EEP/IAS 118 - Introductory Applied Econometrics, Section 13

Pierre Biscaye and Jed Silver

December 2021

Course Evaluations

Please fill out your course evaluations for both lecture and section!

`https://course-evaluations.berkeley.edu`

Agenda

- Review: Panel Data
- Instrumental Variables (IV)

Panel Data and OVB

How can we use panel data to help reduce OVB?

1. Take first differences (or other differences) between observations

$$\Delta y_{it} = \Delta\delta_t + \beta_1 \Delta x_{1it} + \cdots + \beta_k \Delta x_{kit} + \Delta u_{it}$$

2. Use unit fixed effects (equivalent to differencing out the mean)

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \alpha_i + \delta_t + u_{it}$$

Both control for time-constant characteristics of the units of analysis ⇒ reduces the number of potential omitted variables, without having to know anything about what those omitted variables could be.

- Still will be concerned about time-varying omitted variables

First Differences vs Fixed Effects

- Both deal with time-constant characteristics of the unit of analysis and give unbiased estimates if all assumptions hold.
- Both require the assumption of strict exogenity (modified MLR4): $cov(x_{jit}, u_{is}) = 0$ for all t, s, j, not just for $s = t$.
- Identical with 2 time periods.
- Both have correlation of the error terms over time within units, which will lead to different standard errors. Nature of correlation determines which approach will have smaller SEs.
- Fixed effects easier to estimate with R functions (i.e., felm).

Other Panel Data Considerations

- Correlated errors within units over time: we learn less from each observation
  - Solution: **cluster** standard errors to capture level at which errors are correlated
  - in R: felm(regression formula | fixed effects | instruments | *clusters*, data) or use vcovCL
- Can include **lagged variables**; tempting to include lagged dependent variables $\Rightarrow$ control for unobserved factors that affected outcome in prior period and also in current period
  - Don't use together with first differences/fixed effects: control for same thing twice, violate strict exogeneity
- **Unbalanced panels**: no data for some units in certain time periods
  - First differences will lose more data than fixed effects
  - Could have bias if factors determining whether a unit is observed are changing with x

Example

You are interested in the effect of having a child in daycare on work hours, and have data on both parents i from 100 households h in the same zip code over 12 months t.

$$hours_{iht} = \beta_0 + \beta_1 daycare_{ht} + \beta_2 sector_{iht} + \alpha_{ih} + month_t + u_{iht}$$

1. What does the individual fixed effect control for?
2. Where does our identifying variation come from for estimating β_1? In other words, what households are providing the information we use to estimate this coefficient in a unit fixed effects model?
3. What level of SE clustering would be appropriate in this model - which variable would you cluster by?
4. Does this regression recover the causal impact of daycare on work hours? If not, what could be a possible source of omitted variable bias?

Example

$$hours_{iht} = \beta_0 + \beta_1 daycare_{ht} + \beta_2 sector_{iht} + \alpha_{ih} + month_t + u_{iht}$$

1. What does the individual fixed effect control for?
   All characteristics of the individual (and their household) that do not change over time. It also controls for means within individual across the sample period, such as their mean age, mean years of education, mean number of children, etc.

2. Where does our identifying variation come from for estimating β_1?
   With unit fixed effects the coefficient for a particular variable is estimated using within-individual variation across time around its mean. The households that provide the information we use to estimate β_1 is the subset of households that vary their use of daycare in the survey period (i.e. they use day care in some months but not others). For all others there is no variation around the mean.

Example

$$hours_{iht} = \beta_0 + \beta_1 daycare_{ht} + \beta_2 sector_{iht} + \alpha_{ih} + month_t + u_{iht}$$

3. What level of SE clustering would be appropriate in this model - which variable would you cluster by?
   Errors will almost definitely be correlated within individuals, so we could cluster our standard errors at the individual level. But unobserved factors determining daycare use and work hours are likely to be similar for both adults in a household as well. Conservatively then, we could cluster our standard errors at the household level.
4. Does this regression recover the causal impact of daycare on work hours? If not, what could be a possible source of OVB?
   Probably not. There could stll be variables that are changing over time that are correlated with daycare use. For example, a household with a child in daycare could have a new baby and one parent might stop working and remove their child from daycare to care for both children at home.

Instrumental Variables: Intro

The panel data methods are tools that help use deal with omitted variable bias (or endogeneity). However, we often don't have the luxury of panel data (or an RCT, or a treatment eligibility threshold), or we suspect that changes in omitted variables are correlated with changes in x variables.

- Instrumental Variables (IVs) are another method we can try to deal with OVB
- Intuitively, we recognize that our variable of interest X is not as good as random (correlated with u).
- We find another variable that *is* as good as random, Z (the IV), which impacts Y only through X. We use Z to learn about the effect of X on Y using only the variation in X that is as good as random (explained by Z).

Instrumental Variables: Properties

An instrumental variable (Z) must satisfy these two conditions:

1. **Relevance:** Z must be related to our endogenous variable of interest X

   $\Rightarrow \; cov(z, x) \neq 0$

2. **Exogeneity / Exclusion:** Z should be unrelated to Y except indirectly via its effect on X. In other words, Z should be uncorrelated with all omitted variables.

   $\Rightarrow \; cov(z, u) = 0$

These are surprisingly difficult conditions to satisfy at the same time

Testing IV Properties

Can we be sure these properties are satisfied?

1. **Relevance:** We can test simply by running a regression of X on Z:

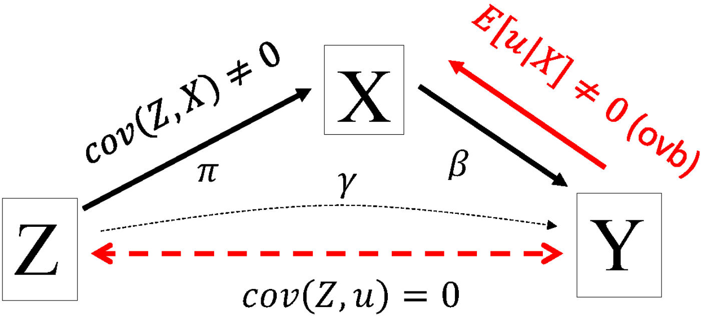
   $$x = \pi_0 + \pi_1 z + v$$

   Here, we want to reject that π_1 is zero - i.e. we want to be confident that there is a correlation between X and Z. The more strongly we reject the null, the stronger our instrument.

2. **Exogeneity / Exclusion:** Because u is unobserved, we cannot (generally) test this condition. Instead we rely on logic and intuition (otherwise known as story telling).

Instrumental Variables: Graphical Interpretation

The graph here may provide some visual intution.



The exclusion restriction implies that the only path from Z to Y is through X, meaning that $\pi\beta = \gamma$. Since Z is exogenous, we can estimate π and γ, which allows us to solve for β (as long as $\pi \neq 0$).

Instrumental Variables: Good or Bad?

The hardest thing about using IVs, is deciding whether a candidate Z satisfies the two conditions.

Example: We want to estimate the effect of school attendance on test scores. But we know that attending school is correlated with many other omitted variables (e.g. family income).

- We think we can use a family's distance to the school as an instrument for attendance. Do you think that this is a good instrument?

Instrumental Variables: Good or Bad?

The hardest thing about using IVs, is deciding whether a candidate Z satisfies the two conditions.

Example: We want to estimate the effect of school attendance on test scores. But we know that attending school is correlated with many other omitted variables (e.g. family income).

- We think we can use a family's distance to the school as an instrument for attendance. Do you think that this is a good instrument?

Maybe, but probably not. Distance likely satisfies the relevance condition - children living close to school attend more. But, distance to the school is likely correlated with omitted variables that matter for attendance - e.g., income, environmental quality, etc.

Instrumental Variables: Good or Bad?

Example: We want to estimate the effect of school attendance on test scores. But we know that attending school is correlated with many other omitted variables (e.g. family income).

- Another possibility is a merit scholarship that some schools offered to provide free college tuition if a student attended more than 95% of school days. Do you think that this is a good instrument?

Instrumental Variables: Good or Bad?

Example: We want to estimate the effect of school attendance on test scores. But we know that attending school is correlated with many other omitted variables (e.g. family income).

- Another possibility is a merit scholarship that some schools offered to provide free college tuition if a student attended more than 95% of school days. Do you think that this is a good instrument?

Maybe. This likely satisfies the relevance conditions - more students will attend due to the offer of the scholarship. But, whether this satisfies the exclusion restriction depends on which schools choose to adopt this policy. If only schools in certain areas (e.g., low-income) offer this program, then this will fail. However, if the program was randomly assigned, then this instrument will satisfy the exlusion restriction as well.

Instrumental Variables: Estimation Mechanics

The equation we wanted to estimate is

$$y = \beta_0 + \beta_1 x + u$$

But x is correlated with omitted variables, therefore we know that $\hat{\beta}_1 \neq \beta_1$ from this regression. So, we use an instrument z. Transforming this equation, we can write:

$$cov(z, y) = \beta_1 cov(z, x) + cov(z, u)$$

If our assumptions are satisfied, we know $cov(z, x) \neq 0$ and $cov(z, u) = 0$, which allows us to solve for β_1:

$$\beta_1 = \frac{cov(z, y)}{cov(z, x)}$$

Instrumental Variables: Estimation Mechanics

We have that

$$\beta_1 = \frac{cov(z,y)}{cov(z,x)}$$

We can estimate this using our sample data:

$$\widehat{\beta_1^{IV}} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

which is our **instrumental variables estimator** of β_1

Example

We want to estimate the return to education for married women in the simple regression model

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Worrying about omitted variable bias, we use a plausible instrument: $fatheduc$. Is it likely to satisfy the exclusion restriction?

Check relevance: does it satisfy the relevance restriction?

$$\widehat{educ} = 10.24 + 0.269 fatheduc$$

$$(0.28)\ (0.029)$$

Using $fatheduc$ as an IV for $educ$ gives:

$$\widehat{log(wage)} = 0.441 + 0.059 educ$$

$$(0.446)\ (0.035)$$

Instrumental Variables: Estimation Mechanics

With the previous assumptions, as well as $E[u^2|z] = \sigma^2 = \mathrm{Var}(u)$, the variance of $\hat{\beta}_1^{IV}$ is

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$$

where

1. σ_x^2 is the population variance of x
2. σ^2 is the population variance of u
3. $\rho_{x,z}^2$ is the square of the population correlation between x and z

With a random sample, you can estimate all these components

The asymptotic standard error of $\widehat{\beta_1^{IV}}$ is the square root of the estimated asymptotic variance, which itself is

$$\frac{\hat{\sigma}^2}{SST_x \cdot R_{x,z}^2}$$

Instrumental Variables: Estimation Mechanics

Compare the variance of the IV estimator to that of the OLS estimator (in the SLR case):

$$\frac{\hat{\sigma}^2}{SST_x \cdot R^2_{x,z}} vs. \frac{\hat{\sigma}^2}{SST_x}$$

Note that, because $R^2 < 1$, the IV variance is always larger than that of OLS

- If $R^2_{x,z}$ is small, then the IV variance can be much larger than that of OLS
- I.e., if the relationship between Z and X is weak, we will have lots of uncertainty about our estimate of the effect of X and Y.

Instrumental Variables: Weak Instruments

One can write the probability limit of the IV estimator as

$$E[\widehat{\beta_1^{IV}}] = \frac{cov(z,y)}{cov(z,x)} = \beta_1 + \frac{cov(z,u)}{cov(z,x)} = \beta_1 + \frac{\text{Corr}(z,u)}{\text{Corr}(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

where σ_u and σ_x are the standard deviations of u and x in the population, respectively. Note that, even if Corr(z,u) is small, a lot of bias can arise if Corr(z,x) is small too (weak instrument). **Both restrictions matter** for IV to generate unbiased estimates.

Treatment Effects with IV

Suppose we are concerned about an endogenous x but have a good instrument z

$$y = \beta_0 + \beta_1 x + u$$
$$x = \pi_0 + \pi_1 z + v$$

We can write

$$\begin{aligned} y &= \beta_0 + \beta_1(\pi_0 + \pi_1 z + v) + u \\ &= (\beta_0 + \beta_1\pi_0) + \beta_1\pi_1 z + (\beta_1 v + u) \end{aligned}$$

- Reduced form regression $y = b_0 + b_1 z + e$ will recover an unbiased estimator: $E[\hat{b}_1] = \beta_1\pi_1$.
- The effect of x on y weighted by the effect of z on y.
- Similar to an **ITT**, where π_1 is a measure of 'compliance'.
- Regressing $x = \pi_0 + \pi_1 z + v$ gives us $\hat{\pi}_1$.
- Can then recover $\widehat{\beta_1^{IV}} = \frac{\hat{b}_1}{\hat{\pi}_1}$: similar to a **TOT**.
- Assume that all of effect of z on y is through effect on x.

Instrumental Variables and Multiple Regression

Simple case: only one explanatory variable is correlated with the error

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + u$$

- This is a **structural equation** which lays out causal relationships we are interested in.
- Assume $E[u] = 0$, and z_1 is exogenous (not correlated with u).
- We suspect x of being correlated with u
- If we measure the above formula with OLS, all estimators will be biased and inconsistent $\Rightarrow$ motivation to find IV for x

Instrumental Variables and Multiple Regression

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + u$$

Can we use z_1 as an IV for x? No (why not?) we need another variable, z_2 that is 1) uncorrelated with u, and 2) correlated with x.

How are our IV equations modified? Write out endogenous explanatory variable as linear function

$$x = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v$$

where (by construction), $E[v] = 0$, $\text{Cov}(z_1, v) = 0$, $\text{Cov}(z_2, v) = 0$ and π_j are unknown

Key identification condition (with those noted above): $\pi_2 \neq 0$

Instrumental Variables and Multiple Regression

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + u$$

$$x = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v$$

Key identification condition (with those noted above): $\pi_2 \neq 0$

This can be tested via OLS through the above **reduced form equation** (writing endogenous variable in terms of exogenous variables).

Unfortunately, still cannot test that z_1 and z_2 are uncorrelated with u

This framework can also be extended to $k - 1$ exogeneous variables.

Two Stage Least Squares (2SLS)

What if we have multiple candidate IVs for the same endogenous variable? Use **2SLS**! Our structural model, as before:

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + u$$

but we have two good IV candidates: z_2 and z_3. So you have

$$x = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v$$

The best IV for x will be

$$\hat{x} = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

For IV to not be perfectly correlated with z_1, need at least $\pi_2 \neq 0$ or $\pi_3 \neq 0 \Rightarrow$ test via F-statistic

Two Stage Least Squares (2SLS)

Procedure

- Regress x on z_1, z_2, and z_3 and obtain the fitted values $\hat{x}$
- Verify that z_2 and z_3 are jointly significant in the reduced form
- Then use $\hat{x}$ as replacement for x, for the regression of y on $\hat{x}$ and z_1

Note: use a command in R (felm can handle 2SLS) to do this, since standard errors and test statistics obtained in this way (regresing $y1$ on $\hat{x}$ and z_1) are not valid without some corrections.

Other notes for IVs

1. Loss in precision can be serious
   - Recall variance of $\hat{\beta} = \sigma^2/[\widehat{SST_x}(1-\hat{R}_x^2)]$
     - With IV estimation, $\hat{R}_x^2$ is the R^2 from a regression of $\hat{x}$ on all other exogenous variable appearing in the structural equation.
     - If the correlation between $\hat{x}$ and the exogenous variables is much higher than the correlation between x and these variables $\Rightarrow$ much larger variance
2. Still worried about weak instruments
   - Rule of thumb: $F > 10$ (only on proposed IVs)
3. Can extend 2SLS to models with more than one endogenous explanatory variable. Note that you will need at least as many instruments as there are endogenous variables.
4. Useful to deal with measurement error (15.4).
5. Can extend to pooled cross sections and panel data.
6. RCTs with imperfect compliance are ideal settings for IVs.