Econometrics 1
Lecture 10: Instrumental variables I

ENSAE 2014/2015

Michael Visser (CREST-ENSAE)

Motivation

This (and the next) lecture studies the so-called *Instrumental Variable* (IV) method.

To motivate this method, let us reconsider the wage/education example.

Suppose that in the population the hourly wage is related to education and ability (*abil*) in the following way:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + \xi. \tag{1}$$

where the error term ξ has the usual satisfactory properties.

If we had a random sample on wages, years of education and ability levels, we could regress $log(wage)$ on a constant, *educ* and *abil*, and obtain unbiased and consistent estimators of the parameters of interest.

Motivation (cntd)

Unfortunately, in most data sets, the ability of individuals is unobserved. If ability is indeed not recorded in the data source, we would consider the simple regression model

$$log(wage) = \beta_0 + \beta_1 educ + u \tag{2}$$

where the error term u contains the missing variable *abil* ($u = \beta_2 abil + \xi$).

The OLS estimator resulting from regressing $log(wage)$ on a constant and *educ* results in biased and inconsistent estimators if *educ* and *abil* are correlated and $\beta_2 \neq 0$.

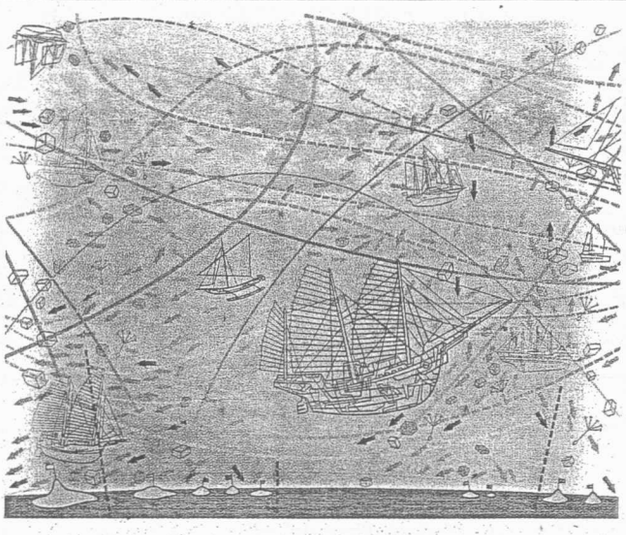
The IV method can be used to solve the above omitted variable bias.

The method requires an additional variable, called an *instrumental variable* for the regressor *educ*.

The instrumental variable, denoted z, needs to satisfy two requirements: i) z is uncorrelated with u; ii) z is correlated with *educ*.

Economics focus | Winds of change

Why economists love empires



IN A speech last year at Oxford University, Manmohan Singh, India's prime minister, caused a stir in his homeland by noting a few "beneficial consequences" of India's years under British rule, including its free press, its civil service and its "notion" of the rule of law. But he also pointed out that India, one of the world's biggest economies in 1700, was impoverished by the time the British left.

However mixed empire's legacy in India, imperialism has recently provided a rich feast for economists. Their interest lies not just in totting up the balance sheet of colonial rule—although that can be fascinating. They are after even bigger game: an explanation of why some countries grow rich and others do not. Of the many proposed solutions to that riddle (technology, geography, the Protestant ethic) the current favourite is rather bland in the abstract: "institutions". In rich economies institutions—meaning the formal laws and unwritten rules that govern society—function rather well on the whole. In poor ones they don't. That much is indisputable.

What is tricky is showing that good institutions are a cause of economic progress rather than a by-product of it. You cannot run controlled experiments in which a particular institution is randomly imposed on some countries, but not on others, in order to compare how they fare. Or at least economists can't. But perhaps imperialists can. Maybe the colonial adventures of the past provide the natural experiments economists need to put their theories to the test.

The imperial powers certainly generated a lot of institutional variety, sprinkling Spanish vassalage, British indirect rule and American paternalism across the globe. But was this variation random? Surely not. Imperialists vied to plant their flag in the most lucrative spots, wherever the spices were rich or the sugar cane tall. Thus a conundrum remains: if, say, America's former colonies have prospered compared with Spain's, was this because America bequeathed the best institutions, or because it found the most promising areas of the world to colonise?

What is ingenious about the recent economic studies of empire is how they overcome this problem. Imperial institutions may determine prosperity, but the reverse may also be true. The trick is to find some third factor that is securely linked to institutions, but entirely unconnected to economic success. Such factors are called "instrumental variables", because the economist is interested in them not for themselves, but for what they tell him about something else.

That name, however, now seems quite ironic. Because all of the fun in the recent spate of papers is in the instruments themselves. Economists are outdoing each other with ever more curious instruments, ranging from lethal mosquitoes to heirless maharajahs, or, most recently, wind speeds and sea currents.

An imperial variable

Guam, which became a Spanish colony in the 17th century and an American one at the end of the 19th, was discovered in 1521 after winds and swells carried Ferdinand Magellan, the Portuguese-born explorer, to its shores. In a recent study* of 80 such islands, all but one of which eventually fell under the imperial yoke, James Feyrer and Bruce Sacerdote of Dartmouth College argue that winds and currents dictated which islands were colonised when. The early colonialists went where their sails took them; only after steamships became the norm in the 19th century could they travel against the wind.

As a result, some islands were colonised early, some late, for reasons that had much to do with meteorology, and rather little to do with any other intrinsic attractions the islands might offer. The two authors show that the accessible islands, which lay on natural sailing routes, have prospered relative to the others. They put this down in part to the longer period these islands spent under colonial rule. A century as a colony is worth a 40% increase in today's GDP, they argue.

But as the authors point out, this striking result disguises a more disturbing fact. On many islands the original population was decimated, or worse, by European contact. After the Spanish colonised Puerto Rico in 1505, the native population fell from 60,000 to 1,500 within 30 years. The island may have since prospered, but the original islanders did not.

The study paints the British as relatively benign rulers compared with the Iberians. But instruments can cut both ways. Lakshmi Iyer of Harvard has used the technique to reveal some unhappy consequences of the Raj that might have made Mr Singh's Oxford audience squirm. The British, she points out, did not wrest direct control of India all at once. From 1848 to 1856, for example, the governor-general pursued a "doctrine of lapse", taking charge of states whenever the native ruler died without an heir. These states, then, came under British rule as a result of patrilineal misfortune, not economic potential. Ms Iyer shows that such areas had fewer schools, clinics and roads as a result of British rule. The effects lingered into the 1980s.

Once just an obscure statistical method, instrumental variables are now popping up all over the place. Daniel Hamermesh, a labour economist at the University of Texas, has joked about the "instrument police", who patrol empirical economics, forever suspicious that causality may run both ways. Indeed, "reverse causality", which was once a frustrating problem, is now seen as a chance to demonstrate ingenuity. Instruments have brought colour to the study of institutions, and sharpened the debate over colonialism, without really resolving it. But whatever the claims of empire, the instrumental variable now enjoys an almost imperial grip on the imagination of economists. ■

*Links to the papers can be found in the online version of this article.

Motivation (cntd)

In practice it is often difficult to find IVs. Some researchers are, however, particularly ingenious in finding them (see the article in The Economist, next slide).

Question: what would be a good instrumental variable in the wage/education example?

The IV for *educ* should be uncorrelated with *abil and* all other variables that affect wages. The IV should also be correlated with *educ*.

A variable such as the last digit of an individual's social security number satisfies the first condition (because it is determined randomly). But this variable is not related to the number of years of education, and is therefore a poor instrument for *educ*.

A variable such as IQ (recorded in some data sources) is correlated with *educ*, but probably also with *abil* and other characteristics that affect hourly wages. IQ is therefore also not an appropriate IV.

Motivation (cntd)

Labor economists argue that some family background variables constitute good IVs for *educ*, e.g., mother's and and father's education. Both these variables are positively correlated with the child's level of education, but uncorrelated (or only weakly) with *abil*.

Labor economists have also used the number of siblings as an IV for *educ*. This variable is associated with lower average levels of education, and is likely to be unrelated to *abil* or other variables affecting hourly wages.

IV in the SLR model

Consider the regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{3}$$

Suppose that we suspect x to be correlated with u (x is then said to be endogenous in model (3)).

A variable z is an instrumental variable for the regressor x if

- $cov(z, u) = 0$ (exclusion restriction)
- $cov(z, x) \neq 0$ (relevance)

When the exclusion restriction holds, z is said to be *exogenous* in model (3). The exclusion restriction is an assumption that cannot be tested without extra information (since u is unobserved).

Instrument relevance can be verified by estimating the following regression

$$x_i = \pi_0 + \pi_1 z_i + v_i$$

and we can (and must!) test relevance through $\mathcal{H}_0 : \pi_1 = 0$ against the two-sided alternative $\mathcal{H}_1 : \pi_1 \neq 0$

IV in the SLR model (cntd)

Now we can identify β_1 by i) noting that

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

and ii) using our exclusion restriction to write

$$\beta_1 = \frac{cov(z, y)}{cov(z, x)} \tag{4}$$

The sample analog of the right hand side of (4) is the *instrumental variable estimator* of β_1 :

$$\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \tag{5}$$

IV in the SLR model (cntd)

Remarks:

- By construction, the IV estimator is consistent for β_1 nut not unbiased.
- The IV estimator of β_0 is $\hat{\beta}_0^{IV} = \overline{y} - \hat{\beta}_1^{IV}\overline{x}$
- If $z = x$, then the IV estimator $\hat{\beta}_1^{IV}$ is the usual OLS estimator of β_1 in model (3). In other words, if $Cov(x, u) = 0$ (i.e., if x is exogenous in model (3)), x can be used as its own IV, and the resulting IV estimator coincides with the OLS estimator.

Wald Estimator

A common (and simple) example of IV is one where the instrument is binary

$$z_i \in \{0, 1\}$$

After some simple algebra the IV estimator (5) becomes (check this)

$$\hat{\beta}_1^{IV} = \frac{\overline{y}_{(z=1)} - \overline{y}_{(z=0)}}{\overline{x}_{(z=1)} - \overline{x}_{(z=0)}}$$

where $\overline{x}_{(z=1)} \equiv (|\{i : z_i = 1\}|)^{-1} \sum_{\{i:z_i=1\}} x_i$ etc. This is the so-called Wald estimator (after the statistician who first proposed the estimator).

Inference

To derive the asymptotic sampling distribution of the IV estimator, it is necessary to impose a homoscedasticity assumption on the error term (just as in the OLS setting).

The homoscedasticity condition is stated conditional on z (and not x):

$$E(u^2|z) = \sigma^2.$$

This implies $E(u^2) = \sigma^2$ and hence $Var(u) = \sigma^2$ (under the nonrestrictive assumption that $E(u) = 0$).

Given the similarity between the IV and OLS estimators, it is not surprising that the IV estimator is also normally asymptotically distributed:

$$\sqrt{n}(\hat{\beta}_1^{IV} - \beta_1) \overset{a}{\sim} N(0, \sigma^2/(\sigma_x^2 \rho_{x,z}^2)).$$

where σ_x^2 is the variance of x, and $\rho_{x,z}^2$ is the square of the correlation between x and z.

Inference (cntd)

So the asymptotic variance of $\hat{\beta}_1^{IV}$ is

$$Avar(\hat{\beta}_1^{IV}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \tag{6}$$

Recall from lecture 5 (on large sample theory) that the asymptotic variance of the OLS estimator of β_1 in model (3) is $\sigma^2/(n\sigma_x^2)$. When $z = x$, the two variances naturally coincide (in this case $\rho_{x,z}^2 = 1$).

We can compare the asymptotic variance of the IV estimator to that of the OLS estimator

$$Avar(\hat{\beta}_1^{IV}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} > \frac{\sigma^2}{n\sigma_x^2} = Avar(\hat{\beta}_1^{OLS})$$

and see that the IV variance

- is always larger than the OLS variance
- depends crucially on the correlation between z and x

Inference (cntd)

All unknown terms appearing in the asymptotic variance of the IV estimator can be estimated using the data:

- $\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$
- $\hat{\rho}_{x,z} = \frac{\sum_{i=1}^n (z_i - \overline{z})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^n (z_i - \overline{z})^2}\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}$
- $\hat{\sigma}^2 = (n-2)^{-1}\sum_{i=1}^n \hat{u}_i^2$ where $\hat{u}_i = y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} x_i$

The asymptotic variance (6) can thus be estimated by

$$\widehat{Avar}(\hat{\beta}_1^{IV}) = \frac{\hat{\sigma}^2}{n\hat{\sigma}_x^2\hat{\rho}_{x,z}^2}$$

Inference (cntd)

Using a well-known result from asymptotic statistical theory, it can be shown that

$$\sqrt{n}\frac{(\hat{\beta}_1^{IV} - \beta_1)}{\sqrt{\frac{\hat{\sigma}^2}{\hat{\rho}_{x,z}^2 \hat{\sigma}_x^2}}} = \frac{(\hat{\beta}_1^{IV} - \beta_1)}{\sqrt{\frac{\hat{\sigma}^2}{\hat{\rho}_{x,z}^2 \sum_{i=1}^n (x_i - \overline{x})^2}}} \overset{a}{\sim} N(0,1)$$

The denominator in this last expression is just the asymptotic standard error of $\hat{\beta}_1^{IV}$:

$$se(\hat{\beta}_1^{IV}) = \sqrt{\frac{\hat{\sigma}^2}{\hat{\rho}_{x,z}^2 \sum_{i=1}^n (x_i - \overline{x})^2}}.$$

The "t statistic"

$$\frac{(\hat{\beta}_1^{IV} - \beta_1)}{se(\hat{\beta}_1^{IV})}$$

thus follows approximately a $N(0,1)$ distribution in large samples (i.e., approximately a t distribution). This last result can be used to perform t tests of hypotheses involving β_1, or construct confidence intervals for β_1.

Example

Let us illustrate the IV method via estimation of wage equation (2) on a sample of 428 married women, using father's education (*fatheduc*) as an instrumental variable for *educ*.

Regressing *educ* on a constant and *fatheduc*, we find that the slope estimate is 0.269 with a standard error equal to 0.029. Since the t statistic equals 9.28, we conclude that the relevance condition holds. Recall that we cannot test the exclusion restriction.

IV estimation of the wage equation gives

$$\widehat{log(wage)} = \underset{(0.446)}{0.441} + \underset{(0.035)}{0.059}\,educ.$$

The variable *educ* is significant (but only at the 10% level). The estimate suggests that an additional year of education augments the hourly wage by almost 6%.

For comparison, the equation estimated by OLS is:

$$\widehat{log(wage)} = \underset{(0.28)}{-0.185} + \underset{(0.014)}{0.109}\,educ.$$

Example (cntd)

Now *educ* is statistically significant at any conventional significance level. The OLS estimate of the return to education is much larger than the IV estimate (11% vs 6%).

This last finding can be interpreted as follows.

If the true wage equation is (1) while we regress $log(wage)$ just on a constant and *educ*, then the OLS estimator has omitted variable bias equal to $\beta_2\tilde{\delta}_1$ (where $\tilde{\delta}_1$ is the slope estimator from the regression of *abil* on a constant and *educ*; see Lecture 3).

Since it is likely that both β_2 and $\tilde{\delta}_1$ are positive, the OLS estimator has an upward bias. On average the OLS estimate is therefore larger than the IV estimate.

IV in the MLR model

The IV method is easily extended to MLR models. Consider for simplicity a MLR model with just two regressors

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \tag{7}$$

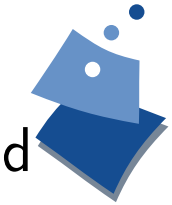
The regressor x_1 is still suspected of being correlated with u (and thus potentially endogenous), but x_2 is thought of as being exogenous (i.e., uncorrelated with u).

OLS regression of model (7) leads to biased and inconsistent estimators estimators of all parameters if indeed x_1 is endogenous.

As in the SLR setting, the method of instrumental variables can be used to solve this problem.

Let z (still) denote the IV for x_1.

IV in the MLR model (cntd)



The variable z is a valid instrument if it is exogenous (uncorrelated with the error term):

$$Cov(z, u) = E(zu) = 0 \tag{8}$$

where the first equality follows from the fact that we can always impose without loss of generality that

$$E(u) = 0 \tag{9}$$

Since the regressor x_2 is assumed to be exogenous, we also have:

$$Cov(x_2, u) = E(x_2 u) = 0 \tag{10}$$

The IV estimators of β_0, β_1, and β_2 solve the sample analogs of (9), (8), and (10):

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} x_{i1} - \hat{\beta}_2^{IV} x_{i2}) &= 0 \\ \sum_{i=1}^{n}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} x_{i1} - \hat{\beta}_2^{IV} x_{i2}) z_{i1} &= 0 \\ \sum_{i=1}^{n}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} x_{i1} - \hat{\beta}_2^{IV} x_{i2}) x_{i2} &= 0 \end{aligned}$$

IV in the MLR model (cntd)

This is a set of 3 equations in 3 unknowns, and can easily be solved.

Remark: If it is believed that x_1 is exogenous, and if we choose $z = x_1$, then the above conditions are in fact the the first order conditions for the OLS estimators in a model where y is regressed on a constant, and x_1 and x_2.

As in the SLR setting, we also need the IV z to be correlated with x_1 in a certain sense. To define the additional assumption that is needed, consider the model

$$x_1 = \pi_0 + \pi_1 z + \pi_2 x_2 + v$$

The additional assumption that must be imposed (along with (9), (8), and (10)) is

$$\pi_1 \neq 0 \tag{11}$$

IV in the MLR model (cntd)

Remarks:

- The relevance condition can be easily tested (regress x_1 on a constant, z, and x_2, and test whether the parameter associated with z is zero). However, the exclusion restriction (8) cannot be tested.
- There are no restrictions on the other parameters, i.e., π_0 and π_2 may take any values (including zero).
- The formulas for the variance and standard error of the IV estimator in the MLR setting will not be given in this course. However, as in the SLR setting, given these formulas, t statistics can be defined, and hypotheses testing can be done in the usual way.

R^2 and IV

Many econometric software packages calculate an R-squared after IV estimation using the standard OLS formula

$$R^2 = 1 - SSR/SST = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $\hat{u}_i^2$ are the squared IV residuals.

However, unlike in the case of OLS, the R-squared from IV estimation cannot be interpreted as the fraction of the sample variation in y that is explained by the regressors.

This is because in the case of IV estimation we cannot decompose SST as the sum of SSR and SSE (explained sum of squares).

So it is not very useful or informative to report an R-squared after IV estimation (it can even be negative because the SSR for IV can be larger than SST*).*

Two-stage least squares (2SLS)

Consider again the model (7) with x_1 the (potentially) endogenous regressor and x_2 the exogenous regressor.

Suppose that we now have $M > 1$ instrumental variables: z_1,..., z_M.

So each IV is uncorrelated with the error term

$$cov(u, z_m) = 0 \quad m = 1, \ldots, M$$

and each IV is partially correlated with x_1 (as explained above, this means that when x_1 is regressed on a constant, x_2, and z_m, the latter has a significant impact).

In principle we could thus calculate M different IV estimates of β_1

This approach has two drawbacks. One is that different IV estimates may be very different in practice, and we would not know which one to select.

2SLS (cntd)

Another drawback is that these "one-instrument" IV estimators are not, generally, asymptotically efficient.

The efficient IV estimator *simultaneously* uses all M instruments.

Since z_1, z_2,..., z_M, and x_2 are each uncorrelated with u, any linear combination of these exogenous variables is also uncorrelated with the error term. Thus any linear combination of the exogenous variables is a potential instrumental variable for x_1.

The best instrumental variable (resulting in the efficient IV estimator) is the linear combination that is most highly correlated with x_1. It turns out that the optimal combination is $\pi_0 + \pi_1 z_1 + \ldots + \pi_M z_M + \pi_{M+1} x_2$, the 'observed' part in the model

$$x_1 = \pi_0 + \pi_1 z_1 + \ldots + \pi_M z_M + \pi_{M+1} x_2 + v \qquad (12)$$

2SLS (cntd)

The variable $\pi_0 + \pi_1 z_1 + \ldots + \pi_M z_M + \pi_{M+1} x_2$ is a valid instrument for x_1 if at least one parameter (among π_1, ..., π_M) is significant. Under this condition the IV is not perfectly correlated with x_2.

After estimating model (12) by OLS, the null

$$\mathcal{H}_0 : \pi_1 = \ldots = \pi_M = 0$$

can be tested against the alternative that at least one of these parameters is different from zero using the usual F test.

The πs are generally unknown (hence the IV is unknown), but a natural idea is to replace them by their OLS estimates, and use the fitted value

$$\hat{x}_1 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \ldots + \hat{\pi}_M z_M + \hat{\pi}_{M+1} x_2$$

as the IV for x_1.

2SLS (cntd)

The IV estimators of β_0, β_1, β_2 solve:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} x_{i1} - \hat{\beta}_2^{IV} x_{i2}) &= 0 \\
\sum_{i=1}^{n} \hat{x}_{i1}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} x_{i1} - \hat{\beta}_2^{IV} x_{i2}) &= 0 \\
\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} x_{i1} - \hat{\beta}_2^{IV} x_{i2}) &= 0.
\end{aligned}$$

This is a set of 3 equations in 3 unknowns $\hat{\beta}_0^{IV}$, $\hat{\beta}_1^{IV}$, $\hat{\beta}_2^{IV}$, and can easily be solved.

2SLS (cntd)

Remark 1: The IV estimator with multiple instruments is also called the *two stage least squares (2SLS) estimator*. The first stage corresponds to OLS regression of x_1 on a constant and all exogenous regressors, the second stage to OLS regression of y_1 on a constant, $\hat{x}_1$ and x_2. This can be understood by rewriting the above equations as

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV}\hat{x}_{i1} - \hat{\beta}_2^{IV}x_{i2}) &= 0 \\
\sum_{i=1}^{n}\hat{x}_{i1}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV}\hat{x}_{i1} - \hat{\beta}_2^{IV}x_{i2}) &= 0 \\
\sum_{i=1}^{n}x_{i2}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV}\hat{x}_{i1} - \hat{\beta}_2^{IV}x_{i2}) &= 0
\end{aligned}$$

(which follows from the first-order conditions associated with OLS estimation of equation (12)).

2SLS (cntd)

The above equations correspond indeed to the first-order conditions of OLS regression of y_1 on a constant, $\hat{x}_1$ and x_2.

Remark 2: The IV estimators can be shown to be consistent and asymptotically normally distributed (variance formulas not given). This result can be used to define t statistics (which, asymptotically, follow t distributions) and test hypotheses in the usual way.

Remark 3: The generalization to the case where there are multiple exogenous regressors in model (7) is direct.

2SLS example

Consider IV estimation of the wage equation

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u \qquad (13)$$

where *exper* (years of experience) and $exper^2$ are thought to be exogenous variables, and *educ* endogenous.

There are two additional exogenous variables (not included in the model): *motheduc* and *fatheduc*. These two variables can be used as IVs for *educ* if in the model

$$\begin{aligned} educ = \pi_0 + \pi_1 exper + \pi_2 exper^2 \\ + \pi_3 motheduc + \pi_4 fatheduc + v \end{aligned} \qquad (14)$$

we have $\pi_3 \neq 0$ and/or $\pi_4 \neq 0$.

2SLS example

Using the sample of 428 married women, we estimate (14) by OLS, and test $H_0 : \pi_3 = 0, \pi_4 = 0$ using the F test. The F statistic equals 55.4, the 5% critical value for the $F_{2,423}$ distribution is 3.0, so we strongly reject the null.

2SLS estimation of (13) gives

$$\widehat{log(wage)} = \underset{(0.4)}{0.048} + \underset{(0.031)}{0.061}\,educ + \underset{(0.013)}{0.044}\,exper - \underset{(0.0004)}{0.0009}\,exper^2.$$

The estimated return to education is 6.1% (compared with an OLS estimate of 10.8%).

Testing for endogeneity

Hausman test

When all regressors are exogenous, both OLS and 2SLS result in consistent estimation of the parameters. However, the 2SLS estimator is less efficient than the OLS estimator (see above, where we compare the asymptotic variances of both estimators in the SLR model (3)).

Thus when the regressor that is suspected of being endogenous is actually exogenous, the use of IV comes at a price: the variance of the IV estimator is (sometimes much) larger than the variance of the OLS estimator.

It is therefore useful to have a test to detect the possible endogeneity of a regressor that shows whether 2SLS is necessary.

To study the test, consider again model (7) with M instruments $z_1,\ldots,z_M$ for x_1 (the test is easily extended to more general settings).

We wish to test $\mathcal{H}_0 : cov(x_1, u) = 0$ against $\mathcal{H}_1 : cov(x_1, u) \neq 0$.

Testing for endogeneity (cntd)

Hausman test

The Hausman test for endogeneity consists in two steps.

First, estimate (12) by OLS and obtain the residuals $\hat{v}$.

Second, regress the following model (model (7) where $\hat{v}$ is added as an additional regressor)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \alpha \hat{v}_i + \text{ error term}$$

It can be shown that testing for endogeneity amounts to testing whether $\alpha = 0$ against $\alpha \neq 0$ using the usual $t-$test.

As an example, let us test for endogeneity of *educ* in (13).

We obtain the residuals $\hat{v}_i$ from the OLS regression of (14) and include these in (13). The estimate of α is $\hat{\alpha} = 0.058$ with a standard error of 0.034. The outcome of the t statistic is 1.67, which is (moderate) evidence in favor of the hypothesis that *educ* is endogenous.

Testing for overidentifying restrictions

Sargan test

If there is just a single instrumental variable, one cannot test the basic requirement that the instrument and error term are uncorrelated.

However, if multiple instrumental variables are available, it is possible to test whether they are uncorrelated with the error term.

To explain the test, consider again model (7) with two instrumental variables for x_1: z_1 and z_2.

The null hypothesis we wish to test is whether the additional variables are indeed exogenous. The null is
$H_0 : E(z_1 u) = E(z_2 u) = 0$.

The test is simple and consists of three steps:

- Estimate model (7) using 2SLS, and obtain the 2SLS residuals $\hat{u}$.
- Regress $\hat{u}$ on a constant and all exogenous variables. Obtain the R-squared from this regression, $R^2_{\hat{u}}$.

Testing for overidentifying restrictions (cntd)

Sargan test

- Under the null, $nR^2_{\hat{u}}$ follows asymptotically a Chi-Square distribution with one degree of freedom: $nR^2_{\hat{u}} \overset{a}{\sim} \chi^2_1$. If $nR^2_{\hat{u}}$ exceeds (say) the 5% critical value for the χ^2_1 distribution, we reject the null and conclude that at least one of the IVs are not exogenous. If we fail to reject the null, we can have some confidence in the set of IVs used.

Remark 1: The test is easily extended to the case where there are M instrumental variables (and an arbitrary number of exogenous regressors in (7)). We then have $nR^2_{\hat{u}} \overset{a}{\sim} \chi^2_{M-1}$.

Remark 2: The test is called the *test of overidentifying restrictions* (proposed by Sargan).

When *motheduc* and *fatheduc* are used as IVs for *educ* in model (13), we have $M = 2$.

Testing for overidentifying restrictions (cntd)

Sargan test

Regressing the 2SLS residuals $\hat{u}$ on a constant, *exper*, $exper^2$, *motheduc* and *fatheduc*, we find $R^2_{\hat{u}} = 0.0009$ and hence $nR^2_{\hat{u}} = 428(0.0009) = 0.3852$, which is much smaller than the 5% critical value of the χ^2_1 distribution (3.84).

The parent's education variables therefore pass the test of overidentifying restrictions.

Adding the husband's education as another IV, we have $M = 3$.

Performing an analogous regression as above, the test statistic $nR^2_{\hat{u}}$ equals 1.11. The 5% critical value of the χ^2_2 distribution is 5.99, so it is apparently a good decision to add the husband's education to the list of IVs.

The 2SLS estimate on *educ* using the three IVs (resp. the two IVs) is 0.08 (resp. 0.061) with a standard error of 0.022 (resp. 0.031). So adding an additional instrument renders *educ* much more significant.

I understand/can apply...

- IV estimator
  - special case: Wald estimator
- Exclusion restriction
- Instrument relevance
- 2SLS estimator
- Hausman & Sargan Test