EC402 - Problem Set 5

Konrad Burchardi

25th of February 2009

Introduction

Today we will talk about

- the **intuition** behind IV estimation,
- **common pitfalls** in IV estimation (Question 1) and
- what we mean with **'weak instruments'** (Question 2).

Suppose you want to estimate the return of schooling (s_i) on earnings (y_i) controlling for a vector of covariates x_i' in

$$y_i = x_i'\gamma + \rho s_i + \eta_i \tag{1}$$

You might be worried that **A3Rsru** is not satisfied in this regression[1], so the simple OLS estimates would not be consistent (and not unbiased anyways).

[1] Why?

You learned that if there is an **instrumental variable** z_i which satisfies

- $Cov(z_i, s_i) \neq 0$
- $Cov(z_i, \eta_i) = 0$

we can still **consistently**[2] estimate ρ with the formula

$$\hat{\rho}_{IV} = (W'X)^{-1}W'y$$

where $W = \begin{bmatrix} x_1' & z_1 \\ x_2' & z_2 \\ . & . \end{bmatrix}$ and $X = \begin{bmatrix} x_1' & s_1 \\ x_2' & s_2 \\ . & . \end{bmatrix}$.

[2] Note: IV is not 'unbiased', see question 2.

Introduction

What you know already

1. First thing you should make sure to know is how in the proof of consistency of $\hat{\rho}_{IV}$ we need these conditions.[3]

2. Secondly, you know that there are two ways of **calculating the IV estimator**:

   ILS/GIV Either we calculate it directly.

   2SLS Or we run a so-called **first stage** of s_i on x_i' and z_i and then a **second stage** of y_i on x_i' and the fitted values from the first stage, $\hat{s}_i$.

   In practise you will use the first procedure, e.g. with the **ivreg** command in STATA. However, the second procedure is useful to understand intuitively what is going on.

[3] Check for example the chapter 4.2 in *Mostly Harmless Econometrics* or the notes from Vassilis.

Introduction

Intuition of IV

I would like you to understand

- What is IV doing intuitively?
- Intuitively, why are the conditions important?
- What precisely do the conditions mean?

Introduction

What do the conditions mean?

$Cov(z_i, s_i) \neq 0$

This says that the instrument does indeed influence s_i. We can actually check whether this condition is satisfied by running the **first stage** and this will surely be done in any paper you read using IV.

Introduction

What do the conditions mean?

$$Cov(z_i, \eta_i) = 0$$

The condition is called the **'exclusionary restriction'** and if it is satisfied we call the instrument 'exogenous'. Since it involves the unknown η_i can never be checked. Any debate about 'the validity of the instrument' is about whether this condition is actually satisfied in the example at hand. And what does it mean?

- The most important part of this condition is that the the instrument does not influence the outcome trough anything we do not control for (and which would hence be part of η_i).
- Secondly the value of the instrument itself cannot be driven by anything that drives as well y_i.

Introduction

Why are the two conditions important?

If the two conditions are satisfied you can think of the instrument as something that comes from outside the system (condition 2), 'shocks' the endogenous variable s_i (condition 1) and otherwise has no influence on y_i (condition 2).[4]

It hence creates *'quasi-random variation'* in s_i or - another way to say this - it singles out the *'exogenous variation'* in s_i.

[4] Strictly speaking: It has no other influence on y_i other than possibly through the x_i' we control for. Cause it needs to be uncorrelated from η_i, not η_i **and** x_i'.

Question 1

Common Pitfalls

Suppose the conditions are satisfied, **what can get wrong** when calculating IV/2SLS?

- When using the standard command, e.g. **ivreg** in STATA: Nothing.
- When manually calculating the first and second stage:
  - A You might not calculate the correct standard error in the second stage.
  - B You might forget to include *all* x_i' in the first stage.[5]

[5] This is important, cause even if you use the standard commands you should calculate the first stage to see whether condition 1 is satisfied. So: Calculate the correct one!

Question 1

A. Which standard errors to calculate

You run the **first stage** regression

$$s_i = x_i'\gamma + \pi z_i + \epsilon_i$$

and obtain the fitted values $\hat{s_i}$. The true **second stage** equation is then[6]

$$y_i = x_i'\gamma + \rho\hat{s_i} + \underbrace{\eta_i + \rho(s_i - \hat{s_i})}_{\nu_i}$$

But if you just run a simple OLS of y_i on x_i' and $\hat{s}_i$ this fails to realize the structure of the error ν_i and would just calculate $s^2(X'X)^{-1}$... which is wrong.

[6] Note that *by construction* s_i is uncorrelated from $s_i - \hat{s_i}$ and *by assumption* s_i and hence $\hat{s_i}$ are uncorrelated from η_i. Hence A3Rsru is satisfied in this equation!

Question 1

B. Which first stage to calculate?

Suppose you run the **correct first stage** regression

$$s_i = x_i'\delta + z_i\pi + \epsilon_i.$$

Then *by construction* the OLS residuals $(s_i - \hat{s}_i)$ will be uncorrelated from x_i' and z_i.

Question 1

B. Which first stage to calculate?

But suppose **you forget to include all** x_i' in the first stage regression and run instead

$$s_i = w_i'\delta + \pi z_i + \epsilon_i$$

where w_i' is a subset of the covariates x_i'. The $(s_i - \hat{s}_i)$ from this is still uncorrelated from w_i' and z_i, but most likely not from the remaining variables in x_i'. Hence in the second stage

$$y_i = x_i'\gamma + \rho\hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i)]$$

x_i' is likely correlated with $(s_i - \hat{s}_i)$ and hence **A3Rsru** does not hold.

Question 2

Understanding weak instruments

We saw how our instrument z_i - to actually 'shock' the endogenous variable s_i - needs to be correlated with it. What is the consequence if this correlation is low?

Let us derive a formulation of the bias of IV which we can interpret.

Forget about the x_i' for this question and suppose we have a matrix of instruments Z. Then the 2SLS estimator is

$$\hat{\rho}_{2SLS} = \rho + (s'P^Z s)^{-1} s'P^z \eta$$

where $P^Z = Z(Z'Z)^{-1}Z$ and its bias is

$$E[\hat{\rho}_{2SLS} - \rho] = E[(s'P^Z s)^{-1} s'P^Z \eta].$$

Substituting the first stage relation $s = Z\pi + \epsilon$ we get[7]

$$\begin{aligned} E[\hat{\rho}_{2SLS} - \rho] &= E[(s'P^Z s)^{-1}[Z\pi + \epsilon]'P^Z \eta] \\ &= E[(s'P^Z s)^{-1}(\pi' Z'\eta)] + E[(s'P^Z s)^{-1}(\epsilon' P^Z \eta)] \end{aligned}$$

[7] Using that $Z'P^Z = (P^Z Z)' = Z'$.

Question 2

A. Rewriting the bias

You have seen many times that generally $E[a \cdot b] \neq E[a] \cdot E[b]$. But in this special case there is a fairly complicated proof that we can rewrite this expression approximately as

$$E[\hat{\rho}_{2SLS} - \rho] \approx (E[s' P^Z s])^{-1} E[\pi' Z' \eta] + (E[s' P^Z s])^{-1} E[\epsilon' P^Z \eta]$$

By condition 2 for a valid instrument $E[\pi' Z' \eta] = 0$ and since moreover $E[\pi' Z' \epsilon] = 0$ we have

$$\begin{aligned}
E[\hat{\rho}_{2SLS} - \rho] &\approx (E[s' P^Z s])^{-1} E[\epsilon' P^Z \eta] \\
&= (E[(\pi' Z' + \epsilon') P^Z (Z\pi + \epsilon)])^{-1} E[\epsilon' P^z \eta] \\
&= (E[\pi' Z' Z\pi + \pi' Z' \epsilon + \epsilon' Z\pi + \epsilon' P^Z \epsilon])^{-1} E[\epsilon' P^Z \eta] \\
&= (E[\pi' Z' Z\pi + \epsilon' P^Z \epsilon])^{-1} E[\epsilon' P^Z \eta]
\end{aligned}$$

Question 2

B. Understanding the bias

Using Vassilis usual trace-trick you can show that

$$E[\epsilon' P^Z \epsilon] = \sigma_\epsilon^2 Q$$

and

$$E[\epsilon' P^Z \eta] = \sigma_{\eta\epsilon} Q$$

where Q is the the rank of P^Z and hence

$$\begin{aligned} E[\hat{\rho}_{2SLS} - \rho] &\approx (E[\pi' Z' Z \pi + \epsilon' P^Z \epsilon])^{-1} E[\epsilon' P^Z \eta] \\ &= (E[\pi' Z' Z \pi] + \sigma_\epsilon^2 Q)^{-1} \sigma_{\eta\epsilon} Q \\ &= \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \left[\frac{E[\pi' Z' Z \pi]/Q}{\sigma_\epsilon^2} + 1 \right]^{-1} \end{aligned}$$

Question 2

B. Understanding the bias

We found

$$E[\hat{\rho}_{2SLS} - \rho] \approx \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \left[\frac{E[\pi' Z' Z \pi]/Q}{\sigma_\epsilon^2} + 1 \right]^{-1}$$

Now we see what creates the bias in 2SLS: $\sigma_{\eta\epsilon}$. Intuitively, since $\hat{s}_i$ is *estimated* it will be fitted towards very high and low errors ϵ_i. But if these are correlated with η_i, then $\hat{s}_i$ is still correlated with η_i.

But now we will see how 'strong' instruments help.

Question 2

B. Understanding 'strong' instruments

We can realize that $\frac{E[\pi' Z' Z \pi]/Q}{\sigma_\epsilon^2}$ is the population explained sum of squares of the **first stage** over the population error sum of squares of the first stage, so it is the **population F-statistic of the first stage**.[8] Hence

$$E[\hat{\rho}_{2SLS} - \rho] \approx \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \left[\frac{E[\pi' Z' Z \pi]/Q}{\sigma_\epsilon^2} + 1 \right]^{-1} = \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \frac{1}{F+1}$$

and hence as $F \to \infty$, Bias $\to 0$.

So with 'strong' instruments the bias vanishes.

[8] The actual F-Stat would be $[\hat{\pi}' Z' Z \hat{\pi}/Q] \cdot (1/\hat{\sigma}_\epsilon^2)$. This is the population analog to which the sample F-statistic will tend if the sample gets very big.

Question 2

C. Understanding 'weak' instruments

What happens if the instruments are 'weak'?

$$E[\hat{\rho}_{2SLS} - \rho] \approx \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \frac{1}{F+1} \rightarrow \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2}, \text{ as } F \rightarrow 0$$

Remember that the bias in the OLS estimate was $\frac{\sigma_{\eta\epsilon}}{\sigma_s^2}$. And how will F be 0? By the instrument having no influence on s_i, or $\pi = 0$. But then $s_i = \epsilon_i$ and hence $\sigma_s^2 = \sigma_\epsilon^2$.

So with very 'weak' instruments, the 2SLS bias tend to the OLS bias.

(Intuition: The instrument doesn't help at all.)

Question 2

D. Is the first stage good?

If we have a small sample and want to see whether our instrument(s) are strong, calculating the F-**Statistic of the excluded instruments** from the first stage is informative.