Problem Set 2
Due Tuesday, November 10 at 11:59 AM (noon) via Canvas
ECON 31720, University of Chicago, Fall 2020

Assignments must be typeset in nicely formatted LaTeX. Programming assignments must use low-level commands, be commented clearly (not excessively), formatted nicely in 80 characters per column, etc. **You will be graded on the exposition of your written answers, the clarity of your code, and the interpretability and beauty of your tables and graphics.** The problem sets are individual assignments, but you may discuss them with your classmates. Submit the problem sets through Canvas in a single zip/tar/rar file. **Late problem sets will not be accepted under any circumstances.**

1. Let $\mathcal{H}$ denote the set of all proper distribution functions for a scalar random variable. Suppose that we observe a non-negative random variable Y with distribution $G \in \mathcal{G}$, where $\mathcal{G}$ is the subset of $\mathcal{H}$ such that $\mathbb{P}_G[Y \geq 0] = 1$. Assume that Y is determined by

$$Y = \max\{U, 0\}, \tag{1}$$

where U is an unobserved random variable with distribution function F. Assume that $F \in \mathcal{F}$, where $\mathcal{F}$ is some subset of $\mathcal{H}$ specified below. Notice that any $F \in \mathcal{F}$ implies a distribution for Y called G_F, defined as

$$G_F(y) \equiv \mathbb{P}_F[\max\{U, 0\} \leq y].$$

Let $\mathcal{F}^\star(G)$ denote the sharp identified set for F, that is

$$\mathcal{F}^\star(G) = \{F \in \mathcal{F} : G_F(y) = G(y) \quad \text{for all } y \in \mathbb{R}\}.$$

Consider the target parameter $\pi : \mathcal{F} \to \mathbb{R}$ defined as $\pi(F) = \mathbb{E}_F[U]$. Let $\Pi^\star(G)$ denote the sharp identified set for π.

   (a) Suppose that $\mathcal{F} = \mathcal{H}$. Determine $\Pi^\star(G)$ for any $G \in \mathcal{G}$. Is the model falsifiable?

   (b) Suppose that $\mathcal{F} = \{F \in \mathcal{H} : F(-1) = 0 \text{ and } F(2) = 1\}$. Determine $\Pi^\star(G)$ for any $G \in \mathcal{G}$. Is the model falsifiable?

   (c) Suppose that $\mathcal{F} = \{F \in \mathcal{H} : F(0) = 1/2\}$. Determine $\Pi^\star(G)$ for any $G \in \mathcal{G}$. Is the model falsifiable?

   (d) Suppose that $\mathcal{F} = \{F \in \mathcal{H} : \mathbb{E}_F[U] = 0\}$. Determine $\Pi^\star(G)$ for any $G \in \mathcal{G}$. Is the model falsifiable?

   (e) Now change the target parameter to $\pi(F) \equiv \text{med}_F(U) \equiv \inf\{u : F(u) \geq \frac{1}{2}\}$ (the median of U when it's distributed like F). Suppose that $\mathcal{F} = \mathcal{H}$. Determine $\Pi^\star(G)$ for any $G \in \mathcal{G}$. Is the model falsifiable?

2. Consider the simple instrumental variables model

$$Y = \alpha X + U,$$

where X is scalar. Suppose that the instrument, Z is also scalar, and that there is homoskedasticity with respect to Z in both the reduced form and first stages. (This is the case used to discuss weak instruments in the supplemental notes.) Let $\hat{F}$ denote the sample first stage F–statistic. Determine the asymptotic bias of $\hat{F}$ as an estimator of the concentration parameter ("μ^2") under weak instrument asymptotics. Construct an alternative estimator that is asymptotically unbiased.
Hint: No need to reinvent what I did in the supplemental notes. You may use any of the derivations there. Constructing the alternative estimator should be easy.

3. Consider the binary treatment potential outcomes model $Y = DY(1)+(1-D)Y(0)$. Suppose that we have a binary treatment Z that is independent of $(Y(0), Y(1), D(0), D(1))$, where $D = ZD(1)+(1-Z)D(0)$. Maintain the Imbens and Angrist (1994) monotonicity condition that $D(1) \geq D(0)$, and assume that $\mathbb{P}[D = 1|Z = 1] - \mathbb{P}[D = 1|Z = 0] > 0$. Let $G \in \{c, a, n\}$ denote the types complier, always-taker and never-taker, i.e. $G = c$. Let $F_d(y|g)$ denote the distribution function for $Y(d)$ conditional on $G = g$.

   (a) Show that $F_1(y|a)$ is point identified for any y.

   (b) Show that $F_0(y|n)$ is point identified for any y.

   (c) Show that both $F_1(y|c)$ and $F_0(y|c)$ are point identified for any y.

   (d) Suppose that both $Y(0)$ and $Y(1)$ are continuously distributed scalar random variables with support over the entirety of the real line. Let F_1 and F_0 denote the unconditional distribution functions of $Y(1)$ and $Y(0)$. Define the random variables $U(0) \equiv F_0(Y(0))$ and $U(1) \equiv F_1(Y(1))$. Assume that $U(0) = U(1) = U$ with probability 1.

   Show that $F_0(y)$ and $F_1(y)$ are point identified for any y.

   (e) Maintain the assumptions of the previous part. Let G denote the unconditional distribution function of $Y(1) - Y(0)$.

   Show that $G(y)$ is point identified for any y.

4. This problem is about the paper "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size" by Angrist and Evans (1998, *The American Economic Review*). The authors define their endogenous variable as a binary indicator for whether a family has exactly two or more than two children. They use two instruments: the same-sex instrument and the twin birth instrument, both separately and together.

   (a) Discuss the interpretation and credibility of the monotonicity condition when using the same-sex instrument separately.

   (b) Discuss the interpretation and credibility of the monotonicity condition when using the twin birth instrument separately.

   (c) Discuss the interpretation and credibility of the monotonicity condition when using the same-sex and twin birth instruments together.

5. Consider an instrumental variables model with a binary treatment $D \in \{0,1\}$ and binary instrument $Z \in \{0,1\}$. The usual potential outcomes are $Y(0)$ and $Y(1)$ and the usual potential treatment choices are $D(0)$ and $D(1)$. Let $X \in \{1,\ldots,K\}$ be a vector of covariates that has a discrete distribution with K points of support. Let β_{tsls} denote the population coefficient on D in the following two stage least squares specification:

$$\begin{aligned} \text{reduced form: } & Y \text{ on } Z \text{ and } \{\mathbb{1}[X = k]\}_{k=1}^{K} \\ \text{first stage: } & D \text{ on } Z \text{ and } \{\mathbb{1}[X = k]\}_{k=1}^{K}, \end{aligned}$$

and assume that β_{tsls} exists. Note that, relative to the case discussed in the slides and notes, the first stage in this specification does not contain interactions between Z and X. Maintain the usual exogeneity condition that Z is independent of $(Y(0), Y(1), D(0), D(1))$ conditional on X. Assume that $\mathbb{P}[D(1) \geq D(0) | X = k] = 1$ for all $k = 1, \ldots, K$.

(a) Explain how the monotonicity condition given here differs from the one discussed in the slides.

(b) Show that

$$\beta_{\text{tsls}} = \mathbb{E}\left[\frac{\text{Cov}(D, Z|X)}{\mathbb{E}[\text{Cov}(D, Z|X)]}\, \mathbb{E}[Y(1) - Y(0) | D(1) = 1, D(0) = 0, X]\right]$$

where $p(z, x) \equiv \mathbb{P}[D = 1 | Z = z, X = x]$ is the propensity score.

(c) Explain what this result shows and provide some intuition.

6. Consider the following data generating process:

$$\begin{aligned} Y &= X + U \\ X &= .3Z_1 + V, \end{aligned}$$

where (U, V) are jointly normal, mean zero, with variance matrix given by

$$\begin{bmatrix} .25 & .20 \\ .20 & .25 \end{bmatrix}.$$

Let $Z_1, Z_2, \ldots, Z_{20}$ be independent standard normals. Run a Monte Carlo experiment that compares the performance of the following estimators:

- The OLS estimator of Y on X and a constant.
- The TSLS estimator with Z_1 as an instrument for X and has only a constant as a control (included instrument).
- The TSLS estimator with $Z_1, \ldots, Z_{20}$ as instruments for X, and only a constant as a control variable.
- The jackknife IV estimator with Z_1 as an instrument for X, and only a constant as a control variable.

- The jackknife IV estimator with $Z_1, \ldots, Z_{20}$ as instruments for X, and only a constant as a control variable.

For each estimator, report the bias, median, standard deviation across simulations. Also, report the coverage rate of a 95% confidence interval. Consider sample sizes, $N = 100, 200, 400, 800$. Design an easy-to-read table to report your results. Explain your results in the context of instrumental variable regressions with many instruments.

7. Read "Semiparametric instrumental variable estimation of treatment response models" by Alberto Abadie (2003, *Journal of Econometrics*). The paper is on Canvas. The data used in Section 6 is also on Canvas.

   1. Reproduce Table 2.
      Note: As usual, getting standard errors exactly right can be difficult, in part because the precise meaning of "robust standard errors" is a bit vague. In particular, several small-sample modifications to the Eicker-Huber-White standard errors have been proposed in the literature, and people are usually not clear on which one they are using. These different refinements have names like HC0, HC1, HC2, etc. The best guess on what someone used is typically the Stata default, which at least today is HC1. This is also typically the form you will see in textbooks, with the degrees of freedom correction for the number of regressors. However, for older papers this is not always so clear due to changes in software over time. If you have trouble getting the standard errors, then just bootstrap!
   2. Compute a 95% Anderson–Rubin confidence interval for the coefficient on the endogenous variable (participation in a 401(k) plan). Compare this confidence interval to the one you would obtain using standard asymptotic approximation for TSLS. Are they similar or different? Explain.
   3. Compute a jackknife IV estimator with bootstrapped standard error. Compare this to the TSLS results reported in Abadie's column (3). Are they similar or different? Explain.