Problem Set 5

Due November 18th, 6:30 pm

1. Linear expansions of the matching estimators

Lecture 12 states that $\hat{\tau}^{\text{mbc}} = n^{-1} \sum_{i=1}^{n} \hat{\psi}_i$ and $\hat{\tau}_{\text{T}}^{\text{mbc}} = n_1^{-1} \sum_{i=1}^{n} \hat{\psi}_{\text{T},i}$ without proving them. Prove these results.

2. Z-bias

Lecture 13 Section 4.2 gives the bias of τ_{unadj} and τ_{adj} without proving them. Prove these results.

3. Cochran's formula or the omitted variable bias formula (for Stat 260 only)

The following result is due to Yule and Fisher, although Sir David Cox calls it Cochran's formula and econometricians call it the omitted variable bias formula. It is also a sister of the Frisch–Waugh–Lovell Theorem.

The formula has two versions. All vectors are column vectors, as in R.

(1) (Population version) Assume $(y_i, x_{1i}, x_{2i})_{i=1}^{n}$ are iid, where y_i is a scalar, x_{i1} has dimension K, and x_{i2} has dimension L.

We have the following OLS decompositions of random variables

$$\begin{aligned} y_i &= \beta_1' x_{i1} + \beta_2' x_{2i} + \varepsilon_i, && (1) \\ y_i &= \gamma' x_{i1} + e_i, && (2) \\ x_{i2} &= \delta' x_{i1} + v_i. && (3) \end{aligned}$$

Equation (1) is called the long regression, and Equation (2) is called the short regression. In Equation (3), δ is a matrix because it is a regression of a vector on a vector. You can view (3) as regression of each component of x_{i2} on x_{i1}.

Show that $\gamma = \beta_1 + \delta\beta_2$.

(2) (Sample version) We have an $n \times 1$ vector Y, an $n \times K$ matrix X_1, and an $n \times L$ matrix X_2. We do not assume any randomness. All results below are purely linear algebra.

We can fit the following OLS, for example, using R, to obtain

$$\begin{aligned} Y &= X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}, \\ Y &= X_1\hat{\gamma} + \hat{e}, \\ X_2 &= X_1\hat{\delta} + \hat{v}, \end{aligned}$$

where $\hat{\varepsilon}, \hat{e}, \hat{v}$ are the residuals. Again, the last OLS fit means the OLS fit of each column of X_2 on X_1, and therefore the residual $\hat{v}$ is an $n \times L$ matrix.

Show that $\hat{\gamma} = \hat{\beta}_1 + \hat{\delta}\hat{\beta}_2$.

4. Simulation for RDD

`RDDnumericexamples.R` simulates potential outcomes from linear models. Change them to nonlinear models, and compare different point estimators and confidence intervals, including the biases and variances of the point estimators, and the coverage properties of confidence intervals.

5. Data analysis: Sommer and Zeger (1991)

Re-analyze the data in Table 23.1 of the Imbens–Rubin book. Note that W_i^{obs} is the D_i and Y_i^{obs} is the Y_i in my lecture notes.

6. Data analysis: a flu shot encouragement design (McDonald et al. 1992)

For the details, you can read Chapter 25.2 of the Imbens–Rubin book. Chapter 25 of the book uses a complicated model-based analysis. You can ignore it.

In `fludata.txt`, the variables are the treatment assigned, treatment received, outcome, and pretreatment covariates. Analyze the data with and without using covariates.

7. Data analysis: the Karolinska data

Rubin (2008) used the Karolinska data as an example for the instrumental variable method. In `karolinska.txt`, whether a patient was diagnosed at large volume hospital can be viewed as an instrumental variable for whether a patient was treated at a large volume hospital. This is plausible at least conditioning on other observed covariates. See Rubin (2008)'s analysis for more details.

Reanalyze the data assuming that the instrumental variable is randomly assigned conditional on observed covariates.

REFERENCES

McDonald, C., Hui, S. L., and Tierney, W. (1992). Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD computing: computers in medical practice*, 9(5):304–312.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, pages 808–840.

Sommer, A. and Zeger, S. L. (1991). On estimating efficacy from clinical trials. *Statistics in medicine*, 10(1):45–52.