

Policy Evaluation Supervision 2

Samuel Lee

Question 1

Whether the expected value of an IV estimate is larger or smaller than that of the OLS estimate entirely depends on what the instrument is and what it is instrumenting for. It is worth stating at the outset that the IV estimator is typically biased even if valid, that is, $E[u_i|Z_i] = 0$. Validity only assures consistency; we have

$$E[\hat{\beta}_{IV}|X, Z] = E[(Z'X)^{-1}Z'y|X, Z] = E[(Z'X)^{-1}Z'E[u|X, Z]] = \beta + (Z'X)^{-1}Z'E[u|X, Z]$$

where $E[u|Z, X] \neq 0$ if $E[u|X] \neq 0$, so iterating expectations won't give us β . This is because

$$E[u|X] = E[E[u|X, Z]|X]$$

which means $E[u|X] \neq 0 \implies E[u|X, Z] \neq 0$. So in the case where the instrument is valid, we can say whether we think the IV estimator will be biased upwards or downwards depending on what we are willing to guess about $E[u|Z, X]$. Still, it is instructive to consider the case of simple linear regression, where we have one explanatory variable x and one instrument z . In this case, we get

$$\hat{\beta}_{IV} = \frac{\sum y_i(z_i - \bar{z})}{\sum x_i(z_i - \bar{z})} = \frac{\sum (\alpha + \beta x_i + u_i)(z_i - \bar{z})}{\sum x_i(z_i - \bar{z})} = \beta + \frac{\sum u_i(z_i - \bar{z})}{\sum x_i(z_i - \bar{z})}$$

Again, we know that the expectation of the second term is not 0 since expectations don't pass through nonlinear functions, but we know from Slutsky's theorem that it will converge in probability to 0 if the numerator converges to 0. Again, before saying anything about whether the IV estimator is likely larger or smaller than the OLS estimate, we'd need some story of what $E[u_i|x_i, z_i]$ looks like.

In the case where the instrument is not valid, such that $E[u_i|z_i] \neq 0$, not even Slutsky's theorem can save us. It does help a bit; it's probably much easier to guess what $\lim_{n \rightarrow \infty} \frac{\sum u_i(z_i - \bar{z})}{\sum x_i(z_i - \bar{z})} = \frac{\text{Cov}[u_i, z_i]}{\text{Cov}[x_i, z_i]}$ is than to guess what $E\left[\frac{\sum u_i(z_i - \bar{z})}{\sum x_i(z_i - \bar{z})}\right]$ is. So we can at least know whether the IV estimate is asymptotically larger or smaller than the OLS estimate, provided we also have a good guess of what $\frac{\text{Cov}[u_i, x_i]}{\text{Var}[x_i]}$ looks like. Knowing the relative magnitudes to any reasonable degree of approximation might still be difficult, but we might still be able say with some certainty whether $\frac{\text{Cov}[u_i, z_i]}{\text{Cov}[x_i, z_i]}$ and $\frac{\text{Cov}[u_i, x_i]}{\text{Var}[x_i]}$ are of different signs.

Question 2

(a)

Ashenfelter and Krueger (1994) use the fixed-effects estimator to try and wash away the endogeneity in their estimates of the return to schooling. This is essentially estimating a separate intercept for every set of twins, or equivalently, regressing differences in outcomes on differences in regressors between twins. This allows to control for genetic differences and family characteristics that may be correlated with earnings. It might be attractive to assume that whatever differences in schooling that remain are essentially due to chance and not correlated with unobserved factors in earnings, but there are reasons to be cautious. We would essentially be saying that differences in ability, among other things, can more or less be ruled out once we hold genes and family background constant. But many things can happen, accidental or otherwise, that affect one twin's innate ability, and therefore schooling and earnings, leading to a spurious correlation even after differencing. For one, Sanderwall, Cesarini, and Johannesson (2014) show in one dataset that the twins-based estimates of returns to schooling are attenuated when IQ or birth weight are added as a proxy for ability.

Apart from the fixed-effects estimator, the authors also attempt to reduce measurement error using twins' reports of their own years of schooling as well as that of the other twin. Averaging the reports reduces the asymptotic bias, while an instrument variable approach could lead to consistent estimates (in this case, using differences in schooling, calculated from each twins' reports of the other twin's years of schooling, as an instrument).

(b)

The OLS estimate is 0.084. The first-difference estimate is 0.092 (the difference is smaller than the standard error on either the OLS or first-difference estimates), while combining the instrument variable and first-difference approaches as described before yields an estimate of 0.167 (though the 95% confidence interval contains the OLS point estimate).

(c)

One of their instruments is $\Delta S'' = S_1^2 - S_2^1$ which instruments for $\Delta S' = S_1^1 - S_2^2$, where S_i^j is twin j 's report of twin i 's years of schooling. If $\Delta S''$ is to be an instrument for $\Delta S'$, we must have $E[\eta|\Delta S''] = 0$ in the equation

$$\Delta y = \beta \Delta S + \gamma \Delta Z + \varepsilon = \beta \Delta S' + \gamma \Delta Z + \underbrace{\varepsilon - \beta(u_1^1 - u_2^2)}_{\text{Composite error } \eta}$$

If the errors in reported schooling are uncorrelated across twins, then

$$\begin{aligned}\Delta S' &= \Delta S + u_1^1 - u_2^2 \\ \Delta S'' &= \Delta S + u_1^2 - u_2^1\end{aligned}$$

and we can see that $E[\eta|\Delta S''] = 0$ if reporting errors are uncorrelated across twins. This seems like a good baseline to work with, but the authors do consider the possibility of correlation in

reporting errors across twins; in this case they instrument $\Delta S^* = S_1^1 - S_2^1$ with $\Delta S^{**} = S_1^2 - S_2^2$. This is a consistent estimator for ΔS which can be shown in steps similar to the above.

(d)

For one, Angrist and Krueger (1991) use birth quarters as an instrument, exploiting the timing of compulsory school attendance requirements, and find that the IV estimates are close to OLS estimates. One problem with IV estimates is that they only identify the local average treatment effect. In this case, the estimates could be attenuated downwards since they only identify the effect of essentially forcing someone, who would otherwise have dropped out, to study for one more year. (There is also a potential issue in this particular paper; Buckles and Hungerman (2013) find that season of birth is associated with maternal characteristics that may affect later outcomes.)

Fixed-effects estimates do not have this problem, but they do face the problem of potentially throwing away too much variation. One possible pitfall is that de-meaning or differencing the data could mostly leave measurement error behind, leading to inflated coefficients. Ashenfelter and Krueger do attempt to account for this by combining the FE and IV approaches.

Question 3

(a)

The first design estimates

$$E \left[y_{old,1}^{NJ} - y_{old,0}^{NJ} \right] - E \left[y_{young,1}^{NJ} - y_{young,0}^{NJ} \right]$$

and someone putting faith in this design would inherently be assuming that secular trends affecting healthcare have the same effect on average whether we look at the young or the old in New Jersey. In other words, they believe that

$$\begin{aligned} y_{old,t}^{NJ} &= \alpha_{old}^{NJ} + \mu_t^{NJ} + \delta \times \mathbb{1}(t = 1) + \varepsilon_{old,t}^{NJ} \\ y_{young,t}^{NJ} &= \alpha_{young}^{NJ} + \mu_t^{NJ} + \varepsilon_{young,t}^{NJ} \end{aligned}$$

The second design estimates

$$E \left[y_{old,1}^{NJ} - y_{old,0}^{NJ} \right] - E \left[y_{old,1}^{PA} - y_{old,0}^{PA} \right]$$

and someone putting faith in this design would inherently be assuming that secular trends affecting healthcare have the same effect on average whether we look at the old in New Jersey or the old in Pennsylvania. In other words, they believe that

$$\begin{aligned} y_{old,t}^{NJ} &= \alpha_{old}^{NJ} + \theta_{old,t} + \delta \times \mathbb{1}(t = 1) + \varepsilon_{old,t}^{NJ} \\ y_{old,t}^{PA} &= \alpha_{old}^{PA} + \theta_{old,t} + \varepsilon_{old,t}^{PA} \end{aligned}$$

We should believe in the first design when we think there is something affecting health equally for the old and young in New Jersey over time. For example, the state government could be cutting

its healthcare budget (though there may be reasons to believe this affects certain age groups more than others).

We should believe in the second design when we think there is something affecting health equally for the old in New Jersey and Pennsylvania over time. For example, the federal government could be cutting its healthcare budget (though there may be reasons to believe this affects certain states more than others).

(b)

We have

$$\begin{aligned}\hat{\delta}_{DID1} &= \left(\hat{y}_{old,1}^{NJ} - \hat{y}_{old,0}^{NJ} \right) - \left(\hat{y}_{young,1}^{NJ} - \hat{y}_{young,0}^{NJ} \right) = (3.2 - 3.5) - (3.9 - 4) = -0.2 \\ \hat{\delta}_{DID2} &= \left(\hat{y}_{old,1}^{NJ} - \hat{y}_{old,0}^{NJ} \right) - \left(\hat{y}_{old,1}^{PA} - \hat{y}_{old,0}^{PA} \right) = (3.2 - 3.5) - (3.3 - 3.4) = -0.2\end{aligned}$$

and it seems it doesn't matter which design we think is more reasonable (though if we were given more than the averages it might matter since the standard errors might differ).

(c)

We might now believe that

$$\begin{aligned}y_{old,t}^{NJ} &= \alpha_{old}^{NJ} + \mu_t^{NJ} + \theta_{old,t} + \delta \times \mathbb{1}(t = 1) + \varepsilon_{old,t}^{NJ} \\ y_{young,t}^{NJ} &= \alpha_{young}^{NJ} + \mu_t^{NJ} + \theta_{young,t} + \varepsilon_{young,t}^{NJ} \\ y_{old,t}^{PA} &= \alpha_{old}^{PA} + \mu_t^{PA} + \theta_{old,t} + \varepsilon_{old,t}^{PA} \\ y_{young,t}^{PA} &= \alpha_{young}^{PA} + \mu_t^{PA} + \theta_{young,t} + \varepsilon_{young,t}^{PA}\end{aligned}$$

We should believe in this when we think there is something affecting states equally (regardless of age group) and something affecting age groups equally (regardless of state). For example, the state governments of New Jersey and Pennsylvania could both be adjusting its healthcare budget in different ways (though there may be reasons to believe these may affect certain age groups more than others), and the federal government could be spending more on the elderly (though there may be reasons to believe this may affect certain states more than others).

We can implement this in the regression specification

$$\begin{aligned}y_{it} &= \alpha + \alpha^t \times \mathbb{1}(t = 1) + \alpha^s \times \mathbb{1}(s = NJ) + \alpha^a \times \mathbb{1}(a = old) \\ &+ \beta^{ts} [\mathbb{1}(t = 1) \times \mathbb{1}(s = NJ)] + \beta^{ta} [\mathbb{1}(t = 1) \times \mathbb{1}(a = old)] + \beta^{sa} [\mathbb{1}(s = NJ) \times \mathbb{1}(a = old)] \\ &+ \delta [\mathbb{1}(t = 1) + \mathbb{1}(s = NJ) + \mathbb{1}(a = old)]\end{aligned}$$

which is a saturated model. But we didn't really have to write that out since we can just calculate it as such

$$\begin{aligned}\hat{\delta}_{DIDID} &= \left[\left(\hat{y}_{old,1}^{NJ} - \hat{y}_{old,0}^{NJ} \right) - \left(\hat{y}_{young,1}^{NJ} - \hat{y}_{young,0}^{NJ} \right) \right] - \left[\left(\hat{y}_{old,1}^{PA} - \hat{y}_{old,0}^{PA} \right) - \left(\hat{y}_{young,1}^{PA} - \hat{y}_{young,0}^{PA} \right) \right] \\ &= [(3.2 - 3.5) - (3.9 - 4)] - [(3.3 - 3.4) - (3.8 - 4.1)] = -0.4\end{aligned}$$

It is not a given that this is a “more robust” analysis. We are now imposing an additional parallel trend assumption, instead of just one if we had gone along with one of the initial designs. Moreover we have to assume that these effects are additive which makes adding these interaction effects a bit more of an onerous restriction. On the other hand, if we assume there are no additional age-specific trends ($\theta_{a,t} = 0$), then the DDD estimator becomes something like a placebo test, since we know the policy must have no effect on the younger group in New Jersey. The drawback is that even if the placebo assumptions are valid, the estimator will have a larger standard error and we will have to work especially hard if we want to convince people of a null result in particular.

Question 4

(a)

The difference-in-differences approach is as described before, and the assumption that makes it work (at least in a 2×2 setting) is one about parallel trends, or in this case common shocks. The details are not much different from what was discussed before, so we calculate the estimate:

$$\hat{\delta}_{DID} = (y_L^{TX} - y_L^{NY}) - (y_H^{TX} - y_H^{NY}) = (9,636 - 7,843) - (9,981 - 9,676) = 1,488$$

Intuitively, the first term reflects both the benefit of getting a GED (since low-scoring Texans get one while low-scoring New Yorkers don't) and the inter-state difference in earnings. The second term only reflects the inter-state difference in earnings (the difference is not assumed to vary by qualification), so subtracting this from the first leaves the benefit of getting a GED (assumed not to differ by state, and assuming that a low-scoring Texans and New Yorkers are balanced on unobserved covariates that affect earnings).

(b)

Those who are just above the cut-off may simply be more hardworking or ambitious (or richer, with a lower opportunity cost of time), given that they might have had similar scores to those who were below and just took the test more times. The low-scoring Texans (defined to be scoring too low for New York but not too low for Texans) might just be more hardworking or whatnot than the low-scoring New Yorkers; the assumption that low-scoring Texans and New Yorkers are balanced on unobserved covariates that affect earnings is violated.

(c)

The regression discontinuity design exploits the fact that someone scoring, say, one mark below the cutoff is likely to be not very different from someone who scores exactly at the cutoff. Any substantial differences in outcomes between these two classes of people is therefore plausibly due to just receiving the GED. In this case the running variable is GED scores, and the we are looking at a “sharp” regression discontinuity design (unlike a “fuzzy” design, the probability of treatment goes from 0 to 1 at the cutoff). The key identifying assumption is the continuity assumption, which assumes that $E[y_i^1 | x_i = x]$ and $E[y_i^0 | x_i = x]$ are continuous in x ; we should not

see any discontinuous jump in the potential outcomes as x , in this case the GED score, varies. This assumes that all the unobserved covariates are continuously related to GED scores; if we observe a discontinuity in outcomes, it must have been due to the treatment.

One way to do this is to estimate the following model

$$y_i = \alpha + \sum_{k=1}^p \beta_k x_i^k + \delta \mathbb{1}(x_i \geq \tilde{x}) + \varepsilon_i$$

though Gelman and Imbens (2016) advise against anything of a higher-order than a quadratic form as this approach introduces bias. We are really letting the polynomial in x_i approximate the continuous function $E[y_i^1 | x_i = x]$; conditional on this, any discontinuity in y_i must be causal.

The robustness checks we must perform are to ensure that the identifying assumptions are valid. In particular, we would want to make sure that the only source of discontinuity is the treatment assignment mechanism and not any other covariate. Using final attained GED scores violates this since we expect some bunching from people who retake the exams. Plotting the observed variables would be helpful to check for this. Another good idea is to do placebo tests based on ‘fake’ cutoff points. If we detect an effect at points we arbitrarily choose, the design is perhaps flawed. Another test in a similar vein is to use the real cutoff point, but check for an effect on different outcome variables which we know should not be affected by the treatment, say, earnings before taking the GED.