Problem Set 2: Treatment Effects

Chris Conlon

Due: Two Weeks

Question 1: Treatment Effects

I have put in this directory a file `datapaps3.txt`. It has 10,000 observations of $(y_i, D_i, x_{1i}, x_{2i})$, where $D_i = 0, 1$ is a treatment variable, y_i the outcome, and x_{1i} and x_{2i} are exogenous explanatory variables.

1. Give me your best considered estimate of the treatment effect $TE(x_1, x_2)$.
2. Test for conditional independence.
3. Compare your estimates under (CIA) and under selection on unobservables.

Question 2: Regression Discontinuity Design

Use the RDD checklist from Lee and Lemieux (2010) and the data **yelp.Rdata** to estimate the causal effect of an additional star (not half-star) on the yelp platform. Your data contain the following variables:

logrev monthly store revenue in (log) dollars.

stars the number of displayed stars on the Yelp site

score this is the true Yelp score that is rounded to produce *stars*

rest_id this is the restaurant identifier (1 to 1500)

time this is the time identifier (1 to 10).

Question 3: Instrumental Variables, Experiments and Quasi-Experiments.

This question is based on the article *The Oregon Experiment - Effects of Medicaid on Clinical Outcomes*, by Katherine Baicker, Ph.D., Sarah L. Taubman, Sc.D., Heidi L. Allen, Ph.D., Mira Bernstein, Ph.D., Jonathan H. Gruber, Ph.D., Joseph P. Newhouse, Ph.D., Eric C. Schneider, M.D., Bill J. Wright, Ph.D., Alan M. Zaslavsky, Ph.D., and Amy N. Finkelstein, Ph.D. (*The New England Journal of Medicine 2013, 368: 1713-1722*)

Despite the imminent expansion of Medicaid coverage, the effects of expanding this coverage are unclear. This question will ask you for alternative procedures to estimate the effect of Medicaid enrollment on the health outcomes of individuals, as measured by their blood-pressure.

Define T_i as a dummy variable that takes value 1 if individual i is enrolled in the Medicare program. Define Y_i as the blood-pressure of individual i. Define W_i as the income of individual i, and S_i as the size of the household to whom individual i belongs. Impose the following assumptions:

1. *Assumption 1.* The expectation of Y_i conditional on (T_i, W_i, S_i) is determined by

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 W_i + \beta_3 S_i + \varepsilon_i, \quad \beta_1 > 0, \beta_2 > 0, \beta_3 > 0,$$

   with $\mathbb{E}[\varepsilon_i | T_i, W_i, S_i] = 0$.

2. *Assumption 2.* Enrollment in the Medicare program is determined through a random lottery drawing among a group of individuals included on a waiting list.

3. *Assumption 3.* Individuals who win the lottery are automatically enrolled in the Medicare program.

4. *Assumption 4.* Entitlement applies only to lottery winners. In particular, it does not automatically include other household members.

5. *Assumption 5.* Each individual on the waiting list had to voluntarily apply to be included in such list, and pay a small fee.

Define G_1 as the group of individuals that were enrolled in Medicare through this lottery (i.e. individuals who won the lottery), G_2 as the group of individuals who were on the waiting list but did not win the lottery, and G_3 as the group of individuals who did not apply to be included on the waiting list. Answer the following questions:

(a) Definition of treatment and control groups.

1. Imagine you observe a random sample of each of the three groups (G_1, G_2, G_3). Under *Assumptions 1 to 5*, which of the three groups, (G_1, G_2, G_3), would you use as treatment group ($T_i = 1$)? Which of the three groups, (G_1, G_2, G_3), would you use as control group ($T_i = 0$)?
2. How would your answer change if, instead of *Assumption 5*, being on the waiting list was also random?

(b) Definition of the regression equation.

For each question below, indicate: (1) the regression you would use to estimate the effect of T_i on Y_i; (2) the parameter of your regression that captures the effect of T_i on Y_i; (3) the estimator you would use to estimate the parameters of your regression.

1. Under *Assumptions 1 to 5*, which regression would you run? Can you run this regression if you do not observe W_i and S_i?
2. How would your answer change if, instead of *Assumption 4*, winning the lottery implied automatic enrollment for those individuals who won the lottery *and* for all the other members of their households. Can you run this regression if you do not observe W_i and S_i?

3. How would your answer change if, instead of *Assumption 3*, winning the lottery only implied a discount in the Medicare premium? Assume that, in this case, not everyone who wins the lottery ends up enrolling in the Medicare program. Define $\tilde{T}_i = 1$ if individual i won the lottery (and $\tilde{T}_i = 0$ otherwise), and remember that $T_i = 1$ if individual i actually enrolled the Medicare program. Assume that both $\tilde{T}_i$ and T_i are observed in your sample.
4. How would your answer change if, instead of *Assumptions 2 and 3*, every individual whose income level is below \$30,000 per year was automatically enrolled in the medicare program? (i.e. assume that there is no lottery determining Medicare enrollment).

Question 4: Local Average Treatment Effect (LATE) and Average Treatment Effect (ATE).

Assume that

1. The following equations hold
$$\begin{aligned} Y_i &= \beta_0 + \beta_{1i} X_i + u_i, \\ X_i &= \pi_0 + \pi_{1i} Z_i + v_i. \end{aligned}$$
2. The vector (β_{1i}, π_{1i}) is independent of the vector (u_i, v_i, Z_i).
3. $\mathbb{E}[u_i|Z_i] = 0$.
4. $\mathbb{E}[v_i|Z_i] = 0$.

Answer the following questions

(a) Prove that
$$\frac{\mathbb{E}\big[(Y_i - \mathbb{E}(Y_i))(Z_i - \mathbb{E}(Z_i))\big]}{\mathbb{E}\big[(X_i - \mathbb{E}(X_i))(Z_i - \mathbb{E}(Z_i))\big]} = \frac{\mathbb{E}\big[\beta_{1i}\pi_{1i}\big]}{\mathbb{E}\big[\pi_{1i}\big]}.$$

(b) Indicate three different **sufficient** assumptions under which:
$$LATE = \frac{\mathbb{E}\big[\beta_{1i}\pi_{1i}\big]}{\mathbb{E}\big[\pi_{1i}\big]} = \mathbb{E}[\beta_{1i}] = ATE$$

(c) Assume that, for every individual i in the population of interest,
$$\begin{aligned} \beta_{1i} &= \beta^H \text{ with probability } 0.5, \\ \beta_{1i} &= \beta^L \text{ with probability } 0.5, \end{aligned}$$
with $\beta^H > \beta^L$. Analogously, assume that, for every individual i in the population of interest,
$$\begin{aligned} \pi_{1i} &= \pi^H \text{ with probability } 0.5, \\ \pi_{1i} &= \pi^L \text{ with probability } 0.5, \end{aligned}$$
with $\pi^H > \pi^L$. Indicate which of the following three statements is true and why

A. LATE > ATE.

B. LATE < ATE.

C. We do not have enough information to know whether LATE > ATE or LATE < ATE.

(d) Assume that, for every individual i in the population of interest,
$$\begin{aligned} (\beta_{1i}, \pi_{1i}) &= (\beta^H, \pi^H) \text{ with probability } 0.25, \\ (\beta_{1i}, \pi_{1i}) &= (\beta^H, \pi^L) \text{ with probability } 0.25, \\ (\beta_{1i}, \pi_{1i}) &= (\beta^L, \pi^L) \text{ with probability } 0.25, \\ (\beta_{1i}, \pi_{1i}) &= (\beta^L, \pi^H) \text{ with probability } 0.25, \end{aligned}$$
with $\beta^H > \beta^L$ and $\pi^H > \pi^L$. Indicate which of the following three statements is true and why

A. LATE > ATE.

B. LATE < ATE.

C. We do not have enough information to know whether LATE > ATE or LATE < ATE.

(e) Assume that, for every individual i in the population of interest,

$$\begin{aligned}
(\beta_{1i}, \pi_{1i}) &= (\beta^H, \pi^H) \text{ with probability } 0.5,\\
(\beta_{1i}, \pi_{1i}) &= (\beta^H, \pi^L) \text{ with probability } 0,\\
(\beta_{1i}, \pi_{1i}) &= (\beta^L, \pi^L) \text{ with probability } 0.5,\\
(\beta_{1i}, \pi_{1i}) &= (\beta^L, \pi^H) \text{ with probability } 0,
\end{aligned}$$

with $\beta^H > \beta^L$ and $\pi^H > \pi^L$. Indicate which of the following three statements is true and why

A. LATE > ATE.

B. LATE < ATE.

C. We do not have enough information to know whether LATE > ATE or LATE < ATE.

(f) Assume that, for every individual i in the population of interest,

$$\begin{aligned}
(\beta_{1i}, \pi_{1i}) &= (\beta^H, \pi^H) \text{ with probability } 0.5,\\
(\beta_{1i}, \pi_{1i}) &= (\beta^H, \pi^L) \text{ with probability } 0.5,\\
(\beta_{1i}, \pi_{1i}) &= (\beta^L, \pi^L) \text{ with probability } 0,\\
(\beta_{1i}, \pi_{1i}) &= (\beta^L, \pi^H) \text{ with probability } 0,
\end{aligned}$$

with $\beta^H > \beta^L$ and $\pi^H > \pi^L$. Indicate which of the following three statements is true and why

(a) LATE > ATE.

(b) LATE < ATE.

(c) We do not have enough information to know whether LATE > ATE or LATE < ATE.