DS421 Stat 261 Project: Regression Model

Group Members: Dan Blaustein-Rejito (GSPP), Ian Bolliger (ERG), Hal Gordon (ARE), Andy Hultgren (ARE), Yang Ju (Landscape Arch. and Env. Planning), Kate Pennington (ARE), Sara Stoudt (Statistics)

4/4/16

---

1 Motivation

Our group has been interested in examining the socioeconomic effects in the U.S. of drought generally, with different group members interested in different particular outcomes. The goal of this document is to describe a regression framework for characterizing county-level resiliency to drought, along dimensions such as per capita mortality, per capita income, and average water expenditures, and unemployment. So, a "resilient" county would, for example, have a small change in mortality associated with a given increase in drought exposure.

Then, in a second stage we would like to explore the (non-causally identified) associations between county characteristics and their sensitivities. County-level characteristics could include mean income, gender, race, etc.

2 Data and Regression Models

Drought data have been pulled from the U.S. Drought Monitor, which categorizes each area of the U.S. as being in a condition of Severe, High, Moderate, Low, or Zero drought for each week from 2000 - 2015. We have defined our Drought variable to be 1 if the drought level is Moderate or higher, and 0 otherwise. From this dataset, we have calculated the number of days each county experienced moderate or higher drought, on an annual basis.

Our ideal (i.e. data-permitting) first stage empirical model is presented below. If lack of variation in the data does not permit the level of fixed effects and time trends described, we may need to relax the specification.

First Stage:

$$y_{i,t} = \beta_i D_{i,t} + \alpha_i + \tau_i t + \gamma_{s,t} + \epsilon_{i,t} \tag{1}$$

Where $D_{i,t}$ refers to the number of days in U.S. Drought Survey bins 2-4 in county i and year t, α_i are county fixed effects controlling for time-invariant differences between counties, τ_i is the coefficient on a county level linear time trend, and $\gamma_{s,t}$ are state-by-year fixed effects controlling for state level time trends common across all counties $i \in s$. Note that the state-by-year fixed effects will non-parametrically account for national trends in the outcome of interest as well as state-level trends. The identifying variation in this model is within-county, annual deviations from the county time trend and from statewide annual average drought levels. Standard errors will need to be corrected for serial correlation over space and time, due to the spatial and temporal nature of droughts (neighboring observations in time and space are not independent draws).

From a causal perspective, this specification would be vulnerable to bias in the coefficient estimates if county-level time trends differed systematically from both the county linear time trend as well as state-level time trends. The following example illustrates one such case. Take annual mean income as an outcome of interest. Say a county happens to have a non-linear decline in income over the sample period (perhaps due to a large employer failing slowly at first and dramatically toward the end of the sample) and county-level drought happens to exhibit a similar pattern. Further, for simplicity imagine the county is small and state-level trends are flat. In this case, residual variation in drought levels and income would both be positive early in the sample period, and would both be negative late in the sample period. This would create a downward bias in the drought coefficient β_i. That is, the trend of declining county income (in fact due to the declining large employer) would be improperly attributed to the trend in county drought exposure. The key point here is that drought and income data that are not well captured by a linear time trend, nor by the state-year fixed effects, can introduce bias.

Second Stage:

$$\beta_i = \rho_0 + \boldsymbol{\delta}\mathbf{X}_i + \nu_i \tag{2}$$

Where the β_i come from Eq.(1) for a given outcome; $\mathbf{X}_i$ represents a vector of county characteristics such as urban/rural, mean age, and home ownership; and $\boldsymbol{\delta}$ is a vector of the associated coefficients.

This regression is cross-sectional and therefore not well identified from a causal perspective. Any omitted variable that happens to covary with both the levels of β_i and the variables $\mathbf{X}_i$ will bias the coefficients $\boldsymbol{\delta}$. However, this model will illustrate how "drought resilience" (a low value of β_i) covaries with a set of common county socioeconomic characteristics. The goal with this stage is to gain some insight regarding what characteristics are commonly associated with drought resilience, or the lack thereof. Because county level characteristics likely are correlated over space, we anticipate correcting our OLS standard errors by clustering over space.