Problem Set 6

Research Design for Causal Inference
Due: June 2, 2015

Part I: Concepts behind RDD and IVE

Question 1: In the context of regression discontinuity designs, what is a forcing variable and how can you use one to identify a causal effect? Provide an example.

Question 2: Still thinking about RDDs, explain the concept of "bandwidth." How does using a larger bandwidth impact your estimate of treatment effects?

Question 3: Describe at least two strategies for checking the validity and robustness of RD estimates of causal effects. For each strategy, be sure to explain the potential threat that it addresses as well as the basic mechanics of how to perform the check.

Question 4: What are the properties of a good instrument (for performing valid instrumental variables estimation)? Feel free to use words, pictures, equations, diagrams, etc. in your response.

Question 5: Explain the "no third path" restriction. What is it? Why does it matter? How do you test for a third path? What is an example of a study we read about or discussed in which this exclusion rule might have been violated?

Question 6: Why is an IVE always LATE? Explain what this means and draw on at least one example to illustrate your explanation.

Part II: RDD Analysis

For this part of the Problem Set, you'll analyze a subset of the data from an unpublished manuscript that I am preparing together with Benjamin Mako Hill.[1]

In the study, we analyze the effect of requiring contributors to online communities to adopt "cheap pseudonyms" (disposable usernames you might create on a website). We are interested in understanding the impact of requiring cheap pseudonyms on the quality of contributions coming into the wiki.

We take advantage a series of quasi-experiments that affected 137 wikis hosted by Wikia. All of these wikis underwent a policy change: from one day to the next, anyone who wanted to edit these wikis was suddenly required to login with a username. This shift was abrupt and (for the vast majority of those involved) both unannounced and unanticipated, making it the sort exogenous shock well-suited to quasi-experimental inference. The fact that the change was implemented in software means that end-users and potential participants in the wikis had no way of avoiding exposure to the "treatment" (in this case, the requirement of using a cheap pseudonym). We estimate the effect of this change within and across all of these wikis by comparing edits before and after the shock.

The overarching goals here are for you to reproduce a slightly-simplified part of our analysis by modeling the effect of the change on two measures of quality using an RD analysis. I also ask you to perform a pseudo-cutoff test to assess the robustness of the findings.

The dataset is available as an RData file at: `http://aaronshaw/teaching/2015/causal/data/ps6.RData`. Each row of the dataset corresponds to a "wiki-week" of observations, that is one week for one wiki. This is the unit of analysis. For almost all of the wikis, there are 24 weeks of data. The variables included in the dataset are described in Table . The DVs are `edits.reverted` (a count of bad edits) and `edits.non.reverted` (a count of good edits). Note that when you load the dataset you will also import a function called `reduced.summary()`. This will come in handy later.

Question 1: Summarize (provide descriptive statistics about) all of the substantively meaningful wiki-level independent variables at the week immediately before the software change (week -1) and the week immediately after (week `0`).

Question 2: Summarize both of the dependent variables across the entire dataset for all wikis. Summarize these same variables at the week immediately before the software change (week -1) and the week immediately after (week `0`). Provide a visual comparison of these pre-/post- weeks.

Question 3: For both dependent variables, calculate a naive estimate of the ATE across all wikis by comparing the average in the week immediately before the block (week -1)

[1] `http://mako.cc`

Variable name	Definition
`wiki`	The name/url of the wiki.
`window.week`	Week relative to the cutoff.
`blocked`	Indicator for whether unregistered contributions were blocked.
`age.weeks`	The age (in weeks) of the wiki-week.
`total.pages`	The total (cumulative) # of pages in the wiki-week.
`total.editors`	The total (cumulative) # of editors in the wiki-week.
`edits.anon`	The # of edits from unregistered contributors in the wiki-week.
`edits.reverted`	The # of edits from that wiki-week which were deleted
`edits.non.reverted`	The # of edits from that wiki-week which were not deleted.

Table 1: Variables in Shaw & Hill's "cheap pseudonyms" dataset

and the week immediately after (week `0`). Perform a statistical test (of your choosing) to determine whether this difference seems different than what you might expect by chance.

Question 4: Summarize your findings from Questions 1-3. Based on what you know about the study, are your naive estimates (from Question 3) biased? Why might you find them credible or not?

Question 5: Explain in words how you might model the effect of the software change using a regression discontinuity approach.

Question 6: Use an RD approach to estimate the Average Treatment Effect for both of the dependent variables (use any kind of regression model you like, but if you're not sure what to go with, I'd recommend good old OLS). In addition to whatever variables you need to model the effect of the block, I recommend you include only the covariates `window.week` and `wiki` as well as their interaction term, `window.week*wiki`. Let's call these "fixed effects" — they control for any variations related to week-specific and wiki-specific factors respectively (I'll explain more in class next week). Note that when you include fixed effects you don't really need to interpret or even report any of the resulting coefficients from your model. The `reduced.summary()` function that Mako and I wrote helps you do this by only displaying the coefficients for the variables you care about in this dataset.

Question 7: Interpret your findings from Question 6. What is the effect of requiring cheap pseudonyms of the number of reverted (bad) edits coming into these wikis? What is the effect on the number of non-reverted (good) edits? What sorts of conclusions should Mako and I draw? What limitations do you see based on this analysis?

Question 8: Test whether the finding is valid using a “pseudo-cutoff.” Our options are somewhat constrained by the amount of data I’ve given you here, but you can create a `fake.blocked` variable and move the block out to a different week than the actual one in which it occurred. Then run the same analysis you did in Question 6 substituting `fake.blocked` for `blocked` in the model. Report and interpret the results of this validity check in light of your findings from Question 6.