Microeconometrics: Chapter 5 Instrumental Variables Estimation

Per Johansson

Yau Mathematical Sciences Center, Tsinghua University
Dept. of Statistics, Uppsala University and IFAU

October 8, 2021

History and motivation of IV

To establish a causal relationship we would ideally like to make a 'ceteris paribus' comparison where we change the variable of interest (say a policy variable) while holding everything else constant.

The problem is that such comparisons are difficult to make because of the various interdependencies in observational data, i.e., the daily life observations we have on individual and firm behavior.

The initial response to this challenge was to model all conceivable interdependencies.

History and motivation of IV

The econometricians of the 1930s and 1940s, thus, built simultaneous equations models with the aim of controlling for all interdependencies.

Economic theory was used to motivate the exclusion of some variables from certain parts of the model of simultaneously determined variables (Haavelmo, 1943).

However, the idea of identification in the simultaneous equation models had been solved by father and son, Philip and Sewal, Wright, almost 20 years earlier, but since forgotten (Stock and Trebbi, 2003).

Philip Wright's interest was on estimating supply and demand elasticizes using data on quantity and price of a good.

History and motivation of IV

They showed that in order to identify the demand elasticity one needs a variable not within the 'the model' that changes the supply. This type of variables is now known as an instrumental variable (IV).

Chapter 7 deals with systems of "Systems of Equations"with no simultaneity problems, while Chapter 8 extends on this chapter, primarily by discussing Generalized methods of moments (GMM) estimation with dynamic models (not to be discussed in this class).

Chapter 9 focus is on identification with potential simultaneity between the variables and with an interest of identifying parameters in two or more simultaneously determined equations (not covered here).

I will however, shortly try to provide an intuition to the problem addressed by the Wright's.

Simultaneous models

Example (Individual labor supply) Consider a labor supply function and a wage offer (inverse labor demand) function:

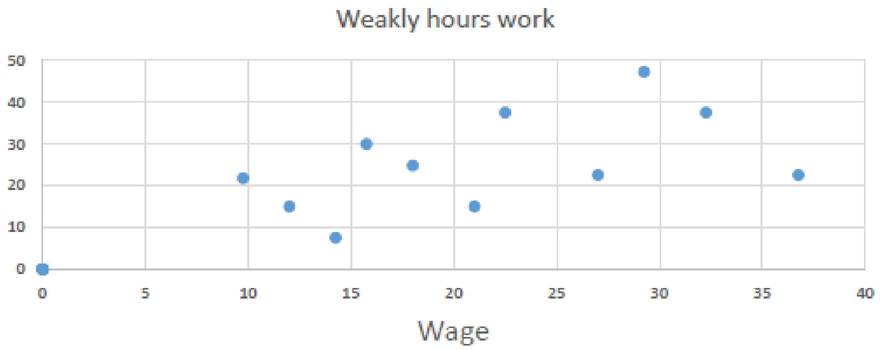
$$
\begin{aligned}
LS : h^s(w) &= \gamma_1 w + \boldsymbol{x}_1 \boldsymbol{\delta}_1 + u_1 \\
WO : w^o(h) &= \gamma_2 h + \boldsymbol{x}_2 \boldsymbol{\delta}_2 + u_2
\end{aligned}
$$

LS is the labor supply function, which shows how much each unit in the population would work at any given wage, w.

If we could run the right experiment, we could use OLS to estimate γ_1

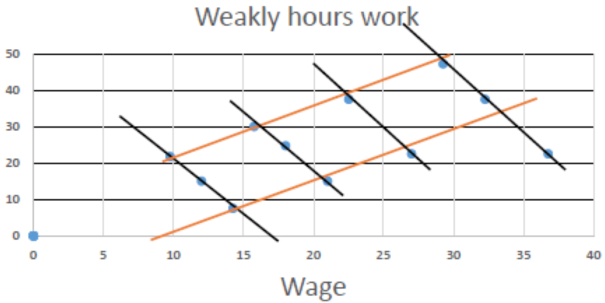
We bring in the wage offer (WO) function to recognize that, for retrospective data, we observe the pair (h_i, w_i), with w_i not being randomly assigned.

Simultaneous models



Simultaneous models

What is a sensible assumption about how (h_i, w_i) are generated? A standard approach is to assume that we observe equilibrium hours and wages for each individual.



History and motivation of IV

That is, the data are generated as

$$
\begin{aligned}
h_i &= \gamma_1 w_i + \boldsymbol{x}_{i1}\boldsymbol{\delta}_1 + u_{i1} \\
w_i &= \gamma_2 h_i + \boldsymbol{x}_{i2}\boldsymbol{\delta}_2 + u_{i2}
\end{aligned}
$$

In the usual case, we assume that the elements of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are exogenous in both equations.

If the set of $\boldsymbol{x}_1$ differ from the set $\boldsymbol{x}_2$, then γ_1 and/or γ_2 are in general identified.

Estimation of all parameters in both equations requires at least two instruments.

Simultaneous models

With the focus on estimating γ_1 we would estimate

$$w_i = \beta_0 + \boldsymbol{\alpha}' \boldsymbol{z}_i + \varepsilon_i \tag{1}$$

where $\boldsymbol{z}_i$ is all exogenous variables and then use $\widehat{w_i}$ as the "exogenized" wage in the LS equation, thus

$$h_i = \gamma_1 \widehat{w_i} + \boldsymbol{x}_{i1} \boldsymbol{\delta}_1 + u_{i1}$$

The OLS estimator on this equation is known as the two stage least square (2SLS) estimator.

IV and selection

Assume interest is in estimating β_1 in

$$y = \beta_0 + \beta_1 w + u, \tag{2}$$

but that ω_i in

$$w_i = \delta_0 + \delta_1 z_i + \omega_i, \tag{3}$$

is correlated with u_i. Then OLS on (2) would be a biased estimator of β_1.

IV and selection

Note that

$$y_i = \alpha_0 + \alpha_1 z_i + \varepsilon_i, \tag{4}$$

where, $\alpha_0 = \delta_0 + \beta_0\gamma_0$ and $\alpha_1 = \beta_1\delta_1$.

Thus, if we first estimate (3) and (4) with OLS we can estimate β_1

$$\widehat{\beta}_1 = \widehat{\alpha}_1/\widehat{\gamma}_1.$$

This is the WALD estimator. In this situation (one instrument and no $\boldsymbol{x}$) WALD = 2SLS. With 2SLS, the predictions from the OLS on (3) is used as a regressor in (2).

IV and selection

Example: Smoking and birth weight of children, (Permut and Hebel, 1989) (an encouragement design).

- Outcome y child's birth weight
- Treatment w: $w_i = 1$ stopped smoking and $w_i = 0$ still smoking
- z_i : a letter, sent (or not) randomly to smoking mothers with information about the danger with smoking.
- 20 % stopped in $z = 0$ while 43% in $z = 1$ why $\widehat{\delta_1} = 0.23$. Estimated α_1, $\widehat{\alpha_1} = 98$

$$\widehat{\beta_1} = 98/0.23 \simeq 426$$

That is, an increase in bwght with around 426 gram if quit smoking.

We will return to this example in the quasi-experiment lectures.

The simple case

Simple linear model in the population:

$$y = \beta_0 + \beta_1 w + u \tag{1}$$

where u is thought to be correlated with w (which can have any kind of features – discrete, continuous, or hybrid).

If $Cov(w, u) \neq 0$ (w is "endogenous") then the OLS estimator will be inconsistent for β_1.

If we have a rich set of controls, we might be able to break the link between u and w – but that is an selection on observable (or 'OLS') solution.

An instrumental variable, z, for w has two properties:

$$Cov(z,u) = 0 \quad \text{(exogeneity)} \tag{2}$$

$$Cov(z,w) \neq 0 \quad \text{(relevance)} \tag{3}$$

A key difference is that while we must take (2) on faith (or have other ways of checking it), we can test the null that z and w are uncorrelated in the sample.

Apply the covariance operator to $y = \beta_0 + \beta_1 w + u$ and use the fact that it is a linear operator

$$Cov(z,y) = \beta_1 Cov(z,w) + Cov(z,u)$$

and then impose (2) and (3) to get

The simple case

$$\beta_1 = \frac{Cov(z, y)}{Cov(z, w)}. \quad (4)$$

Equation (4) establishes that β_1 is identified because it is a function of population moments of variables we observe.

Replacing the population covariances with the sample covariances gives us the so-called instrumental variables estimator for the simple regression model:

$$\begin{aligned}\hat{\beta}_{1,IV} &= \frac{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(w_i - \overline{w})} \\ &= \beta_1 + \frac{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})u_i}{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(w_i - \overline{w})}\end{aligned}$$

The simple case

Consistency of $\hat{\beta}_{1,IV}$ follows by the law of large numbers and the algebra of plims.

Even if we restrict attention to consistency, it is not true that one should use a "slightly" endogenous instrument rather than OLS. Why?

The simple case

Using $Cov(w,u) = Corr(w,u)\sigma_w\sigma_u$ it can be shown that

$$p\lim \hat{\beta}_{1,OLS} = \beta_1 + \frac{\sigma_u}{\sigma_w} \cdot Corr(w,u) \tag{5}$$

and

$$p\lim \hat{\beta}_{1,IV} = \beta_1 + \frac{\sigma_u}{\sigma_w} \cdot \frac{Corr(z,u)}{Corr(z,w)} \tag{6}$$

If $Corr(z,w)$ is small – that is, z is a "weak instrument" – then even a small correlation between z and u the asymptotic bias of IV can be larger than for the OLS. [Also common to see IV estimates that are larger in magnitude than OLS estimates.]

The simple case

Under a homoskedasticity assumption (to be made precise later),

$$Avar\sqrt{n}(\hat{\beta}_{1,IV} - \beta_1) = \frac{\sigma_u^2}{\sigma_w^2 \rho_{z,w}^2}, \qquad (7)$$

where $\rho_{z,w} = Corr(z, w)$. Thus, the asymptotic standard deviation of $\sqrt{n}(\hat{\beta}_{1,IV} - \beta_1)$ is

$$\frac{\sigma_u}{\sigma_w} \cdot \frac{1}{|\rho_{z,w}|}$$

When $|\rho_{z,w}|$ is small, the asymptotic variance can be very large. The formula for the OLS estimator omits $|\rho_{z,w}|$.

Weak instruments also make the usual normal asymptotics highly suspect. See (Nelson and Startz, 1990b,a; Bound et al., 1995; Stock et al., 2002; Staiger and Stock, 1997). Imbens and Wooldridge (2009) contains a recent discussion.

The general case

Start again with the population model

$$y = \boldsymbol{c}\beta + u, \tag{9}$$

where $\boldsymbol{c} = (c_1, c_2, ..., c_H)$ is $1 \times H$ vector of covariates and β the corresponding parameter vector, and in all cases, $c_1 = 1$.

Note: no such thing as an OLS "model" or IV "model." OLS and IV are different *estimation* methods that can be applied to the same model. They are consistent under different assumptions.

The general case

Let $\boldsymbol{z} = (z_1, z_2, ..., z_L)$ be a $1 \times L$ vector, where $z_1 = 1$. Further, $\boldsymbol{z}$ contains all exogenous elements of $\boldsymbol{c}$. But, if one or more elements of $\boldsymbol{c}$ is correlated with u, $\boldsymbol{z}$ must contain some outside variables. $\boldsymbol{z}$ is exogenous;

$$E(\boldsymbol{z}'u) = \mathbf{0}. \tag{10}$$

How does this assumption – these moment or orthogonality conditions – allow us to identify β?

The general case

Suppose $L = H$. For example, $\boldsymbol{c} = (1, x_2, ..., x_K, w)$ and $\boldsymbol{z} = (1, x_2, ..., x_K, z)$, so that only w is (possibly endogenous) and z as an IV for w. Using (9) and (10),

$$
\begin{aligned}
E(\boldsymbol{z}'y) &= E(\boldsymbol{z}'\boldsymbol{c})\beta + E(\boldsymbol{z}'u) \\
&= E(\boldsymbol{z}'\boldsymbol{c})\beta
\end{aligned}
$$

The general case

If we assume the *rank condition*

$$rank\ E(\boldsymbol{z}'\boldsymbol{c}) = H \tag{12}$$

then

$$\beta = [E(\boldsymbol{z}'\boldsymbol{c})]^{-1}E(\boldsymbol{z}'y), \tag{13}$$

which extends the moment condition for OLS (which is the special case $\boldsymbol{z} = \boldsymbol{x}$).

The general case

Given a random sample,

$$\hat{\beta}_{IV} = \left(n^{-1} \sum_{i=1}^{n} \boldsymbol{z}_i' \boldsymbol{c}_i \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} \boldsymbol{z}_i' y_i \right). \tag{14}$$

Using the algebra of plims and the WLLN, $p\lim(\hat{\beta}_{IV}) = \beta$ under (10) and (12).

In matrix notation we can write

The general case

$$\hat{\beta}_{IV} = (\boldsymbol{Z}'\boldsymbol{C}/n)^{-1}(\boldsymbol{Z}'\boldsymbol{Y}/n) = (\boldsymbol{Z}'\boldsymbol{C})^{-1}\boldsymbol{Z}'\boldsymbol{Y},$$

where

$$\underset{n\times H}{\boldsymbol{C}} = \begin{pmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \\ \vdots \\ \boldsymbol{c}_n \end{pmatrix}, \underset{n\times 1}{\boldsymbol{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \underset{n\times H}{\boldsymbol{Z}} = \begin{pmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \\ \boldsymbol{z}_n \end{pmatrix}.$$

The general case

The condition $E(\mathbf{z}'u) = \mathbf{0}$ is called *instrument exogeneity*; the rank condition (12) is also called *instrument relevance*.

Without extra information we cannot test $E(\mathbf{z}'u) = \mathbf{0}$ because u is unobserved.

Write the *reduced form* of w as

$$w = \gamma_1 + \gamma_2 x_2 + \ldots + \gamma_K x_K + \theta_1 z + r_H, \tag{16}$$

where, by *definition*,

The general case

$$E(r_H) = 0,\ Cov(x_k, r_H) = 0,\ k = 2, ..., K,\ Cov(z, r_H) = 0. \quad (17)$$

In other words, the linear projection of x_K on $(1, x_2, ..., x_K, z)$ is

$$L(w|1, x_2, ..., x_K, z) = \gamma_1 + \gamma_2 x_2 + ... + \gamma_K x_K + \theta_1 z. \quad (18)$$

The rank condition (12) holds if and only if

$$\theta_1 \neq 0. \quad (19)$$

OLS consistently estimates the parameters of a linear projection.

The general case

Need to reject

$$H_0 : \theta_1 = 0 \tag{20}$$

in favor of (19). Heteroskedasticity-robust inference can be used.

We do not care about the γ_k in (18), but $x_2, ..., x_K$ must be partialled out. (z could be correlated with w, but we require that w is *partially* correlated with z.)

The general case

w can be discrete, continuous, or a hybrid. Regardless of the nature of w, the linear projection is well-defined.

The IV estimator is consistent under $E(\mathbf{z}'u) = \mathbf{0}$ if

$$L(W|1, x_2, ..., x_K, z) \neq L(W|1, x_2, ..., x_K).$$

This is just another way to say that, in a linear sense, z helps to predict w controlling for the other exogenous variables.

Regressing w on $1, x_2, ..., x_K, z$ using the data is often called the **first-stage regression**.

It should be done to establish sufficient partial correlation between w and z.

The general case

A reduced form also exists for y, and can be written

$$y = \gamma_1 + \gamma_2 x_2 + \ldots + \gamma_K x_K + \gamma_H z + v$$

If, e.g., w is years of education and z is mothers education, the reduced form for y, estimates the "effect" of mothers education on y.

The "structural" equation, $y = \beta_1 + \beta_2 x_2 + \ldots + \beta_H w + u$, estimated by IV, attempts to get at the causal effect years of schooling.

Two Stage Least Squares

In some cases, we have more instruments than we need. For example, if we can use mother's education as an IV, why not father's education, too?

Again write

$$
\begin{aligned}
y &= \boldsymbol{c}\beta + u \\
E(\boldsymbol{z}'u) &= \boldsymbol{0}
\end{aligned}
$$

where $L = \dim(\boldsymbol{z}) \geq \dim(\boldsymbol{c}) = H$.

When $L > H$, have more than one IV estimator. We say the model is (potentially) **overidentified**.

When $L = H$ and the rank condition holds, the model is **just identified**.

Two Stage Least Squares

Suppose z_1 and z_2 are IVs for w. Which should we use?

Under a homoskedasticity assumption, the best IV for w is the linear combination of *all* exogenous variables defined by the linear projection.

In general, the best vector of IVs for $\boldsymbol{c}$ is the vector of linear projections of each element of $\boldsymbol{c}$ on $\boldsymbol{z}$.

Write the LPs in error form as

$$\boldsymbol{c} = \boldsymbol{z}\boldsymbol{\Pi} + \boldsymbol{r},$$

where $\boldsymbol{\Pi}$ is the $L \times H$ matrix

$$\boldsymbol{\Pi} = [E(\boldsymbol{z}'\boldsymbol{z})]^{-1}[E(\boldsymbol{z}'\boldsymbol{c})]$$

Two Stage Least Squares

and

$$E(\boldsymbol{z}'\mathbf{r}) = \mathbf{0}.$$

For each c_h we can write

$$c_h = \boldsymbol{z}\boldsymbol{\pi}_h + r_h \equiv c_h^* + r_h$$

where $\boldsymbol{\pi}_h$, the $(L \times 1)$ vector, is the h^{th} column of $\mathbf{\Pi}$. For any $c_h \in \boldsymbol{z}$, $c_h^* = c_h$, so exogenous variables act as their own instruments. In the general case, use

$$\boldsymbol{c}^* = \boldsymbol{z}\mathbf{\Pi},$$

as the $1 \times H$ vector of instruments for $\boldsymbol{c}$. Because $\boldsymbol{z}$ is exogenous, so is $\boldsymbol{c}^*$:

$$E(\boldsymbol{c}^{*\prime}u) = \mathbf{0}.$$

Two Stage Least Squares

The rank condition becomes

$$rank\ E(\boldsymbol{c}^{*\prime}\boldsymbol{c}) = H.$$

But

$$E(\boldsymbol{c}^{*\prime}\boldsymbol{c}) = \boldsymbol{\Pi}' E(\boldsymbol{z}'\boldsymbol{c}) = E(\boldsymbol{c}'\boldsymbol{z})[E(\boldsymbol{z}'\boldsymbol{z})]^{-1} E(\boldsymbol{z}'\boldsymbol{c}).$$

Two Stage Least Squares

Formally, here are the first two assumptions for 2SLS, stated in the population. (We assume access to a random sample).

Assumption 2SLS.1 (Exogenous Instruments): $E(\mathbf{z}'u) = \mathbf{0}$.

Assumption 2SLS.2 (Rank Condition): (a) $rank\ E(\mathbf{z}'\mathbf{z}) = L$; (b) $rank\ E(\mathbf{z}'\mathbf{c}) = H$.

Part (a) rules out perfect collinearity among the exogenous variables (which means we cannot use linear combinations of exogenous variables as additional instruments).

Part (b) is the practically important restriction, and requires $L \geq H$.

Two Stage Least Squares

Deriving 2SLS: With

$$\beta = [E(\boldsymbol{c}^{*\prime}\boldsymbol{c})]^{-1}E(\boldsymbol{c}^{*\prime}Y),$$

need to worry about unknown $\mathbf{\Pi}$ because $\boldsymbol{c}^* = \boldsymbol{z}\mathbf{\Pi}$.

Two-step estimation: (1) Run the regression $\boldsymbol{c}_i$ on $\boldsymbol{z}_i$, $i = 1, ..., n$ to obtain $\hat{\mathbf{\Pi}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{C}$. Obtain the vector fitted values,

$$\hat{\boldsymbol{c}}_i = \boldsymbol{z}_i\hat{\mathbf{\Pi}},\ i = 1, ..., n.$$

This is the same as regressing each element of $\boldsymbol{c}_i$ not in $\boldsymbol{z}_i$ on $\boldsymbol{z}_i$, and obtaining the fitted values. Any element of $\boldsymbol{c}_i$ in $\boldsymbol{z}_i$ is used as its own fitted value.

Two Stage Least Squares

and (2) Use $\hat{\boldsymbol{c}}_i$ as the vector of IVs for $\boldsymbol{c}_i$:

$$\hat{\beta}_{IV} = \left(n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{c}}_i'\boldsymbol{c}_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{c}}_i' y_i\right).$$

We can write this differently. Because

$$\begin{aligned}\boldsymbol{c}_i &= \hat{\boldsymbol{c}}_i + \hat{\boldsymbol{r}}_i \\ \sum_{i=1}^{n}\hat{\boldsymbol{c}}_i'\hat{\boldsymbol{r}}_i &= \boldsymbol{0} \text{ (by OLS FOCs).}\end{aligned}$$

Two Stage Least Squares

So

$$\sum_{i=1}^{n} \hat{\mathbf{c}}_i' \mathbf{c}_i = \sum_{i=1}^{n} \hat{\mathbf{c}}_i' \hat{\mathbf{c}}_i$$

and then the IV estimator can be written as a 2SLS estimator:

$$\hat{\beta}_{2SLS} = \left(n^{-1} \sum_{i=1}^{n} \hat{\mathbf{c}}_i' \hat{\mathbf{c}}_i \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} \hat{\mathbf{c}}_i' Y_i \right).$$

Two Stage Least Squares

The first-stage regression is $\boldsymbol{c}_i$ on $\boldsymbol{z}_i$ to get the fitted values, $\hat{\boldsymbol{c}}_i$.

The second-stage regression is y_i on $\hat{\boldsymbol{c}}_i$.

The 2SLS algorithm is not really the best way to think about the estimator.

(1) standard errors from second-stage regression are not correct.

(2) using the fitted values as IVs is not the same as 2SLS approach for some panel data applications.

(3) the two-step approach can lead to abuse for simple nonlinear models.

Two Stage Least Squares

The 2SLS residuals are defined as

$$\hat{u}_i = Y_i - \boldsymbol{c}_i \hat{\beta}_{2SLS},$$

where it is $\boldsymbol{c}_i$, not $\hat{\boldsymbol{c}}_i$, multiplying $\hat{\beta}_{2SLS}$. An algebraic fact is that the $\hat{u}_i$ are orthogonal to $\hat{\boldsymbol{c}}_i$ in the sample.

This is the condition that determines $\hat{\beta}_{2SLS}$:

$$\sum_{i=1}^{n} \hat{\boldsymbol{c}}_i' \hat{u}_i = \mathbf{0}$$

Two Stage Least Squares

Using full data matrices and some algebra, we can write

$$\begin{aligned}\hat{\beta}_{2SLS} &= (\hat{\boldsymbol{C}}'\hat{\boldsymbol{C}})^{-1}\hat{\boldsymbol{C}}'\boldsymbol{Y} \\ &= [(\boldsymbol{C}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{C})]^{-1}(\boldsymbol{C}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{Y}) \\ &= \beta + [(\boldsymbol{C}'\boldsymbol{Z}/n)(\boldsymbol{Z}'\boldsymbol{Z}/n)^{-1}(\boldsymbol{Z}'\boldsymbol{C}/n)]^{-1}(\boldsymbol{C}'\boldsymbol{Z}/n)(\boldsymbol{Z}'\boldsymbol{Z}/n)^{-1}(\boldsymbol{Z}'\boldsymbol{U}/n)\end{aligned}$$

where the last expression can be used to show consistency by applying the WLLN to each term, along with the rank condition and $E(\boldsymbol{z}'u) = \boldsymbol{0}$.

Two Stage Least Squares

Key Result: Under 2SLS.1 and 2SLS.2, $\hat{\beta}_{2SLS}$ on a random sample is consistent for β.

For inference, it is useful to show

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) = \left(n^{-1} \sum_{i=1}^{n} \mathbf{c}_i^{*\prime} \mathbf{c}_i^{*} \right)^{-1} \left(n^{-1/2} \sum_{i=1}^{n} \mathbf{c}_i^{*\prime} u_i \right) + o_p(1)$$

where the $\mathbf{c}_i^* = \mathbf{z}_i \mathbf{\Pi}$.

Two Stage Least Squares

It follows that

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1})$$

$$\begin{aligned} \boldsymbol{A} &= E(\boldsymbol{c}_i^{*\prime}\boldsymbol{c}_i^*) \\ \boldsymbol{B} &= E(u_i^2\boldsymbol{c}_i^{*\prime}\boldsymbol{c}_i^*). \end{aligned}$$

Two Stage Least Squares

Assumption 2SLS.3 (Homoskedasticity): u^2 is uncorrelated with all elements of $\boldsymbol{z}$ as well as z_l^2 and $z_l z_{l'}$, $l \neq l'$.

That is,

$$E(u^2\boldsymbol{z}'\boldsymbol{z}) = E(u^2)E(\boldsymbol{z}'\boldsymbol{z}) \equiv \sigma^2 E(\boldsymbol{z}'\boldsymbol{z}).$$

Key Result: Under 2SLS.1, 2SLS.2, and 2SLS.3,

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \boldsymbol{A}^{-1})$$

Consistent (not unbiased) estimators of σ^2 and $\boldsymbol{A}$:

Two Stage Least Squares

$$\hat{\sigma}^2 = (n - H)^{-1} \sum_{i=1}^{n} \hat{u}_i^2 \xrightarrow{p} \sigma^2$$

$$\hat{\boldsymbol{A}} = n^{-1} \sum_{i=1}^{n} \hat{\boldsymbol{c}}_i' \hat{\boldsymbol{c}}_i$$

Under 2SLS.1, 2SLS.2, and 2SLS.3, we can use

$$\widehat{Avar}(\hat{\beta}_{2SLS}) = \hat{\sigma}^2 \hat{\boldsymbol{A}}^{-1} / n = \hat{\sigma}^2 (\hat{\boldsymbol{C}}' \hat{\boldsymbol{C}})^{-1}$$

Heteroskedasticity-Robust Inference

$$\widehat{Avar}(\hat{\beta}_{2SLS}) = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^{-1}/n = \left(\sum_{i=1}^{n} \hat{\boldsymbol{c}}_i'\hat{\boldsymbol{c}}_i\right)^{-1}\left(\sum_{i=1}^{n} \hat{u}_i^2\hat{\boldsymbol{c}}_i'\hat{\boldsymbol{c}}_i\right)\left(\sum_{i=1}^{n} \hat{\boldsymbol{c}}_i'\hat{\boldsymbol{c}}_i\right)^{-1}$$

which is not the same as using the heteroskedasticity-robust inference in the second stage regression Y_i on $\hat{\boldsymbol{c}}_i$, as then

$$\widehat{\hat{u}}_i = Y_i - \hat{\boldsymbol{c}}_i\hat{\beta}_{2SLS},$$

Efficiency: 2SLS has the smallest asymptotic variance among all IV estimators using linear functions of $\boldsymbol{z}_i$ as instruments under 2SLS.1, 2SLS.2, and 2SLS.3. (Only meaningful when $L > H$.)

Weak IVs and homogeneous treatment effects

If IV's are weak all IV estimators (including the 2SLS) are non-normal, can be badly biased and t-test may fail to control the size (cf. (Nelson and Startz, 1990b,a; Bound et al., 1995; Stock et al., 2002)).

Based on the idea of test inversion, a method for valid inference also with weak IVs in just identified models were suggest already in 1949 (Anderson and Rubin, 1949). See Andrews et al. (2019) for a recent review of valid IV methods when the IVs are week.

However, the tradition is to use the F-statistic (t with one IV) of the importance of the instrument(s) in a regression of the treatment on instrument(s).

The conditional likelihood ratio test (Moreira, 2003; Blundell et al., 2005; Andrews et al., 2006) is a test that overcomes the distortion in standard t-test (of an effect) based on the IV estimator by adjusting the critical value for the hypothesis test to depend on the given sample.

However, as these test are complicated, most often the rule of thumb suggested in Staiger and Stock (1997) is used. Here an F-statistic above 10 are consider to be not weak and relevant.

This procedure have been shown Andrews et al. (2019) to lead to size distortions of the test of a treatment effect.

Anderson, T. and Rubin, H. (1949). stimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20:46–63.

Andrews, D. W. K., Moreira, M. J., and Stock, J. H. (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica*, 74(3):715–752.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753.

Blundell, R., Newey, W., and Persson, T., editors (2005). *Inference with Weak Instruments*. Cambridge University Press, Cambridge. Prepared for the 2005 World Congress of the Econometric Society.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048.

Nelson, C. R. and Startz, R. (1990a). The distribution of the instrumental variable estimator and its t ration when the instrument is a poor one. *Journal of Business*, 63(63):124–140.

Nelson, C. R. and Startz, R. (1990b). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica*, 58(4):967–976.

Permut, T. and Hebel, J. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 45(2):619–622.

Staiger, D. and Stock, J. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Stock, J. H. and Trebbi, F. (2003). Retrospectives who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics*, 20(4):518–529.