Introduction to Causal Inference

Problem Set 4

Professor: Teppei Yamamoto

Due Tuesday, July 19 (at beginning of class)

Only the required problems are due on the above date. The optional problems will not directly count toward your grade, though you are encouraged to complete them as your time permits. If you choose not to work on the optional problems, use them for your self-study during the summer vacation.

Required Problems

Problem 1

In this problem we will assess one of the most famous social science articles using instrumental variables, Acemoglu, Johnson, & Robinson's 2001 paper "The Colonial Origins of Comparative Development: An Empirical Investigation" (henceforth AJR).[1]

First, we will begin with a stylized characterization of the study. Assume that AJR use the following variables for any country i that was previously colonized:

> Instrument $Z_i \in \{0, 1\}$: Mortality in the 17th, 18th, and early 19th centuries, 0 if low mortality, 1 if high.
>
> Treatment $D_i \in \{0, 1\}$: Modern property rights institutions, 0 if weak, 1 if strong.
>
> Outcome Y_i: Modern log GDP per capita.

In our stylized characterization, assume that AJR use instrumental variables to estimate the effect of D_i on Y_i by instrumenting for D_i with Z_i. They find that having strong modern property rights institutions causes higher GDP per capita. (**Note:** As we will see in a minute, AJR include various specifications in which they also control for some pre-treatment covariates, but for now we will focus on the "simplest" empirical strategy.)

(a) Assuming that their empirical strategy is valid, draw a simple DAG to represent the instrumental variables approach used by AJR. Include a hypothetical unobserved confounder (U_i) that creates a back-door path between treatment and outcome. Why is it important to include this hypothetical unobserved confounder?

[1] This paper has over 8,000 citations, and is heavily debated in a range of disciplines including economics, political science, and history. This problem set question is highly stylized, and if you are really interested in the substantive and methodological details of the paper we encourage you to read the paper and surrounding debates carefully. Also remember, it is always easier to criticise something than to build it yourself!

(b) Name the five assumptions underpinning instrumental variables as a strategy for identifying the effect of an endogenous treatment on the outcome for compliers. Write out each assumption formally in terms of Z_i, D_i, and Y_i. In your own words, interpret each assumption with regard to the specific setup of AJR's study. Finally, discuss the plausibility of each assumption. (**Hint:** It may be useful to refer to your DAG from (a) in interpreting and assessing some assumptions.)

(c) We will replicate the main specifications from AJR using their publicly available replication data. Download the data `ajr_data.dta` from the course website and read it into `R`. The data has the following variables:

`shortnam`: Three letter country code for each unit.
`logpgp95`: Log purchasing power parity GDP per capita, 1995.
`avexpr`: Average protection against expropriation risk.
`logem4`: Log settler mortality.
`lat_abst`: Absolute value of latitude of capital city.
`africa`: Dummy=1 if African.
`asia`: Dummy=1 if Asian.
`other`: Dummy=1 if Other.
`america`: Dummy=1 if American.

Estimate the effect of `avexpr` on `loggp95` in two ways using naïve regression approaches. First, run a linear regression of `loggp95` on `avexpr` as the lone regressor (do not include any other covariates). Second, do the same but include, linearly and additively, `lat_abst`, `africa`, `asia`, and `other`. Present the results in a table, including HC2 robust standard errors. Interpret the direction and statistical significance of the estimates. Why should we be concerned about whether these are good estimates of the causal quantity of interest?

(d) Now, use the same two approaches as in part (c) and estimate the effect of `logem4` on `loggp95`. Interpret the direction and statistical significance of the estimate of the causal effect. What does this "reduced form" estimator purport to estimate? Under what conditions can we interpret this result as causal?

(e) Finally, use the `ivreg` function in the `AER` package to estimate the complier local average treatment effect (LATE) of `avexpr` on `loggp95`, with `logem4` used as the instrument for `avexpr`. As in (c) and (d), first include no covariates, and second include linearly and additively `lat_abst`, `africa`, `asia`, and `other`. In your result, be sure to report and interpret the F-statistic from a test for weak instrumentation, which you can retrieve from an output of `ivreg`. What do you find?

Problem 2

De Kadt and Rosenzweig, graduate students at MIT, investigate partisan ties and their effects on national resource allocation in Ghana. The specific research question of interest is whether

constituencies in Ghana that elect MPs who are from the same party as the President (the "ruling party") receive more electrification over the next four years. Using electoral data from the 1996 parliamentary elections and nightlights data the authors use a sharp regression discontintuity (RD) design to investigate the effect of the treatment (an MP from the ruling party winning the 1996 election) on the outcome (change in nightlights over the next four years). The forcing variable the authors use is voteshare of the ruling party MP candidate. The unit of analysis is the constituency. In this problem you will similarly conduct a sharp RD analysis using the dataset `Ghana_RD.csv`.

The dataset contains 152 observations each corresponding to a constituency in Ghana, with the following variables:

- `constit`: Name of consistuency
- `voteshare`: Voteshare (% votes obtained) for the ruling party MP candidate
- `treatment`: The treatment status indicator (1 if the ruling party MP won the 1996 election in that constituency, 0 if the ruling party candidate lost)
- `changeNL_1996_2000`: Change in nightlights between 1996 and 2000
- `mean_1995`: Mean level of nightlights in 1995

(a) First, let's look at the data to see if sharp RD makes sense for our dataset. Plot treament (y-axis) as a function of the forcing variable (x-axis), where the forcing variable is the margin of victory/loss for the ruling party MP candidate. In other words, adjust the variable `voteshare` such that the cutpoint lies at 50%. Does it seem appropriate to use a sharp regression discontinuity design in this case?

(b) Estimate the local average treatment effect (LATE) at the threshold using a linear model with common slopes for treated and control units. What are the additional assumptions required for this estimation strategy? Provide a plot of the change in nightlights from 1996 to 2000 (y-axis) and forcing variable (x-axis for the range of -30 to 50) in which you show the fitted curves and the underlying scatterplot of the data. Interpret your resulting estimate.

(c) Conduct the same analysis as in part (b) except that you should now use a linear model with *different* slopes for treated and control units.

(d) Conduct the same analysis as in part (b) except that you should now use a quadratic model with different model coefficients for the treated and control groups.

(e) Use the `rdd` package in `R` to estimate the LATE at the threshold using a local linear regression with a triangular kernel. Note that the function `RDestimate` automatically uses the Imbens-Kalyanamaran optimal bandwidth calculation. Report your estimate for the LATE and an estimate of uncertainty.

(f) How do the estimates of the LATE at the threshold differ based on your results from parts (b) to (e)? In other words, how robust are the results to different specifications of the regression? What other types of robustness checks might be appropriate?

(g) Conduct one such test by examining the density of the forcing variable around the cutoff. First, create a histogram plot of the forcing variable (binned by 1 percentage points). Second, conduct a formal test of the difference in density of the forcing variable around the cutoff by using McCrary's test (`DCdensity` function in the `rdd` package will be helpful) and report the p-value from the test. Why is this analysis a good diagnosis of the identification assumption for the sharp RD design? What can you say about the plausibility of no manipulation in voteshare?

(h) Finally conduct a placebo test using nightlights measured in 1995 as the outcome in a sharp RD analysis for the 1996 election. Use local linear regression as you did in part (e). Also graph the points for the forcing variable (x-axis) and 1995 nightlights (y-axis). What does this placebo test say about the relationship between the 1996 election of ruling party MPs and nightlights measured in 1995?

Optional Problems

Problem A

(a) Suppose that we are interested in the effect of a potentially endogenous causal variable X_i on an outcome variable of interest Y_i. Assume that we have another variable Z_i, which is binary and is an instrumental variable for X_i. Show that the IV estimator for the effect of X_i on Y_i

$$\hat{\beta}_{IV} = \frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, X_i)}$$

Can be written as

$$\frac{(\bar{Y}_1 - \bar{Y}_0)}{(\bar{X}_1 - \bar{X}_0)},$$

where $\text{cov}(\cdot)$ is the sample covariance; $\bar{Y}_0$ and $\bar{X}_0$ are the sample averages of Y_i and X_i over the part of the sample with $Z_i = 0$; and $\bar{Y}_1$ and $\bar{X}_1$ are the sample averages of Y_i and X_i over the part of the sample with $Z_i = 1$.

(b) Let $\mathbf{X} = [1, X_1, X_2, \cdots, X_k, D]$ and $\mathbf{Z} = [1, X_1, X_2, \cdots, X_k, Z]$. The matrix $\mathbf{X}$ contains the covariates (including a vector of 1s) and your treatment vector D, and $\mathbf{Z}$ is a matrix of the same covariates and the instrument for the treatment variable in place of the actual treatment. Y is a vector of observed outcomes. We can construct the following system of linear equations, with error terms u_2 and u_1 respectively:

$$\begin{aligned} Y &= \mathbf{X}\beta + u_2 \\ D &= \mathbf{Z}\pi + u_1 \end{aligned}$$

with coefficient vectors $\beta = [\beta_0, \beta_1, \cdots, \beta_k, \beta_D]$ and $\pi = [\pi_0, \pi_1, \cdots, \pi_k, \pi_Z]$.

The IV estimator can be obtained by:

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'Y$$

i) What are the conditions under which the treatment effect estimate $\hat{\beta}_{D,IV}$ is consistent?

ii) You can obtain the Two-stage Least Squares estimator by the following steps.

(a) Run the first stage regression: $D = \mathbf{Z}\pi + u_1 \Rightarrow \hat{\pi} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'D$

(b) Get fitted values: $\hat{D} = \mathbf{Z}\hat{\pi}$

(c) Regress Y on $\hat{\mathbf{X}} = [1, X_1, X_2, \cdots, X_k, \hat{D}]$: $Y = \hat{\mathbf{X}}\beta_{2SLS} + u_3$

Show formally that $\hat{\beta}_{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'Y = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'Y = \hat{\beta}_{IV}$. Comment on the steps along the way to reach your conclusion. If you need any additional assumptions, please state them.

Problem B

An alternative way of thinking about regression discontinuity designs is to think of them as "locally randomized" experiments. Consider the following setup:

$$\begin{aligned} Y &= \tau D + \delta_1 W + U \\ D &= 1[X \geq c] \\ X &= \delta_2 W + V \end{aligned}$$

where Y is the dependent variable, $D \in \{0, 1\}$ is the treatment indicator, X is the forcing variable, and c is the cutpoint. V and U are unobserved variables, and W is a set of predetermined and observable characteristics. Making no further assumptions about the correlations between V, U, and W, the identification assumption for the RDD under this approach can be stated as:

$$\Pr(W = w, U = u | X = x) \text{ is continuous in } x.$$

(**Hint:** Refer to sections 2 and 3 of Lee & Lemieux (2010) for more detail about this setup and the identification assumption.)

(a) Interpret the identification assumption given above, in light of the setup. Why does it represent "local randomization"? Intuitively, why does this assumption allow us to identify the causal estimand, the LATE at the discontinuity threshold?

(b) Prove that the LATE at the threshold can be identified under the above assumption as:

$$\tau_{RDD} = \lim_{\eta \downarrow 0} \mathbb{E}[Y \mid X = c + \eta] - \lim_{\eta \uparrow 0} \mathbb{E}[Y \mid X = c + \eta]$$

where η is a real number in the range of X.