Problem Set 5

Research Design for Causal Inference
Due: May 12, 2015

For this Problem Set, you'll re-analyze a subset of the data from:

> Blattman, Christopher, and J. Annan. 2010. "The Consequences of Child Soldiering," *Review of Economics & Statistics*, 92 (4): 882-898. [Available via MIT Press]

The study consists of a panel survey of male youth in several war-afflicted areas of Uganda. The authors attempt to estimate the impact of forced military service – in this case, abduction into the Lord's Resistance Army – on young men's educational and labor market outcomes. Blattman and Annan describe the abductions (2010: 883):

> Abduction was large-scale and seemingly indiscriminate; 60,000 to 80,000 youth are estimated to have been abducted and more than a quarter of males currently aged 14 to 30 in our study region were abducted for at least two weeks. Most were abducted after 1996 and from one of the Acholi districts of Gulu, Kitgum, and Pader.
>
> Youth were typically taken by roving groups of 10 to 20 rebels during night raids on rural homes. Adolescent males appear to have been the most pliable, reliable and effective forced recruits, and so were disproportionately targeted by the LRA. Youth under age 11 and over 24 tended to be avoided and had a high probability of immediate release. Lengths of abduction ranged from a day to ten years, averaging 8.9 months in our sample. Youth who failed to escape were trained as fighters and, after a few months, received a gun. Two thirds of abductees were forced to perpetrate a crime or violence. A third eventually became fighters, and a fifth were forced to murder soldiers, civilians, or even family members in order to bind them to the group, to reduce their fear of killing, and to discourage disobedience.

Assignment to "treatment" (a.k.a. abduction) was not *truly* randomized, but (according to Blattman & Annan) was "as-if" randomized, supposedly resulting in conditional inde-

pendence of treatment assignment on any of the observed or unobserved covariates. The overarching goals here are to (1) assess the as-if random assumption; and (2) estimate treatment effects using multiple techniques in order to address any potential breakdowns in the as-if random assumption.

The dataset is available as a comma-separated value (.csv) file at: `http://aaronshaw.org/teaching/2015/causal/data/blattman.csv`. The variables included in the dataset are described in Table 1. Note that `educ`, `distress`, and `log.wage` are all post-treatment outcomes.

Table 1: Variables in Blattman & Annan (2010) dataset

Variable name	Definition
`abd`	1 if respondent was abducted by the LRA (treatment)
`c_ach -- c_pal`	Location indicators corresponding to a geographic sub-district
`age`	Age (years)
`fthr_ed`	Father's education (years)
`mthr_ed`	Mother's education (years)
`orphan96`	Indicator if parents died before 1997
`hh_fthr_frm`	Indicator if father is a farmer
`hh_size96`	Household size in 1996
`educ`	Respondent's education (years)
`distress`	Emotional distress (index 0–15)
`log.wage`	Log of average daily wage over previous 4 weeks

Question 1 – Naive estimates of the ATE

Part a Use OLS regression to calculate estimates of the Average Treatment Effect of abduction on education. Generate an estimate with and without covariate adjustment. Interpret the results in a few sentences.

Part b This study is not the product of a "pure" natural experiment (treatment was not randomly assigned). Given that fact (and what you know about observational studies in general), what are some of the potential threats to the validity of your estimates in Part a?

Part c Assess the covariate balance between the abducted and non-abducted youth respondents in the study for all pre-treatment covariates. Include any descriptive statistics you think are relevant and report the results of any statistical tests you use in your assessment.

Part d Visually assess balance on a few covariates (of your choice) by using a box-plot. The default graphics command for a box-plot in R is: `boxplot(variable ~ categories)`,

where `categories` is some categorical variable and `variable` is some continuous variable you are comparing across each category in `categories`.

Part d Discuss the implications of your findings in Parts c and d for your findings in Part a. How credible do you find Blattman and Annan's claims that treatment assignment was "as-if random"?

Question 2 – Propensity score weighting

For this question, you'll use an inverse-propensity score weighting technique very similar to that employed by Blattman and Annan to attempt to improve the precision and validity of the estimates you calculated in Question 1, Part a.

The intuition guiding this effort comes from the work of Don Rubin and Paul Rosenbaum, who in a 1983 article introduced the "propensity score" as a means of generating improved balance between treated and un-treated units conditional on some combination observed covariates. *The propensity score (p-score) is an estimate of the probability that any unit in an observational study was selected into either treatment or control, given some combination of observed covariates.* For details on propensity score adjustment and the properties of different propensity score adjusted estimators, see the Hirano, Imbens, and Ritter (2003) paper cited by Blattman and Annan.

To generate improved estimates from Question 1, I ask you to (1) calculate the propensity score for every unit using a logistic regression model; (2) "trim" the dataset to remove units with extreme p-score values; (3) assess covariate balance in the trimmed dataset; (4) estimate treatment effects using a regression model that incorporates the *inverse* propensity scores as weights.[1]

Part a Use a logistic regression model to estimate the probability that subjects in Blattman & Annan's study were treated (abducted). A generic R command for performing logistic regression is as follows:

```
model <- glm( y ~ x1 + x2 + ... + xN, data = DATA,
         family = binomial(link = logit))
```

where `y` is the dependent variable; each `x` is an independent variable and `DATA` is your data frame. Two comments: (1) in theory, the more covariates you use in calculating your p-score, the better; however, in practice you may discover that it makes sense to eliminate some covariates from your p-score model if they include many missing values or empty categories (not a problem here). (2) An intuitive p-score for each unit in your analysis is

[1] Why use the inverse p-scores? Think about this one for a minute and if it doesn't make sense to you let's be sure to discuss it in class.

its fitted value from the logistic regression model (you might also use the exponentiated fitted value, the inverse of the fitted value, or other quantities depending on the circumstances). Fitted values from the results of a regression model in R (e.g., using the output of the command above) can be accessed using `model$fitted.values`.[2]

Part b Assess balance on p-scores across treated and untreated units. Inlcude any descriptive statistics you think are relevant as well as the results of any statistical tests you use in this assessment.

Part c Visually assess balance on the propensity score with a box-plot or a density plot.[3]

Part d Trim the dataset so that only units with propensity score values between the 10^{th}—90^{th} percentiles of all propensity scores remain. You'll want to use the `quantile()` command to locate the values of the p-scores corresponding to these percentiles.

Part e Assess the covariate balance between treatment and control groups in the trimmed dataset (use the original covariates, not the p-scores). Be sure to conduct any statistical and visual comparisons you deem relevant. Compare your results here to the results of Question 1, Part c. Discuss what you find in this comparison. What sorts of units did the trimming procedure remove from the dataset?

Part f Estimate the average treatment effect (ATE) of abduction on educational attainment using a linear regression model with inverse propensity score weights applied to all of the units. Create the weights as a new variable equal to the inverse of the propensity score for all units in the treatment condition and the inverse of 1 minus the propensity score for all units in the control condition. When you run the regression, you can pass this weights variable as an additional argument to the `lm()` or `glm()` command (e.g., `weights = d$my.weights$`).

Part g Discuss the results of Question 2 Part e. Consider the differences between your estimates after propensity-score weighting with the results of "naive regression" in Question 1, Part a. Which of these estimates do you (not) find to be credible? Why? What are the limitations, if any, of your analysis given the particular characteristics of the research design and the balancing procedure you have pursued here?

---

[2] Why use fitted values? Have a think on this one too and be prepared to discuss it in class.

[3] Before you run this in R, step back for a moment: what do you expect the graph(s) to look like? Why?

Key Concepts for Next Class

- Regression Discontinuity Design (RDD).
- Forcing variable (again).
- Bandwidth.
- Threats to the validity of RDD studies.
- Pseudo-outcomes.
- Pseudo-cutoffs (placebo tests).