ECMA 31100: Problem Set 1

Due January 29 by 11:59PM

Question 1 You observe wages (y_i), whether an individual took job training ($d_i \in \{0, 1\}$), and a scalar covariate x_i measuring prior labor market outcomes in an iid sample of n individuals. Define the potential outcomes y_{i0}, y_{i1} as

$$y_{i0} = \text{wage of individual } i \text{ without job training},$$
$$y_{i1} = \text{wage of individual } i \text{ with job training}.$$

You are interested in studying the average treatment effect, $\mathrm{E}(y_{i1} - y_{i0})$. Assume that d_i is randomly assigned conditional on x_i. That is:

$$(y_{i0}, y_{i1}) \perp\!\!\!\perp d_i | x_i.$$

a) How plausible is the assumption that $y_{i1} - y_{i0}$ is constant across individuals?

b) How plausible are the assumptions $(y_{i0}, y_{i1}) \perp\!\!\!\perp d_i$ and $(y_{i0}, y_{i1}) \perp\!\!\!\perp d_i | x_i$?

Suppose in addition that $\mathrm{E}(y_{i1} - y_{i0}|x_i) = c$, for some constant c.

ci) Is the assumption $\mathrm{E}(y_{i1} - y_{i0}|x_i) = c$ stronger or weaker than assuming $y_{i1} - y_{i0} = c$?

cii) Show that $\mathrm{E}(y_i|d_i = 1, x_i) - \mathrm{E}(y_i|d_i = 0, x_i) = c$.

Suppose in addition that $\mathrm{E}(y_{i0}|x_i) = \beta_0 + \beta_1 x_i$ for some unknown constants β_0, β_1.

d) Show that we may write

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + e_i, \tag{1}$$

where $\mathrm{E}(e_i|d_i, x_i) = 0$. Show that β_2 equals the average treatment effect.

Now drop the assumption that $\mathrm{E}(y_{i1} - y_{i0}|x_i)$ is constant, but instead assume that x_i takes two values $x_i \in \{0, 1\}$.

e) Argue that the representation in part d) is not necessarily valid, but the following representation is valid:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 d_i x_i + v_i; \quad \mathrm{E}(v_i|d_i, x_i) = 0. \tag{2}$$

What is β_2? Show that β_3 is a difference between (conditional) average treatment effects.

f) Suppose you estimate the coefficients in the regression

$$y_i = b_0 + b_1 x_i + b_2 d_i + e_i,$$

using OLS. Show that $\hat{b}_2$ converges to a weighted average of conditional average treatment effects, and generally not the ATE.

Question 2 Consider the potential outcomes framework

$$Y = Dy_1 + (1 - D)\, y_0$$

and suppose that $(y_0, y_1) \perp D|X$, where X is a vector of covariates. Define $ATE(x) := \mathrm{E}(y_1 - y_0|X = x)$. You are given an iid sample of $\{Y_i, D_i, X_i\}_{i=1}^N$. Assume that all relevant moments exist.

a) In class we provided an identification argument for the ATE. Do the same for $ATE(x)$. Use this to provide an identification argument for the ATT and ATU.

From now on assume $\mathrm{E}(y_d|X = x) = \alpha_d + x'\beta_d$ for $d = 0, 1$.

b) How do your answers to part a) change?

c) Write $\mathrm{E}(Y|D, X)$ in terms of $\alpha_0, \alpha_1, \beta_0, \beta_1, D, X$. Propose consistent estimators of the ATE, ATT and ATU and prove their consistency.

d) Show that

$$\mathrm{E}(Y|D, X) = \gamma_0 + \gamma_1 D + X'\gamma_2 + D \cdot (X - \mathrm{E}(X))'\gamma_3$$

for some values of $\gamma_0, \gamma_1, \gamma_2, \gamma_3$ you derive, and show that γ_1 represents the ATE. How would you modify the above construction to obtain γ_1 as the ATT or ATU?

e) How might you estimate γ_1 directly? You will need to replace $\mathrm{E}(X)$. Show that your estimator of γ_1 is consistent.

f) Derive $Var(Y|D, X)$ in terms of $Var(y_0|X), Var(y_1|X)$ and D. Is it reasonable to assume homoskedasticity (i.e. that $Var(Y|D, X)$ is constant)?

g) Suppose $\mathrm{E}(X)$ is known, so that it is in fact possible to run the regression

$$Y_i = W_i'\gamma + \epsilon_i; \quad \mathrm{E}(\epsilon_i|W_i) = 0,$$

where $W_i = \left(1, D_i, X_i', D_i \cdot [X_i - \mathrm{E}(X)]'\right)'$. Show that

$$\sqrt{n}\,(\hat{\gamma}_{OLS} - \gamma) \xrightarrow{d} \mathcal{N}(0, V),$$

for some matrix V you specify. Provide a consistent estimate of V under homoskedasticity and another under heteroskedasticity.

h) Show that when $\mathrm{E}(X)$ is replaced by $\bar{X}_n$, so that the regression actually run is feasible, you still obtain

$$\sqrt{n}\,(\hat{\gamma}_{1,OLS} - \gamma_1) \xrightarrow{d} \mathcal{N}(0, \theta),$$

for some θ. Construct a consistent estimate of θ.

Hint: Find the joint asymptotic distribution of the OLS estimates $\left(\hat{\alpha}_0, \hat{\alpha}_1 - \hat{\alpha}_0, \hat{\beta}_0, \hat{\beta}_1 - \hat{\beta}_0\right)$ and $\bar{X}_n$, then use the continuous mapping theorem.

i) Use your previous answers to construct a test of $H_0 : \gamma_1 = 0$ vs. $H_1 : \gamma_1 \neq 0$ that is asymptotically of size α and justify your procedure formally.

j) Download the dataset jtrain98.dta from Canvas, and reproduce the regression results using the covariates used in class, this time allowing for the conditional means to have different slopes. Does this alteration drastically change the estimate of the ATE from the one presented in class? Is the result statistically significant at a 5% level?

Note: You will need to compute your consistent estimate of θ.

Question 3 Using the same setting as in Question 2, define $p(x) := \mathrm{P}(D = 1|X = x)$ and show that we can represent the ATE as

$$ATE = \mathrm{E}\left(\frac{Y(D - p(X))}{p(X)(1 - p(X))}\right).$$

Find similar representations for the ATT and ATU.

Question 4 The dataset card.dta (available on canvas) contains data on wages, years of schooling and other observed characteristics of 2946 men with at least than 8 years of education. Consider the following model:

$$y = \beta_0 + \beta_1 x_1 + w'\gamma + u,$$

where $y = \ln(wage)$, $x_1 =$ years of schooling and w is a set of included instruments that varies by specification. Estimate the following specifications using 2SLS and the optimal GMM estimator from Lecture 3. In each specification the dependent variable is y, and the regressors include a constant, the endogenous variable x_1 and the following sets of included instruments. Report the coefficient estimate on x_1, as well as 95% confidence intervals for this coefficient, assuming homoskedasticity in the TSLS case and heteroskedasticity for the GMM estimates.

Specification 1: Included instruments: none; Excluded instruments: nearc4

Specification 2: Included instruments: south, smsa; Excluded instruments: nearc4

Specification 3: Included instruments: south, smsa; Excluded instruments: nearc4, nearc2

Specification 4: Included instruments: south, smsa, libcrd14, IQ, KWW, exper, expersq; Excluded instruments: nearc4, nearc2.