ECMA 31100: Intro to Empirical Analysis II

Anderson Rubin Test; Many Instruments

Joe Hardwick

University of Chicago

Winter 2022

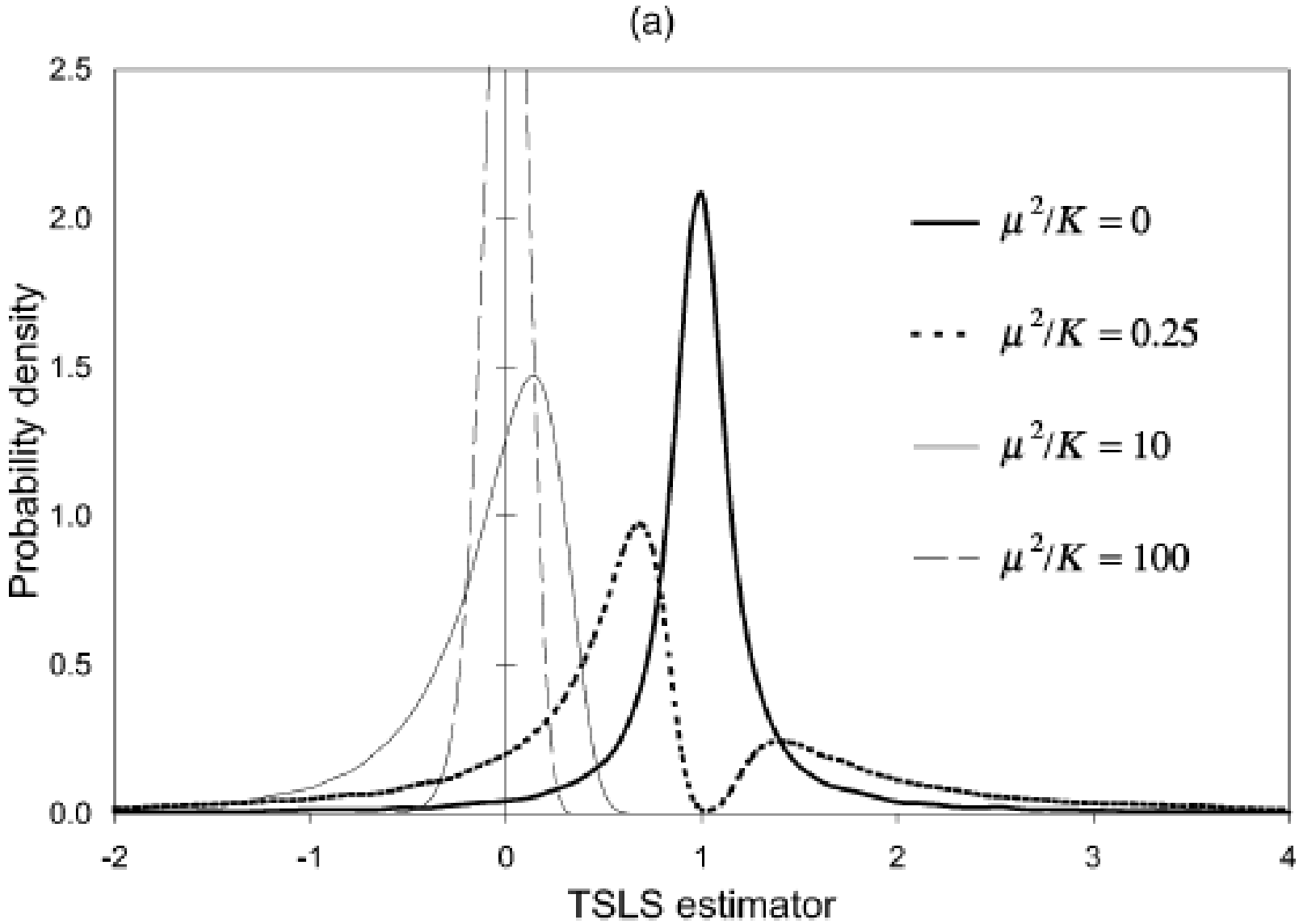
Stock,Wright and Yogo (2002)

- Stock, Wright and Yogo (2002) simulate the finite sample distribution of $\hat{\beta}_{IV}$:

$$y = \beta x + u;$$
$$x = \pi z + v;$$
$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}\right),$$

and $\beta = 0$. Instruments non-random. In this case $\hat{\beta}_{OLS} \xrightarrow{p} \frac{\sigma_{uv}}{\sigma_v^2} = 0.99$. $\mu = \pi \left(\sum_{i=1}^{n} z_i^2\right)^{1/2}$. $K = \dim(z) = 1$.

Stock, Wright and Yogo (2002)



Conducting inference on β

- Two-step methods that check the first stage F-statistic and then use $\hat{\beta}_{IV}$ to conduct inference on β are often unreliable, and it's generally difficult to come up with a reasonable value for the F-statistic.
- Rule of thumb F-stat≥ 10 reasonable for absolute relative bias of 10% and for the specific case of three or more instruments (where the mean actually exists, rather than being approximated by a nagar bias). Stock and Yogo (2005) show the critical value lies between 9 and 12 for number of instruments between 3 and 30.
- In joint normal errors/fixed instruments case, $\hat{\beta}_{IV}$ has number of moments equal to the number of excluded instruments - number of endogenous variables (Kinal (1980)).

Conducting inference on β

- Other methods of defining weak instruments exist, for example actual size of t-test should be 'close' to specified significance level.
- With several endogenous variables, the bias of the 2SLS estimate becomes the norm of the vector of biases for the coefficients on each of the endogenous variables.
- Under heteroskedasticity, Montiel Olea and Pflueger (2013) suggest a modification to the F-stat, but it's no longer the case that $\hat{\beta}_{IV}$ is centered at the prob. limit of OLS. See Pflueger and Wang (2015) for a Stata implementation.

Conducting inference on β

- We use a test that is robust to weak instruments, called Anderson-Rubin test. Suppose $y = x'\beta + u$.
- Idea: $\mathrm{E}(zu) = 0$ implies $\mathrm{E}(z(y - x'\beta)) = 0$, so under $H_0 : \beta = \beta_0$,

$$\mathrm{E}(zu(\beta_0)) = 0,$$

where $u(\beta_0) = y - x'\beta_0$. We can write this equivalently as

$$\mathrm{E}(zz')^{-1}\mathrm{E}(zu(\beta_0)) = 0,$$

and this is the vector of coefficients of a regression of $u(\beta_0)$ on z.

Conducting inference on β

- Reject if coefficient estimates are significantly different from zero.
- Substitute $y = x'\beta + u$ to give

$$\begin{aligned} \mathrm{E}\left(zz'\right)^{-1}\mathrm{E}\left(zu\left(\beta_0\right)\right) &= \mathrm{E}\left(zz'\right)^{-1}\mathrm{E}\left(zx'\left(\beta-\beta_0\right)+zu\right) \\ &= \mathrm{E}\left(zz'\right)^{-1}\mathrm{E}\left(zx'\right)\left(\beta-\beta_0\right). \end{aligned}$$

  which equals zero iff $\beta = \beta_0$.
- Therefore, should expect such a test to have power against any alternative $\beta \neq \beta_0$.

Conducting inference on β

- Suppose $z \in \mathbb{R}^l$ and run the regression

$$y_i - x_i'\beta_0 = z'\gamma + \epsilon; \quad \mathrm{E}(z\epsilon) = 0.$$

Under H_0,

$$\gamma = \mathrm{E}(zz')^{-1}\,\mathrm{E}\left(z\left(y_i - x_i'\beta_0\right)\right) = 0.$$

Assuming all relevant moments exist:

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

and so

$$n\left(\hat{\gamma} - \gamma\right)' \hat{V}^{-1}\left(\hat{\gamma} - \gamma\right) \xrightarrow{d} \chi^2_l.$$

Conducting inference on β

- Under H_0, we have

$$T_n = n\left(\hat{\gamma}\right)' \hat{V}^{-1}\left(\hat{\gamma}\right) \xrightarrow{d} \chi^2_l,$$

  and we reject if $T_n > \chi^2_{l,1-\alpha}$.
- This asymptotic argument does not depend on whether instruments are weak, because under H_0:

$$\sqrt{n}\left(\hat{\gamma} - \gamma\right) = \left(\frac{1}{n}\sum_{i=1}^{n} z_i z_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i.$$

Conducting inference on β

- Let's separate the included and excluded instruments: $z = (z_1, z_2)$ where z_1 are included and z_2 are excluded.
- Now write $x = (x_1, z_1)$, where x_1 denote the endogenous variables. Then

$$y = x_1'\beta_1 + z_1'\beta_2 + u$$
$$x_1 = z_1'\pi_1 + z_2'\pi_2 + v.$$

- Typically only interested in conducting inference on β_1. Under $H_0 : \beta_1 = \beta_{1,0}$:

$$y - x_1'\beta_{1,0} = z_1'\beta_2 + u,$$

where $\mathrm{E}(uz_1) = 0$ and $\mathrm{E}(uz_2) = 0$

Aside: Overidentification tests

- Instead of testing whether $\beta = \beta_0$, we can check whether the validity condition $\mathrm{E}(zu) = 0$ holds when $\dim(x_1) < \dim(z_2)$.
- Suppose $\dim(z_2) = l_2 > k_1 = \dim(x_1)$. We say there are $l_2 - k_1$ "overidentifying restrictions".
- Now we don't know β, but under $H_0 : \mathrm{E}(zz')^{-1}\mathrm{E}(zu) = 0$, any GMM estimator is consistent.
- Run the regression

$$y_i - x_i'\hat{\beta}_{GMM} = z_i'\gamma + \epsilon_i.$$

Aside: Overidentification tests

- Assuming homoskedasticity and using the 2SLS estimator yields that the F statistic is asymptotically distributed as $\chi^2_{l_2-k_1}$ under H_0. This is known as Sargan's test.
- Alternatives robust to heteroskedasticity exist (see Wooldridge text p134).
- If H_0 is rejected, either our model specification is wrong, one or more of the instruments is invalid, but it does not tell us which instrument is invalid.
- If we fail to reject, it could be that both instruments are invalid but cause similar bias! Can only reliably reject H_0 when some subset of instruments is known to be valid.

Back to inference on β

- Since $\mathrm{E}(zu) = 0$, we have

$$y - x_1'\beta_{1,0} = z_1'\beta_2 + z_2'\gamma + u; \quad \mathrm{E}(zu) = 0$$

  where $\gamma = 0$. If $\gamma \neq 0$ either the specification is incorrect, z_2 is not valid or H_0 is false.
- Suppose $z_1 \in \mathbb{R}^{l_1}, z_2 \in \mathbb{R}^{l_2}$, where $l_1 + l_2 = l$.
- Run this regression and note that for some V:

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

  so

$$n(\hat{\gamma} - \gamma)' \hat{V}^{-1} (\hat{\gamma} - \gamma) \xrightarrow{d} \chi^2_{l_2}.$$

Conducting inference on β

- Recall that

$$\begin{aligned}\mathrm{E}\left(zz'\right)^{-1}\mathrm{E}\left(zu\left(\beta_0\right)\right) &= \mathrm{E}\left(zz'\right)^{-1}\mathrm{E}\left(zx'\left(\beta-\beta_0\right)+zu\right)\\ &= \mathrm{E}\left(zz'\right)^{-1}\mathrm{E}\left(zx'\right)\left(\beta-\beta_0\right).\end{aligned}$$

- If the instruments are not relevant $\mathrm{E}\left(zx'\right)$ isn't full rank, so can't distinguish between β and β_0.
- If the instruments are weak, this can be done but power may be poor. AR test controls null rejection probability but does not correct for the fundamental issue of having little information about β from the instruments.

Conducting inference on β

- Critical value $\chi^2_{l_2,1-\alpha}$ increasing in l_2, so may expect finite sample power to be poor if there are many excluded instruments.
- Other alternatives under overidentification discussed in Andrews, Stock and Sun (2019).
- Test has power against $\beta \neq \beta_0$ but also against violations of overidentifying restrictions.
- Testing not just whether $\beta = \beta_0$, but whether in fact there exists a β satisfying model assumptions.

Many Instruments

- Consider the first stage

$$x = z'\pi + v; \quad \mathrm{E}(zv) = 0.$$

- Estimate π by OLS to give fitted values

$$\hat{x}_i = z_i'\hat{\pi}.$$

  If we knew π, the fitted values would be exogenous, but $\hat{\pi}$ depends on v, which is correlated with u.

Many Instruments

- The 2SLS estimator is given by

$$\begin{aligned}\hat{\beta}_{2SLS} - \beta &= \left(X' P_Z X\right)^{-1} X' P_Z U \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i u_i\end{aligned}$$

- If $\hat{x} = z'\pi$ then $\hat{x}$ is uncorrelated with the error since $\mathrm{E}(zu) = 0$.
- Asymptotically, 2SLS is consistent because $\hat{\pi} \xrightarrow{p} \pi$, but what about when the number of instruments is proportional to the sample size?

Many Instruments

- One correction suggested is Jackknife instrumental variables estimator:

$$\hat{\beta}_{JIVE} - \beta = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{x}_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \hat{x}_i u_i,$$

  where this time $\hat{x}_i = z_i' \hat{\pi}_{-i}$. $\hat{\pi}_{-i}$ is the OLS estimator in the first stage computed after dropping observation i.
- Because observations are iid, $z_i \hat{\pi}_{-i}$ is now uncorrelated with u_i.
- Chao et. al (2012) work out the details for JIVE showing consistency and asymptotic normality under heteroskedasticity with many instruments.

Many Instruments

- For now, let's derive the inconsistency in OLS/2SLS assuming $\dim(x) = 1$ and

$$\frac{\dim(z)}{n} := \frac{K}{n} \to \alpha \in (0,1).$$

- Note that

$$\mathrm{E}(x^2) = \mathrm{E}\left(\left[z'\pi + v\right]^2\right) = \mathrm{E}\left(\left[z'\pi\right]^2\right) + \mathrm{E}(v^2).$$

We could keep the instrument strength (concentration parameter fixed) by assuming $\mathrm{E}\left(\left[z'\pi\right]^2\right) \to H > 0$. What we actually need:

$$\frac{1}{n}\sum_{i=1}^{n} \pi' z_i z_i' \pi \xrightarrow{p} H.$$

Many Instruments

- Next suppose $\mathrm{E}(v|z) = 0$, $\mathrm{E}\left(v^2|z\right) = \sigma_v^2$ and $\mathrm{E}\left(v^4|z\right) \leq B$ for some constant B. Assume similar for u, and $\mathrm{E}(uv|z) = \sigma_{uv}$. Then

$$Var\left(\frac{1}{n}\sum_{i=1}^{n}\pi' z_i v_i | Z\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left(\pi' z_i\right)^2 Var\left(\pi' z_i v_i | Z\right)$$
$$= \sigma_v^2 \cdot \frac{1}{n^2}\sum_{i=1}^{n} Var\left(z'\pi\right) \to 0.$$

Since $\mathrm{E}(\pi' z_i v_i | Z) = 0$ and convergence in L^2 implies convergence in probability:

$$\frac{1}{n}\sum_{i=1}^{n}\pi' z_i v_i \xrightarrow{p} 0; \quad \frac{1}{n}\sum_{i=1}^{n}\pi' z_i u_i \xrightarrow{p} 0$$

Many Instruments

- Moreover:

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' = \frac{1}{n}\sum_{i=1}^{n} \pi' z_i z_i' \pi + \frac{2}{n}\sum_{i=1}^{n} \pi' z_i v_i + \frac{1}{n}\sum_{i=1}^{n} v_i^2$$
$$\xrightarrow{p} H + \sigma_v^2$$

and

$$\frac{1}{n}\sum_{i=1}^{n} x_i u_i = \frac{1}{n}\sum_{i=1}^{n} \pi' z_i u_i + \frac{1}{n}\sum_{i=1}^{n} u_i v_i,$$

so

$$\hat{\beta}_{OLS} - \beta \xrightarrow{p} \frac{\sigma_{uv}}{H + \sigma_v^2}.$$

Many Instruments

- For $2SLS$:

$$\begin{aligned}\hat{\beta}_{2SLS} - \beta &= \left(X' P_Z X\right)^{-1} X' P_Z U \\ &= \left(\frac{X' P_Z X}{n}\right)^{-1} \left[\frac{\left(Z\Pi + V\right)' P_Z U}{n}\right].\end{aligned}$$

  Since $Z' P_Z = Z$, we have

$$\frac{\left(Z\Pi + V\right)' P_Z U}{n} = \frac{\Pi' Z U + V' P_Z U}{n} = \frac{1}{n}\sum_{i=1}^{n} \pi' z_i u_i + \frac{V' P_Z U}{n}.$$

- Usually the last term would converge in probability to zero, but now the dimension of Z is growing with n.

Many Instruments

- We have

$$\begin{aligned}
\mathrm{E}\left(V'P_zU\middle|Z\right) &= tr\left(\mathrm{E}\left(V'P_ZU\middle|Z\right)\right) \\
&= tr\left(P_Z\mathrm{E}\left(UV'|Z\right)\right) \\
&= tr\left(P_Z\cdot\sigma_{uv}I_n\right) \\
&= \sigma_{uv}\left(tr\left(P_Z\right)\right) \\
&= \sigma_{uv}\left(tr\left(\left[Z'Z\right]^{-1}Z'Z\right)\right) \\
&= \sigma_{uv}\cdot K
\end{aligned}$$

  Therefore

$$\mathrm{E}\left(\frac{V'P_ZU}{n}\middle|Z\right) = \sigma_{uv}\cdot\frac{K}{n}\to\sigma_{uv}\alpha.$$

Many Instruments

- Can also show that

$$Var\left(\frac{V'P_ZU}{n}\middle| Z\right) \to 0,$$

so

$$\frac{V'P_ZU}{n} \xrightarrow{p} \sigma_{uv}\alpha,$$

therefore the 2SLS estimator is inconsistent with many instruments.

Questions?

Causal parameters from estimators

- Return to the model

$$y_i = \beta_0 + \beta_1 d_i + u_i,$$

  where y_i is the observed outcome of individual i and $d_i = 1$ if individual i is treated and 0 otherwise.
- A causal interpretation of this model implies that the ceteris paribus effect of the treatment, β_1, is the same for everybody.
- On the other hand, since d_i is binary we can always write

$$y_i = b_0 + b_1 d_i + \epsilon_i; \quad \mathrm{E}(\epsilon_i | d_i) = 0.$$

- The latter model does not necessarily have a causal interpretation, but the parameter b_1 may be of interest.
- The error terms u_i and ϵ_i are interpreted differently: u_i contains unobserved determinants of y_i, ϵ_i is the projection residual. It is not necessarily true that $\mathrm{E}(u_i | d_i) = 0$.

Causal parameters from estimators

- We may write

$$y_i = y_{i1} d_i + y_{i0} (1 - d_i) = y_{i0} + d_i (y_{i1} - y_{i0}),$$

  where y_{i0}, y_{i1} are the potential outcomes for individual i.
- y_{i0} is the outcome we would observe if the individual is not treated, and y_{i1} the outcome if the individual is treated.
- The difference $y_{i1} - y_{i0}$ is the individual treatment effect.
- The average across all individuals, $\mathrm{E}(y_{i1} - y_{i0})$, is the average treatment effect (ATE).

Causal parameters from estimators

- So far we have argued that if treatment is randomly assigned then the OLS estimate gives the ATE. Adding the assumption that $y_{i1} - y_{i0} = c$ allowed us to estimate c and to interpret the regression as a causal model.
- One reason to be interested in 2SLS is that while we may ask for a parameter of interest (such as the ATE) and then look at how to estimate it, we can go the other way around: Use the estimation method anyway and then ask what parameter it identifies.
- For 2SLS/IV the answer (under some assumptions) is a Local Average Treatment Effect.

Causal parameters from estimators

- Suppose we allow for heterogeneous treatment effects, so that $y_{i1} - y_{i0} = \beta_{i1}$, and the baseline outcome $\beta_{i0} = y_{i0}$ may also differ across individuals. Then

$$y_i = \beta_{i0} + \beta_{i1} d_i.$$

- Such a model is called a random coefficients model, because the fact we are sampling individuals at random implies that the actual value of β_{i0}, β_{i1} drawn is random also.

Causal parameters from estimators

- The assumption of random assignment to treatment, $(y_{i0}, y_{i1}) \perp\!\!\!\perp d_i$, implies that in the model

$$y_i = b_0 + b_1 d_i + \epsilon_i, \quad \mathrm{E}\left(\epsilon_i | d_i\right) = 0,$$

  we have

$$\begin{aligned} b_1 &= \mathrm{E}\left(y_i | d_i = 1\right) - \mathrm{E}\left(y_i | d_i = 0\right) \\ &= \mathrm{E}\left(y_{i1} | d_i = 1\right) - \mathrm{E}\left(y_{i0} | d_i = 0\right) \\ &= \mathrm{E}\left(y_{i1} - y_{i0}\right). \end{aligned}$$

- Random assignment implies OLS estimate of b_1 is a consistent estimate of the ATE.

Causal parameters from estimators

- If, on the other hand, there is selection into treatment, we obtain merely

$$b_1 = \mathrm{E}\left(y_{i1}|d_i = 1\right) - \mathrm{E}\left(y_{i0}|d_i = 0\right).$$

- Want to identify some feature of the distribution of $y_{i1} - y_{i0}$.
- Suppose we have a binary instrument $z \in \{0, 1\}$. For example, y might be worker output, $d = 1$ if the worker elects job training and 0 otherwise, and $z = 1$ if the worker is offered job training and 0 otherwise. z may be independent of (y_0, y_1) even if d isn't.

Causal parameters from estimators

- Now we not only have potential outcomes but also potential treatments:

$$d_i = d_{i1} z_i + d_{i0} (1 - z_i),$$

  where the potential outcome $d_{i1} = 1$ if individual i would take job training when they are offered it and 0 otherwise. $d_{i0} = 1$ if individual i would take job training when they are not offered it, and 0 otherwise.
- Question: "OLS produces $b_1 = \mathrm{E}(y_{i1}|d_i = 1) - \mathrm{E}(y_{i0}|d_i = 0)$, which is not a feature of $y_{i1} - y_{i0}$, but what if I do 2SLS/IV with my instrument?"

Causal parameters from estimators

- First we will make some relevance and validity assumptions analogous to those in the linear model.
- Validity: $(y_0, y_1, d_0, d_1) \perp\!\!\!\perp z$.
- Relevance: Two assumptions
  - Monotonicity assumption: $\mathrm{P}(d_1 \geq d_0) = 1$. This implies

$$\begin{aligned} Cov(d, z) &= \mathrm{E}(dz) - \mathrm{E}(d)\mathrm{E}(z) \\ &= \mathrm{P}(d = 1, z = 1) - \mathrm{P}(d = 1)\mathrm{P}(z = 1) \\ &= [\mathrm{P}(d = 1|z = 1) - \mathrm{P}(d = 1|z = 0)]\mathrm{P}(z = 1)\mathrm{P}(z = 0) \\ &= [\mathrm{P}(d_1 = 1) - \mathrm{P}(d_0 = 1)]\, Var(z) \\ &= \mathrm{P}(d_1 > d_0)\, Var(z) \geq 0. \end{aligned}$$

  - $\mathrm{P}(d_0 \neq d_1) > 0$. This says that the instrument must alter treatment choice for some positive fraction of individuals. Together with monotonicity, this implies $\mathrm{P}(d_1 > d_0) > 0$.

Causal parameters from estimators

- Note that the first stage regression produces fitted values

$$\hat{d}_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i,$$

where

$$\hat{\gamma}_1 = \frac{\hat{Cov}(d, z)}{\hat{Var}(z)} \xrightarrow{p} \mathrm{P}(d_1 > d_0) > 0.$$

- The second stage regression runs OLS on

$$y_i = \alpha + \beta \hat{d}_i + v_i.$$

- This gives:

$$\hat{\beta}_{2SLS} = \frac{\hat{Cov}(y, z) / \hat{Var}(z)}{\hat{Cov}(d, z) / \hat{Var}(z)} \xrightarrow{p} \frac{Cov(y, z)}{Cov(d, z)}.$$

Causal parameters from estimators

- A similar argument shows that

$$
\begin{aligned}
Cov\left(y,z\right) &= \left[\mathrm{E}\left(y|z=1\right)-\mathrm{E}\left(y|z=0\right)\right] Var\left(z\right) \\
&= \mathrm{E}\left(y_1 d_1 + y_0\left(1-d_1\right)|z=1\right) Var\left(z\right) \\
&- \mathrm{E}\left(y_1 d_0 + y_0\left(1-d_0\right)|z=0\right) Var\left(z\right) \\
&= \mathrm{E}\left(y_1 d_1 + y_0\left(1-d_1\right)\right) Var\left(z\right) \\
&- \mathrm{E}\left(y_1 d_0 + y_0\left(1-d_0\right)\right) Var\left(z\right) \\
&= \mathrm{E}\left(\left[y_1-y_0\right]\left[d_1-d_0\right]\right) Var\left(z\right) \\
&= \mathrm{E}\left(y_1-y_0|d_1>d_0\right)\mathrm{P}\left(d_1>d_0\right) Var\left(z\right).
\end{aligned}
$$

Causal parameters from estimators

- It follows that

$$\hat{\beta}_{2SLS} \xrightarrow{p} \frac{Cov(y,z)}{Cov(d,z)} = \mathrm{E}(y_1 - y_0 | d_1 > d_0).$$

- This parameter is the average treatment effect among those who would be switched from no treatment to treatment if the value of the instrument changed.
- In the job training example, it is the average effect among those who would elect training if offered it and not elect it otherwise.
- This parameter may or may not be of interest, and this interpretation depends on the monotonicity assumption.
- The interpretation changes if the instrument is changed (unless, of course, $y_1 - y_0$ is constant).

Questions?