Problem Set 7

Professor: Christopher Lucas

Due Monday, March 16

You are required to use either `knitr` or `Markdown`, these are fully compatible with `R` and LaTeX.

Please (1) bring a printed version of your answers to class and (2) upload an electronic version on Canvas. Hard copies of your solutions should be handed in at the start of class and electronic copies are due before class begins at 4:00 pm. Late submissions will not be accepted. Please, comment our code and show your work in questions that require calculations.

Please set your seed using set.seed(02139) when the question requires simulations.

1 The IV estimator

(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 $\hat{\beta}_{IV}$ is consistent?

ii) Now let's obtain the Two-stage Least Squares estimator. We can do that following these next 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.

2 Coding the IV estimator in R

Write your own function for instrumental variables regression in R. The setup is as follows: you have one outcome variable, one endogenous treatment variable with one instrument, and you may have additional regressors (covariates). You will have $\mathbf{X} = [1, X_1, X_2, \cdots, X_k, D]$ and $\mathbf{Z} = [1, X_1, X_2, \cdots, X_k, Z]$ where D is a column vector indicating treatment status, Z is a column vector for the instrument for the treatment, and the other X's are column vectors of covariates. You will also have Y, a column vector for the outcome.

Your function should:

- Take in $\mathbf{X}$, $\mathbf{Z}$, and Y (as matrices or dataframes)
- Compute the IV estimator directly using one step (not by doing the two stages regression)
- Output the vector of the IV estimates of regression coefficients, $\hat{\beta}_{IV}$
- Output the standard errors for those coefficients

3 The Colonial Origins of Comparative Development

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:

[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!

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?

(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 now replicate the main specifications from AJR using their publicly available replication data. Download the data `ajr_data.dta` from Canvas 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.

Using OLS, estimate the effect of `avexpr` on `loggp95` in two ways, **without** using instrumental variables regression. First, estimate a linear regression with `loggp95` as the dependent variable, and `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? Broadly, are these concerns issues of "estimation" or "identification"?

(d) Now, again using OLS, estimate the effect of `logem4` on `loggp95`. First, estimate a linear regression with `loggp95` as the dependent variable, and `logem4` 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 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) Using the function you coded in Problem 2, use instrumental variables regression to estimate the (Conditional) Local Average Treatment Effect (LATE) of `avexpr` on `loggp95`, using `logem4` 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`. (**Hint**: You are replicating columns 1 and 8 from Panel A of Table 4 in the American Economic Review publication of AJR. Please do not be surprised if the results are very slightly different from the published version.)

(f) Finally, verify your function using `ivreg` in the package `AER` by replicating your results from (e). In your table, be sure to report and interpret the F-statistic from a test for weak instrumentation (you can retrieve this quantity from the objects generated by `ivreg`). What do you find?