_{it} \tag{2}$$

where i indexes the district and t index the half-year. v is the number attacks, g is CERP spending, and z is a vector of control variables.

(a) Create a merged dataset of attacks against Coalition and Iraqi forces (per 1,000) and reconstruction spending per capita at the district - half year level from January 2004 through December 2008. Make sure to read the codebook to identify the correct variables and don't forget to check for things like missing data, duplicates, and the like when cleaning/merging data. How many districts and half-years are there, and how many total observations?

(b) Run a simple regression of incidents on CERP spending, weighted by population, with standard errors clustered by district. Report the regression output. What do the results suggest? *Note: if using R, see here for clustered standard errors and replicating STATA SE's.*

(c) Run the same regression but control for the following things:

- Economic grievances among civilians, measured as the mean change in household income between 2002 and 3004.
- Opportunity cost of rebellion using (1) unemployment rate and (2) the proportion of the district's population in the bottom two income quintiles.
- The ethnosectarian distribution of each governorate, noisily captured by the Sunni and Shia vote share in the December 2005 election.

Report the results. How did they change compared to the results from (a)? Why do you think controlling for these variables had this effect on the results?

(d) Include additional time controls in your model along with these basic controls, specifically year indicators and their interactions with Sunni vote share. Report and explain your results.

(e) Estimate a first-differenced version of Equation (2) with time controls, population weights, and standard errors clustered at the district level. Run a second version of the first differences model that includes additional controls for pre-existing trends (in the previous half-year):

$$\Delta v_{it} = \alpha_i + \beta \Delta g_{it} + \gamma' \Delta z_{it} + \phi \Delta v_{it-1} + \Delta \epsilon_{it} \tag{3}$$

What happens to your results when you move from levels to first differences? Why do you think that happens?

(f) Using the lessons learned from this exercise, explain when it is appropriate to use differences instead of levels, and what kinds of problems fixed effects can or cannot account for.