Business 41903

Instructor: Christian Hansen

Problem Set 5

1. Verify that $\widehat{\tau} = \widehat{\text{LATE}}$ where $\widehat{\tau}$ is the IV estimator of τ from the model

$$Y = \delta + \tau D + \varepsilon$$

using $Z \in \{0,1\}$ as instrument for $D \in \{0,1\}$ and $\widehat{\text{LATE}} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0}$ with $\bar{Y}_1$ the sample average of Y in the subsample with $Z = 1$, $\bar{Y}_0$ the sample average of Y in the subsample with $Z = 0$, and $\bar{D}_1$ and $\bar{D}_0$ defined analogously.

2. The average treatment effect on the treated is defined as $\text{ATT} = \text{E}[Y_1 - Y_0 | D = 1]$ where Y_1 and Y_0 are potential outcomes in the treatment and control states and D is a treatment indicator. Assume that (i) $\text{E}[Y_0|D, X] = \text{E}[Y_0|X]$ and that (ii) for all $x \in \mathcal{X}$, $\Pr(D = 1|X = x) \leq 1 - \eta$ for some $\eta > 0$.

a. Show that ATT is identified under assumptions (i) and (ii).

b. A regression adjustment estimator for the ATT is $\widehat{\text{ATT}}_{reg} = \frac{1}{\sum_{i=1}^n D_i} \sum_{i=1}^n D_i[\widehat{g}_1(X_i) - \widehat{g}_0(X_i)]$ where $g_1(X) = \text{E}[Y_1|D = 1, X]$ and $g_0(X) = \text{E}[Y_0|X]$. Suppose we are willing to model $\text{E}[Y_1|D = 1, X] = g_1(X) = X'\beta_1$ and $\text{E}[Y_0|X] = g_0(X) = X'\beta_0$. Derive the asymptotic properties of the estimator - i.e. show it is consistent for ATT, asymptotically normal, and provide a consistent estimator of the asymptotic variance.

c. A propensity score weighted estimator for the ATT is $\widehat{\text{ATT}}_{prop} = \frac{1}{n} \sum_{i=1}^n \frac{(D_i - \widehat{p}(X_i))Y_i}{(n_1/n)(1-\widehat{p}(X_i))}$ where n_1 is the number of treated observations and $p(X) = \text{E}[D|X]$ is the propensity score. Suppose we are willing to model $\text{E}[D|X] = p(X) = \Lambda(X'\gamma)$ where $\Lambda(\cdot)$ is the logistic CDF. Derive the asymptotic properties of the estimator - i.e. show it is consistent for ATT, asymptotically normal, and provide a consistent estimator of the asymptotic variance.

d. A double robust estimator of the ATT is

$$\widehat{\text{ATT}}_{dr} = \frac{\frac{1}{n} \sum_{i=1}^n \left[(D_i - \frac{\widehat{p}(X_i)}{1-\widehat{p}(X_i)}(1 - D_i) \right] (Y_i - \widehat{g}_0(X_i))}{\frac{1}{n} \sum_{i=1}^n D_i}.$$

Use the models for $g_1(X)$, $g_0(X)$ and $p(X)$ from parts b. and c. Derive the asymptotic properties of the estimator - i.e. show it is consistent for ATT, asymptotically normal, and provide a consistent estimator of the asymptotic variance.

3. In experimental studies, non-compliance arises when participating in the treatment cannot be enforced. For example, a person can be assigned to a treatment group for receipt of a job training program but cannot be obligated to actually show up for the training. Let $T \in \{0, 1\}$ denote random assignment to the treatment group and $D \in \{0, 1\}$ denote receipt of treatment. By random assignment, we mean that both the exclusion restriction - $Y_{1,1} = Y_{1,0} = Y_1$ and $Y_{0,1} = Y_{0,0} = Y_0$ for $Y_{d,t}$ the potential outcome in participation state d and treatment assignment state t - and independence condition - $(Y_{1,1}, Y_{1,0}, Y_{0,1}, Y_{0,0}, D_1, D_0) \perp T$ where D_t are potential participation states in treatment state t - are satisfied. Let Y denote the observed outcome. Suppose that compliance is one-sided in that the treatment cannot be received by an individual who was not assigned to the treatment group. Note that this implies that $D_0 \equiv 0$ and $D_1 \geq D_0$ by construction.

a. Explain why the average effect of the treatment is not generally identified or estimable within this context.
b. As an alternative to the average effect of treatment, a commonly estimated object is the intention to treat effect (ITT), ITT $= \mathrm{E}[Y|T = 1] - \mathrm{E}[Y|T = 0]$. Show that under the one-sided noncompliance setup above, $|ITT| \leq |ATT|$ where ATT denotes the average treatment effect on the treated, $\mathrm{E}[Y_1 - Y_0|D = 1]$.
c. Show that LATE = ATT in the one-sided noncompliance setup. State in one sentence without too much jargon why this makes sense. (Hint: You might want to use the jargon "compliers" and talk about who the compliers are in this setup.)

4. Use the data in jtrain3.raw to answer this question. The treatment variable in this example is *train*. The outcome of interest is *unem78*.

a. Obtain an estimate of the effect of training on unemployment assuming that training is randomly assigned and the associated standard error. Under these assumptions, is there evidence that training has the anticipated average effect on unemployment?

b. Obtain a regression adjusted estimate of the average treatment effect and associated standard error. How do these compare to the results in a.? (You may assume that a linear model in the controls *age*, *educ*, *black*, *hisp*, *re74*, and *re75* among the treated and control observations is sufficient.)

c. Repeat the exercise in part b. using only the subsample of males who were unemployed in both 1974 and 1975. How do these estimates compare to the estimates in part b.?

d. Estimate the propensity score using a logit model and the explanatory variables from part b. How many outcomes are perfectly predicted? What does this result mean in terms of the overlap assumption? What would the consequences be for forming the propensity score weighted treatment effect estimate?

e. Estimate the propensity score using only observations with $avgre \leq 15$. Form the propensity score weighted estimator of the average treatment effect (using only observations with $avgre \leq 15$ of course). How do the results compare to those in b. and c.? Is it fair in general to compare such results? (I.e. carefully explain what each estimator is estimating in each part and how they are the same or different.)

f. Calculate the regression adjustment estimator of the average treatment effect using only those observations with $avgre \leq 15$ and the specification from b. and the associated standard error. How do these results compare to those in e.? Does this comparison have any bearing on assumptions underlying the validity of the estimators for understanding causal effects? Explain.

g. Calculate the doubly robust of the average treatment effect using only those observations with $avgre \leq 15$,, the regression functions from f., and the logit model from e. for the propensity score and the associated standard error. How do these results compare to those in e. and f.? Does this comparison have any bearing on assumptions underlying the validity of the estimators for understanding causal effects? Explain.

5. Use the data in fertil.raw (a sample of women of child-bearing age in Botswana) to answer this question. In these data, we would like to understand the causal effect of education on fertility. Use *children* as the dependent variable and *educ7* (a binary variable for having at least 7 years of education) as the treatment variable. Education is unlikely to be exogenous relative to fertility

decisions. A candidate for an instrument is *frsthlf* - a dummy for being born in the first half of the year.

a. What conditions do we need *frsthlf* to satisfy if we are to use it as an instrument to identify the LATE? Do these conditions seem likely to be satisfied in this example? Justify, commenting briefly on each condition.
b. Maintaining that *frsthlf* is a valid instrument, provide an estimate of the LATE and associated standard error. Provide an estimate of the strength of the underlying first-stage as well.
c. Maintaining that *frsthlf* is a valid instrument, discuss the features of the complier population underlying the LATE. How representative does this population seem to be? Does looking at these features help in interpreting the LATE? Explain in the context of this example.

6. Use the data in headstart.dta to answer this question. These data come from Calonico, Cattaneo, Farrell, and Titiunik (2016) reanalysis of Ludwig, J., and Miller, D. L. (2007), "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design". In this study, the unit of observation is the U.S. county, the treatment is receiving technical assistance to apply for Head Start funds, and the running variable is the county-level poverty index constructed in 1965. The RD design arises because the treatment was given only to counties whose poverty index was $x^* = 59.1984$ or above, a cutoff that was chosen to ensure that the 300 poorest counties received the treatment. The outcome of interest is the child mortality rate (for children of age five to nine) due to causes affected by Head Start's health services component. In the data set, the outcome variable is "mort_age59_related_postHS" (y) and the running variable is "povrate60" (x). There are also a few additional county-level demographic variables in the data.

a. Form a plot of y against x that provides visual evidence regarding the magnitude of discontinuity in $\mathrm{E}[Y|X = 59.1984]$. Form a similar plot with each of the county level demographic variables as the outcome. Does it appear that there is a discontinuity in the expected value of the county level demographics at x^*? Explain how looking at these additional plots provides a robustness check for the identifying assumption for the regression discontinuity design.

b. Miller and Ludwig (2005, NBER Working Paper Version) suggest a cross-validation procedure targeted to estimation in the regression discontinuity setting. Let $\widehat{m}_h(x) = 1(x < c)\widehat{\alpha}_{h,-}(x) + 1(x \geq c)\widehat{\alpha}^{h,+}(x)$ where $\widehat{\alpha}_{h,-}(x)$ is the local linear estimator of $\mathrm{E}[Y|X = x]$ using bandwidth h and only observations with $X_i < x$ and $\widehat{\alpha}_{h,+}(x)$ is the local linear estimator of $\mathrm{E}[Y|X = x]$ using bandwidth h and observations with $X_i > x$. Further let $\mathrm{med}(X)_-$ be the median of the X_i with $X_i < c$ and $\mathrm{med}X_+$ be the median of the X_i with $X_i \geq c$. The cross-validation criterion is then $CV(h) = \sum_{i:\mathrm{med}(X)_- \leq X_i \leq \mathrm{med}(X)_+} (Y_i - \widehat{m}_h(X_i))^2$. [In other words, observations are deleted so that the forecast being evaluated is always a boundary forecast, where the forecast is for an upper bound when $X_i < c$ and for a lower bound when $X_i \geq c$.] Use this criterion to calculate a bandwidth, $\widehat{h}$, for this example. Use a boundary (triangular) kernel, $K_h(u) = h^{-1}\mathbf{1}(|u/h| \leq 1)(1 - |u/h|)$. Obtain point estimates and standard errors for the treatment effect at the cutoff point using $.5\widehat{h}$, $\widehat{h}$, and $1.5\widehat{h}$. How many observations go into estimation (above and below the cutoff) in each case. Are the results substantively different? Which results do you think one should prefer? Explain.