Recovering Distributions in DID models with a Linear Factor Structure

Bonhomme and Sauder (REStat, 2011)

- Estimators proposed in Attey and Imbens (Econometrica, 2006) relies on assumptions that are not satisfied in many situations

Bonhomme and Sauder (REStat, 2011)

- Estimators proposed in Attey and Imbens (Econometrica, 2006) relies on assumptions that are not satisfied in many situations
- In this lesson we discuss an alternative estimator for models with a linear factor such as those usually considered in education (human capital) production function literature

Bonhomme and Sauder (REStat, 2011)

- Estimators proposed in Attey and Imbens (Econometrica, 2006) relies on assumptions that are not satisfied in many situations
- In this lesson we discuss an alternative estimator for models with a linear factor such as those usually considered in education (human capital) production function literature
- Bonhomme and Sauder (REStat, 2011) use this model to measure the effect of selective education on children's outcomes

Bonhomme and Sauder (REStat, 2011)

- Estimators proposed in Attey and Imbens (Econometrica, 2006) relies on assumptions that are not satisfied in many situations
- In this lesson we discuss an alternative estimator for models with a linear factor such as those usually considered in education (human capital) production function literature
- Bonhomme and Sauder (REStat, 2011) use this model to measure the effect of selective education on children's outcomes
- Better performance of selective schools relative to nonselective ones is essentially due to differences in pupil's composition

The setting

- Local Education Authority (LEA) decides which school each kid should attend after age 11
- Selective (grammar school or secondary modern) Vs comprehensive system
- All schools in selective system since 1945: some started to switch in 1965 with substantial variation within and between LEA

The setting

- Local Education Authority (LEA) decides which school each kid should attend after age 11
- Selective (grammar school or secondary modern) Vs comprehensive system
- All schools in selective system since 1945: some started to switch in 1965 with substantial variation within and between LEA
- Many studies used a value-added methodology to evaluate this policy

The setting

- Local Education Authority (LEA) decides which school each kid should attend after age 11
- Selective (grammar school or secondary modern) Vs comprehensive system
- All schools in selective system since 1945: some started to switch in 1965 with substantial variation within and between LEA
- Many studies used a value-added methodology to evaluate this policy
- They compare outcomes for students passing through either type of school, controlling for achievement levels at the time of entering secondary education
- Potential issues with this approach?

The setting

- Unlikely to successfully eliminate selection effects in who attends what type of school
- Manning and Pischke(2006) find + effect of selective school on test scores at age 11 using VA approach. What does it mean?

The setting

- Unlikely to successfully eliminate selection effects in who attends what type of school
- Manning and Pischke(2006) find $+$ effect of selective school on test scores at age 11 using VA approach. What does it mean?
- $\Longrightarrow$ attending selective school is likely to be correlated with unobservables that affect latter outcomes: need to deal with that

A Model of Test Scores

- 2 periods: period 1 (age 11 - before secondary education); period 2 (age 16 - after)
- Y_{it}: Test score measured at period t and D_i treatment variable (attending a selective school)
- Test score model (simplified): Linear factor model

$$\begin{aligned} Y_{i2}^0 &= g_2^0(X_i, \eta_i, \nu_{i2}^0) \sim \alpha_2^0 + \eta_i + \nu_{i2}^0 \\ Y_{i1} &= g_1(X_i, \eta_i, \nu_{i1}) \sim \alpha_1 + \eta_i + \nu_{i1} \end{aligned}$$

- X_i: Observed characteristics (parental, school, local characteristics)
- η_i: child endowment (cognitive ability), potentially correlated with X_i and D_i
- νs potentially correlated with each other

- η_i: acts as a confounder since it's not observed by the econometrician
- X_i does not include characteristics of secondary school attended: why? what does it mean for estimated effect?

- η_i: acts as a confounder since it's not observed by the econometrician
- X_i does not include characteristics of secondary school attended: why? what does it mean for estimated effect?
- Estimated effect will capture differences in school characteristics (teacher quality, class size, ...) and other factors (grouping students by ability levels,...)

- η_i: acts as a confounder since it's not observed by the econometrician
- X_i does not include characteristics of secondary school attended: why? what does it mean for estimated effect?
- Estimated effect will capture differences in school characteristics (teacher quality, class size, ...) and other factors (grouping students by ability levels,...)
- Standard approaches do not apply
- Given data on (Y_{i2}, Y_{i1}, D_i), we can study identification of the entire counterfactual distribution of potential outcomes $Y_{i2}^0 | D_i = 1$

- **Assumption 1:** ν_{i1} and ν_{i2}^0 are independent of D_i
- Differences in pretreatment outcomes reflect only differences in η_i (same for Y_{i2}^0 and Y_{i2})
- Assumption 1 $\implies$ we can recover mean potential outcome

$$E(Y_{i2}^0|D_i = 1) = E(Y_{i1}|D_i = 1) + (E(Y_{i2}|D_i = 0) - E(Y_{i1}|D_i = 0)) \quad (1)$$

- ATT is given by stantard DID estimator under Assumption 1
$\Delta = E(Y_{i2}|D_i = 1) - \left[E(Y_{i1}|D_i = 1) + (E(Y_{i2}|D_i = 0) - E(Y_{i1}|D_i = 0))\right]$

- Need to make another assumption to recover distribution of $Y_{i2}^0|D_i = 1$
- **Assumption 2:** ν_{i1} and ν_{i2}^0 are independent of η_i given D_i
- **Assumption 3:** The characteristic function of $Y_{i1}|D_i = 0$ is nonvanishing on $\mathbb{R}$
- The characteristic function of a RV W is a complex-valued function: $\psi_W(t) = E(exp(jtW))$ where $j = \sqrt{-1}$
- And pdf of W is given by the inverse Fourier Transform

$$f_W(w) = \frac{1}{2\pi}\int exp(-jtw)\psi_W(t)dt \tag{2}$$

- Assumption 3 is a technical assumption that just requires that the characteristic function not to have any real zero

- **Theorem 1:** Under Assumptions 1, 2 and 3, we have

$$\psi_{Y_{i2}^0|D_i=1}(t) = \frac{\psi_{Y_{i1}|D_i=1}(t)}{\psi_{Y_{i1}|D_i=0}(t)}\psi_{Y_{i2}|D_i=0}(t) \tag{3}$$

- Each of the characteristic function on the RHS can be estimated given a random sample of (Y_{i2}, Y_{i1}, D_i)

- Taking logs of Equation (3) shows that Theorem 1 generalizes Equation(1) to the entire distribution

$$log\left[\psi_{Y_{i2}^0|D_i=1}(t)\right] = log\left[\psi_{Y_{i1}|D_i=1}(t)\right] - log\left[\psi_{Y_{i1}|D_i=0}(t)\right] + log\left[\psi_{Y_{i2}|D_i=0}(t)\right]$$

- First derivative previous equation at 0 yields Equation (1)

- Same logic in both equations: We need to correct $Y_{i2}|D_i = 0$ for the fact that treatment and control group do not have same distribution of unobservables

- This is done by adding the distributional characteristic of $Y_{i1}|D_i = 1$ and subtracting the one of $Y_{i1}|D_i = 0$

A Model of Test Scores

- pdf of the entire distribution of the potential outcome is therefore identified: just replace the identified characteristic function into Equation (2)

$$f_{Y_{i2}^0|D_i=1}(y) = \frac{1}{2\pi}\int exp(-jty)\left[\frac{\psi_{Y_{i1}|D_i=1}(t)}{\psi_{Y_{i1}|D_i=0}(t)}\psi_{Y_{i2}|D_i=0}(t)\right]dt \qquad (4)$$

- Quantile treatment effect is therefore define as:

$$\Delta(\tau) = F^{-1}_{Y_{i2}|D_i=1}(\tau) - F^{-1}_{Y_{i2}^0|D_i=1}(\tau), \tau \in [0,1] \qquad (5)$$

- Why is it important to identify the entire distribution of the potential outcome?

A Model of Test Scores

- pdf of the entire distribution of the potential outcome is therefore identified: just replace the identified characteristic function into Equation (2)

$$f_{Y_{i2}^0|D_i=1}(y) = \frac{1}{2\pi}\int exp(-jty)\left[\frac{\psi_{Y_{i1}|D_i=1}(t)}{\psi_{Y_{i1}|D_i=0}(t)}\psi_{Y_{i2}|D_i=0}(t)\right]dt \qquad (4)$$

- Quantile treatment effect is therefore define as:

$$\Delta(\tau) = F^{-1}_{Y_{i2}|D_i=1}(\tau) - F^{-1}_{Y_{i2}^0|D_i=1}(\tau), \tau \in [0,1] \qquad (5)$$

- Why is it important to identify the entire distribution of the potential outcome?
- selective system is split into grammar and secondary modern schools
- Children at different points of the distribution could benefit differently from attending selective schools

Proof Theorem 1

- We use the independence property of characteristic functions: $W_1 \perp W_2 \implies \psi_{W_1+W_2}(t) = \psi_{W_1}(t)\psi_{W_2}(t)$
- Assumption 2 implies for $t \in \mathbb{R}$

$$\begin{aligned}
\psi_{Y_{i2}^0|D_i=1}(t) &= exp(j\alpha_2^0 t)\psi_{\eta_i|D_i=1}(t)\psi_{v_{i2}^0|D_i=1}(t) \\
\psi_{Y_{i2}^0|D_i=0}(t) &= exp(j\alpha_2^0 t)\psi_{\eta_i|D_i=0}(t)\psi_{v_{i2}^0|D_i=0}(t)
\end{aligned}$$

- Assumption 1 implies

$$\begin{aligned}
\psi_{Y_{i2}^0|D_i=1}(t) &= exp(j\alpha_2^0 t)\psi_{\eta_i|D_i=1}(t)\psi_{v_{i2}^0}(t) \\
\psi_{Y_{i2}^0|D_i=0}(t) &= exp(j\alpha_2^0 t)\psi_{\eta_i|D_i=0}(t)\psi_{v_{i2}^0}(t)
\end{aligned}$$

- Taking ratios: $\psi_{Y_{i2}^0|D_i=1}(t) = \frac{\psi_{\eta_i|D_i=1}(t)}{\psi_{\eta_i|D_i=0}(t)}\psi_{Y_{i2}^0|D_i=0}(t)$

Proof Theorem 1

- Assumption 3 implies this expression is well defined
- Apply the procedure to the expression of Y_{i1} gives $\psi_{Y_{i1}|D_i=1}(t) = \frac{\psi_{\eta_i|D_i=1}(t)}{\psi_{\eta_i|D_i=0}(t)} \psi_{Y_{i1}|D_i=0}(t)$
- Substituting the expression of the ratio give the result. QED

Comparison with AI (2006) CIC estimand

- AI(200) CIC estimand relies on:

$$F_{Y^0_{i2}|D_i=1}(y) = F_{Y_{i1}|D_i=1}\left[F^{-1}_{Y_{i1}|D_i=0}\left(F_{Y_{i2}|D_i=0}(y)\right)\right] \tag{6}$$

- This equation is satisfied in the test score model at hand here only if
  - Distribution of ν_{i1} and ν^0_{i2} are identical
  - or η_i independent of D_i
- These assumptions are too restrictive in this case
- Time invariance assumption (Assumption 3 in previous lesson) implies that test scores have the same shape and dispersion in each group across time
- The extra flexibility in BS(2011) comes at the cost of imposing additivity in test score model
- The new estimand is not invariant to monotone transformations of test score variable

Identification: Allowing for observed covariates

- Want to allow for the effect of covariates that are associated with the change in outcomes and are not similarly distributed between treated and controls
  - Children attending comprehensive school come from parents with low background and live in poorer areas
  - Consider X_i: set of pretreatment characteristics
  - Assumptions 1, 2 and 3 are assumed valid conditional on X_i
  - $p_D = P(D_i = 1)$ and $p_D(x) = P(D_i = 1 | X_i = x)$
  - Assumption 4: $p_D > 0$ and $p_D(X_i) < 1$ with probability 1
  - Restrict support of the propensity score
  - Restricts correlation between time-varying shocks and treatment but leave correlation between $\eta_i, \nu_{i1}, \nu_{i2}^0$ and X_i unrestricted
    - parents take η_i into account when choosing type of primary school (in X_i) for instance

Identification: Allowing for observed covariates

- Conditional characteristic function: $\psi_{W|Z}(t|z) = E\big(exp(jtW)|Z = z\big)$

Theorem 2. *Let assumptions 1, 2, and 3 hold given X_i (almost everywhere), and let assumption 4 hold. Then:*

$$\Psi_{Y_{i2}^0|D_i=1,X_i}(t|x) = \frac{\Psi_{Y_{i1}|D_i=1,X_i}(t|x)}{\Psi_{Y_{i1}|D_i=0,X_i}(t|x)} \Psi_{Y_{i2}|D_i=0,X_i}(t|x), \quad (11)$$

and

$$\Psi_{Y_{i2}^0|D_i=1}(t) = \frac{1}{p_D} \mathbb{E}[\omega(t|X_i)(1 - D_i)\exp(jtY_{i2})], \quad (12)$$

where we have denoted as

$$\omega(t|X_i) \equiv \frac{p_D(X_i)}{(1 - p_D(X_i))} \frac{\Psi_{Y_{i1}|D_i=1,X_i}(t|X_i)}{\Psi_{Y_{i1}|D_i=0,X_i}(t|X_i)} = \frac{\mathbb{E}[D_i \exp(jtY_{i1})|X_i]}{\mathbb{E}[(1 - D_i)\exp(jtY_{i1})|X_i]}. \quad (13)$$

Proof of Theorem 2. The proof of equation (11) is very similar to that of theorem 1. Indeed:

$$\begin{aligned}
\Psi_{Y_{i2}^0|D_i=1}(t) &= \mathbb{E}[\Psi_{Y_{i2}^0|D_i=1,X_i}(t|X_i)|D_i = 1] \\
&= \int \Psi_{Y_{i2}^0|D_i=1,X_i}(t|X_i)dP(X_i|D_i = 1) \\
&= \mathbb{E}\left[\frac{p_D(X_i)}{p_D}\Psi_{Y_{i2}^0|D_i=1,X_i}(t|X_i)\right] \\
&= \mathbb{E}\left[\frac{p_D(X_i)}{p_D}\frac{\Psi_{Y_{i1}|D_i=1,X_i}(t|X_i)}{\Psi_{Y_{i1}|D_i=0,X_i}(t|X_i)}\Psi_{Y_{i2}|D_i=0,X_i}(t|X_i)\right] \\
&= \frac{1}{p_D}\mathbb{E}[\omega(t|X_i)(1 - p_D(X_i))\Psi_{Y_{i2}|D_i=0,X_i}(t|X_i)] \\
&= \frac{1}{p_D}\mathbb{E}[\omega(t|X_i)(1 - D_i)\exp(jtY_{i2})],
\end{aligned}$$

where going from the second to the third line requires use of Bayes' rule, and the last equality comes from applying the law of iterated expectations.

Identification: Allowing for observed covariates

- Model (3): selection on both observables and unobservables
- Estimators based on only selection on observables are biased if distribution of η_i changes between treatment and controls

Identification: Allowing for Different Returns to Unobservables

- Benchmark model imposes that coefficient of η_i is same in equations of pre- and postreatment outcomes
- Here we may want to allow for different coefficients: ability can be differently rewarded at age 11 and 16 and to have specific return in comprehensive system

$$Y_{i2}^0 = \alpha_2^0 + \beta_2^0 \eta_i + \nu_{i2}^0,$$
$$Y_{i1} = \alpha_1 + \beta_1 \eta_i + \nu_{i1},$$

$$Y_{i2}^0 = \alpha_2^0 - \rho\alpha_1 + \rho Y_{i1} + \nu_{i2}^0 - \rho\nu_{i1},$$

where $\rho = \beta_2^0/\beta_1$ is the ratio of returns to η_i.

- Y_{i1} endogenous in this equation because of presence of contemporaneous shock ν_{i1}
- Use panel IV to deal with it:

Identification: Allowing for Different Returns to Unobservables

Assumption 5. *There exists a variable $\widetilde{Y}_{i0}$ such that*

$$\begin{cases} v_{i1} \text{ and } v_{i2}^0 \text{ are uncorrelated with } \widetilde{Y}_{i0} \text{ given } D_i = 0, \\ Y_{i1} \text{ and } \widetilde{Y}_{i0} \text{ are correlated given } D_i = 0. \end{cases}$$

- $\tilde{Y}_{i0}$ not assumed independent of η_i or of potential outcome
- In application: use lagged test scores in different subjects
- Under Assumption 5

$$\rho = \frac{\mathrm{Cov}(\widetilde{Y}_{i0}, Y_{i2} | D_i = 0)}{\mathrm{Cov}(\widetilde{Y}_{i0}, Y_{i1} | D_i = 0)}.$$

- With ρ identified, we get the entire distribution

$$f_{Y_{i2}^0|D_i=1}(y) = \frac{1}{2\pi}\int \exp(-jty)\left[\frac{\Psi_{Y_{i1}|D_i=1}(\rho t)}{\Psi_{Y_{i1}|D_i=0}(\rho t)}\Psi_{Y_{i2}|D_i=0}(t)\right]dt,$$

Non-linearities in production function

- Linearity of production function is a common assumption but can be a strong one
- Unlike in AI(2006), estimator used here not invariant to monotone transformations of test scores
- Need to test robustness of result to other transformations in practice

$$h(Y_{i2}^0; \lambda_2^0) = \alpha_2^0 + \beta_2^0\eta_i + v_{i2}^0,$$
$$h(Y_{i1}; \lambda_1) = \alpha_1 + \beta_1\eta_i + v_{i1},$$

- h, λ_1 and λ_2^0 are known so we can recover distribution of potential outcome

Estimation

- Assume we have a random sample $(Y_{i2}, Y_{i1}, D_i), i = 1, ..., N$
- Pointwise estimate of Equation (9):

$$\hat{f}_{Y^0_{i2}|D_i=1}(y) = \frac{1}{2\pi}\int_{-T_N}^{T_N} exp(-jty)\left[\frac{\hat{\psi}_{Y_{i1}|D_i=1}(t)}{\hat{\psi}_{Y_{i1}|D_i=0}(t)}\hat{\psi}_{Y_{i2}|D_i=0}(t)\right]dt \qquad (7)$$

- Empirical characteristic function in control group is for instance:

$$\hat{\psi}_{Y_{i2}|D_i=0}(t) = \frac{1}{N_0}\sum_{i:D_i=0} exp(jtY_{i2})$$

- T_N is a trimming parameter that ensures integral in Equation (7) is finite

Estimation

- Estimator has relative slow rate of convergence: price to pay for not making more assumptions
- See paper for empirical application!