Instrumental Variables Estimators (IV) in Simple Model

Consider the following simple model:

$$y_i = \beta_1 + \beta_2 x_{2,i} + u_i \tag{1}$$

and suppose that $Cov(x_{2,i} u_i) \neq 0$. Now, suppose there is a random variable z_i, such that

$$Cov(z_i, u_i) = 0 \tag{2}$$

and

$$Cov(z_i, x_{2,i}) \neq 0 \tag{3}$$

Then, z_i is called an *instrumental variable for* $x_{2,i}$. Thus, an instrument for $x_{2,i}$ is a random variable which is correlated with $x_{2,i}$ but is uncorrelated with the error.

The first requirement $Cov(z_i, u_i) = 0$ cannot be tested, as u_i is not observable. The second requirement, non zero correlation between $x_{2,i}$ and z_i can instead be tested. Consider the auxiliary model,

$$x_{2,i} = \pi_1 + \pi_2 z_{2,i} + \epsilon_i, \tag{4}$$

we can then test the null hypothesis $H_0 : \pi_2 = 0$ vs $H_1 : \pi_2 \neq 0$. If we do not reject the null, then z_i is uncorrelated with $x_{2,i}$ and so it CANNOT be used as an instrument. If we reject the null, then we know that z_i is correlated with $x_{2,i}$.

Write model (1) in terms of deviations from mean, i.e.:

$$y_i - \mu_y = \beta_2 \left(x_{2,i} - \mu_{x_2} \right) + u_i. \tag{5}$$

where $\mu_y = E(y_i)$ and $\mu_{x_2} = E(x_{2,i})$. Now, multiply both sides by $(z_i - \mu_z)$,

$$\begin{aligned} & (z_i - \mu_z) \left(y_i - \mu_y \right) \\ = \quad & \beta_2 \left(x_{2,i} - \mu_{x_2} \right) (z_i - \mu_z) + u_i (z_i - \mu_z) \end{aligned}$$

and take the expectation (mean) of both sides, recalling that $E\left((z_i - \mu_z) \left(y_i - \mu_y \right) \right) = Cov(z, y)$,

$$Cov(z, y) = \beta_2 Cov(z, x_2) + Cov(u, z).$$

Thus, if $Cov(u, z) = 0$ and $Cov(z, x) \neq 0$, i.e. if z_i is a valid instrument,

$$\beta_2 = \frac{Cov(z, y)}{Cov(z, x_2)}$$

We now define the instrumental variable IV estimator $\widehat{\beta}_{2,IV}$ as the sample analog of the right hand side above, that is:

$$\begin{aligned} \widehat{\beta}_{2,IV} &= \frac{\widehat{Cov(z, y)}}{\widehat{Cov(z, x_2)}} \\ &= \frac{\sum_{i=1}^{n} \left((z_i - \widehat{\mu}_z) \left(y_i - \widehat{\mu}_y \right) \right)}{\sum_{i=1}^{n} \left((z_i - \widehat{\mu}_z) \left(x_{2,i} - \widehat{\mu}_{x_2} \right) \right)}. \end{aligned}$$

Thus, if z_i satisfies conditions (2) and (3), then

$$\begin{aligned} p\lim \widehat{\beta}_{2,IV} &= \frac{p\lim n^{-1}\sum_{i=1}^{n}\left((z_i-\widehat{\mu}_z)\left(y_i-\widehat{\mu}_y\right)\right)}{p\lim n^{-1}\sum_{i=1}^{n}\left((z_i-\widehat{\mu}_z)\left(x_{2,i}-\widehat{\mu}_{x_2}\right)\right)} \\ &= \frac{Cov(z,y)}{Cov(z,x_2)} = \beta_2, \end{aligned}$$

where $\widehat{\mu}_z = n^{-1}\sum_{i=1}^{n} z_i$, $\widehat{\mu}_y = n^{-1}\sum_{i=1}^{n} y_i$ and $\widehat{\mu}_{x_2} = n^{-1}\sum_{i=1}^{n} x_{2,i}$. Thus, $\widehat{\beta}_{2,IV}$ is consistent for β_2.

Incidentally, in the case of error uncorrelated with regressors, by multypling both sides of (5) by $\left(x_{2,i}-\mu_{x_2}\right)$ and take the expectation, we have (as well known...)

$$\beta_2 = \frac{Cov(x_2,y)}{Var(x_2)}$$

and so

$$\beta_2 = \frac{Cov(x_2,y)}{Var(x_2)} = \frac{Cov(z,y)}{Cov(z,x_2)}.$$

Thus, it follows that OLS, when error are uncorrelated with regressors, and IV converge towards the same probability limit β_2. We see later that in this case is better to use OLS instead of IV, as it is more efficient. In fact,

$$\begin{aligned} \widehat{\beta}_{2,IV} &= \frac{\sum_{i=1}^{n}\left((z_i-\widehat{\mu}_z)\left(y_i-\widehat{\mu}_y\right)\right)}{\sum_{i=1}^{n}\left((z_i-\widehat{\mu}_z)\left(x_{2,i}-\widehat{\mu}_{x_2}\right)\right)} \\ &= \frac{\sum_{i=1}^{n}(z_i-\widehat{\mu}_z)\left(\beta_2\left(x_{2,i}-\widehat{\mu}_{x_2}\right)+u_i\right)}{\sum_{i=1}^{n}\left((z_i-\widehat{\mu}_z)\left(x_{2,i}-\widehat{\mu}_{x_2}\right)\right)} \\ &= \beta_2 + \frac{\sum_{i=1}^{n}(z_i-\widehat{\mu}_z)u_i}{\sum_{i=1}^{n}\left((z_i-\widehat{\mu}_z)\left(x_{2,i}-\widehat{\mu}_{x_2}\right)\right)} \end{aligned}$$

Thus,

$$\begin{aligned} & n^{1/2}\left(\widehat{\beta}_{2,IV}-\beta_2\right) \\ =\ & \frac{n^{-1/2}\sum_{i=1}^{n}(z_i-\widehat{\mu}_z)u_i}{n^{-1}\sum_{i=1}^{n}\left((z_i-\widehat{\mu}_z)\left(x_{2,i}-\widehat{\mu}_{x_2}\right)\right)} \\ \simeq\ & \frac{n^{-1/2}\sum_{i=1}^{n}(z_i-\widehat{\mu}_z)u_i}{Cov(z,x_2)}, \end{aligned}$$

given that $p\lim n^{-1}\sum_{i=1}^{n}\left((z_i-\widehat{\mu}_z)\left(x_{2,i}-\widehat{\mu}_{x_2}\right)\right) = Cov(z,x_2)$. Now, in the IV setting, the assumption of conditional homoskedasticity writes as $E(u_i^2|z_i) = \sigma_u^2$,

$$\begin{aligned} & Var\left(n^{-1/2}\sum_{i=1}^{n}(z_i-\widehat{\mu}_z)u_i\right) \\ =\ & n^{-1}\sum_{i=1}^{n} E\left((z_i-\widehat{\mu}_z)^2 u_i^2\right) \\ =\ & E\left((z_i-\widehat{\mu}_z)^2 E(u_i^2|z_i)\right) = Var(z)\sigma_u^2 \end{aligned}$$

Thus,

$$avar\left(n^{1/2}\left(\widehat{\beta}_{2,IV}-\beta_2\right)\right)=\frac{Var(z)\sigma_u^2}{Cov(z,x_2)^2}$$

First note that the avar of IV estimators crucially depends on the correlation between $x_{2,i}$ and $z_{,i}$, thus if the correlation is low (weak instrument problem), the IV avar can be very large. As a consequence also IV standard error are very large, and so inference based on IV can be not very reliable.

Compare the avar of IV with

$$avar\left(n^{1/2}\left(\widehat{\beta}_{2,ols}-\beta_2\right)\right)=\frac{\sigma_u^2}{Var(x_2)}$$

Now,

$$\begin{aligned} & \frac{avar\left(n^{1/2}\left(\widehat{\beta}_{2,IV}-\beta_2\right)\right)}{avar\left(n^{1/2}\left(\widehat{\beta}_{2,ols}-\beta_2\right)\right)} \\ = & \frac{Var(z)Var(x_2)}{Cov(z,x_2)^2}=\frac{1}{\rho^2_{x_2,z}}\geq 1 \end{aligned}$$

As $-1\leq\rho_{x_2,z}\leq 1$, being $\rho_{x_2,z}$ the coefficient of correlation between x_2 and z.

Do not worry, if $E(u_i^2|z_i)=h(z_i)$ can use White SE also for IV!

Though, in the presence of conditional heteroskedasticity, OLS are not necesarily more efficient than IV.

Example

We want to study the return on education for married women. The issue is that wages depend on the individual ability, which is not observed. Thus, we have an omitted variable issue, and as ability is likely to be correlated with education, chances are that OLS estimator are inconsistent. Need to find a good instrument.

Model:

$$\log(wage_i)=\beta_1+\beta_2 educ_i+u_i \tag{6}$$

First consider OLS estimation:

$$\log\widehat{(wage)}_i=-.185(.185)+.109educ_i$$

so that according to the model estimated with OLS, an extra year of education leads to 10.9% more in wages. As instrument, we try the father's education, $fatheduc$; the idea is that an educated father is more likely to have an educated dauther; though the education of the father is in general uncorrelated with the innate daughter ability. Thus, $fatheduc$ is a good candidate for instrument. To check wether $fatheduc$ is correlated with $educ$, we run a OLS regression of $educ$ on $fatheduc$,

$$\widehat{educ}_i=10.24(.28)+.269(.029)fatheduc_i$$

Thus we can safely reject the hypothesis that $fatheduc$ is uncorrelated with $educ$. Now we estimate the wage equation for married women using IV and using $fatheduc$ as instrument. We have,

$$\log(\widehat{wage_i}) = -.441(.446) + .059(.035)educ_i$$

We note that the IV estimator for the coefficient on education is 0.059 while the OLS estimator is 0.109...the two are quite far away apart. This is a clear signal that the error in model (6) were correlated with the error. Here the issue is the omission of the variable $abil_i$, whose coefficient is positive. Also, as ability and education are positively correlated, we expect the OLS estimator to be upward biased.

Though, not that the SE on the IV estimator is much bigger than the SE of OLS...To really see whether IV and OLS estimators converge to different plim need a formal test. Be patient!

When the correlation between z and $x_{2,i}$ is low, we say that z_i is a weak instrument. We have see that

$$\widehat{\beta}_{2,IV} - \beta_2 = \frac{n^{-1}\sum_{i=1}^{n}(z_i - \widehat{\mu}_z)u_i}{n^{-1}\sum_{i=1}^{n}\left((z_i - \widehat{\mu}_z)\left(x_{2,i} - \widehat{\mu}_{x_2}\right)\right)}$$

Thus, even if for given sample size n the numerator is very tiny, if the instrument is weak the denominator is very tiny too. As a result $\widehat{\beta}_{2,IV} - \beta_2$ can be quite far away from zero even for relatively large samples. In this cases, it is not clear whether is better to use OLS or IV with weak instruments.

Example

We want to analyze the effect of mother's smoking on birth weight. Model:

$$\log(bweight)_i = \beta_1 + \beta_2 packs_i + u_i$$

where $packs$ indicates the number of packs smoked by the mother during pregnancy. We may suspect that $packs_i$ is correlated with other omitted variables, related to health consciousness of mother. The average price of cigarettes per state of residence is likely to be uncorrelated with the error. Economic theory suggest a negative relationship between quantity demanded and price, so we try price of cigarettes as an instrument for pack. We have:

$$\widehat{pack}_i = .067(.103) + .0003(.0008)cigprice_i$$

where $cigprice_i$ is the average price per packet in the state in which mother i is resident. It seems clear that packests and average price are uncorrelated, thus $cigprice$ is what we call a weak instrument. Let's see what happens if we use it as instrument for packet.

$$\log(bweight)_i = 4.45(.91) + 2.99(8.7)cigprice_i$$

nonsense!!!! wrong sign, huge estimated coefficient, and super huge SE!!!

Instrumental Variables Estimators (IV) in Multiple Regression Model

Variables correlated with the error are called *endogeneous,* while variables uncorrelated with the error are called *exogeneous.* Consider the following model:

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + u_i \tag{7}$$

and $Cov(u, x_2) \neq 0$, while $Cov(u, x_3) = 0$, so $x_{2,i}$ is endogeneous, while $x_{3,i}$ is exogeneous. The idea is that both y_i and $x_{2,i}$ are jointly determined via a two-dimensional structural model (think at price and quantities, for example), but here we are interested in estimating only the equation with y_i as dependent variable.

In terms of the familiar wage equation,

$$\log(wage)_i = \beta_1 + \beta_2 educ_i + \beta_3 \exp er_i + u_i$$

we may think at $educ$ as an endogeneous variable (correlated with the error) and at $\exp er$ at an exogeneous variable, uncorrelated with the error.

If we estimate (7) by OLS, all estimators will be inconsistent, not only that for the coefficient of the endogeneous variable.

We proceed in the following manner. We choose an instrument for $x_{2,i}$ say $z_{2,i}$ such that $Cov(z_2, y_2) \neq 0$ and $Cov(z_2, u) = 0$, while as an instrument of $x_{3,i}$ we use itself (if $x_{3,i}$ is uncorrelated with the error the instrument having the highest correlation with $x_{3,i}$ is clearly $z_{3,i}$ itself). Define:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{2,1} & x_{3,1} \\ 1 & x_{2,2} & x_{3,2} \\ 1 & x_{2,3} & x_{3,3} \\ . & . & . \\ 1 & x_{2,n} & x_{3,n} \end{pmatrix}$$

$$Z = \begin{pmatrix} 1 & z_{2,1} & x_{3,1} \\ 1 & z_{2,2} & x_{3,2} \\ 1 & z_{2,3} & x_{3,3} \\ . & . & . \\ 1 & z_{2,n} & x_{3,n} \end{pmatrix}$$

and (as usual)

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ y_n \end{pmatrix}$$

Define:

$$\widehat{\boldsymbol{\beta}}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\,\mathbf{Z}'\mathbf{y}$$

where in the present example $\mathbf{Z}$ in $n \times 3$, $\mathbf{X}$ is $n \times 3$, and $\mathbf{y}$ is $n \times 1$.

Note that,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{IV} &= \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right)\\ &= \boldsymbol{\beta}+\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{u}\end{aligned}$$

If z_2 is a valid instrument,

$$\begin{aligned}p\lim\widehat{\boldsymbol{\beta}}_{IV} &= \boldsymbol{\beta}+\left(\mathbf{p}\lim\left(\mathbf{Z}'\mathbf{X}\right)\right)^{-1}p\lim\left(\mathbf{Z}'\mathbf{u}\right)\\ &= \boldsymbol{\beta}\end{aligned}$$

Example

We have a sample of 3010 working men, and we want to explain their wage using education ($educ$), experience, ($\exp er$), the squared of experience ($\exp er^2$), plus a dummy equal to 1 if they are black ($black$), a dummy equal to 1 if they live in a metrpolitan area ($SMSA$), a dummy equal to 1 if they live in the south ($south$). As usual, the issue is that education is correlated with the error. Card (1995) suggest to use proximity to 4yrs college as instrument for education. Let $nearc4$ be a dummy variable equal to 1 if the individual grew up within a given distance from college. We expect $nearc4$ to be positively correlated to $educ$ (the student can go to college and live at home, saving on accomodation), but clearly uncorrelated with the omitted variable ability, and so uncorrelated with the error. We now regress $educ$ on $nearc4$, and all other exogeneous variables. We get

$$\widehat{educ_i} = 16.6(.24) + .32(.088)nearc4_i - .413(.034)\exp er_i + ...$$

Thus, $nearc4_i$ is correlated with $educ_i$ and thus is a valid instrument for education to use.

OLS estimates:

$$\begin{aligned}&\log(\widehat{wage_i})\\ =\ &const + .075(.003)educ_i + .085(.007)\exp er_i - 0.023(.0003)\exp er_i^2\\ &-.199(.018)black_i + .136(.020)SMSA_i - .148(.026)south_i\end{aligned}$$

IV estimates:

$$\begin{aligned}&\log(\widehat{wage_i})\\ =\ &const + .132(.055)educ_i + .108(.024)\exp er_i - 0.023(.0003)\exp er_i^2\\ &-.147(.054)black_i + .112(.032)SMSA_i - .145(.027)south_i\end{aligned}$$

We note that:

(i) Not only the coefficient on the endogeneous variable $educ$ differ in the OLS and IV case, but also all other coefficients, but for $\exp er^2$ and $south$.

(ii) Contrary to the married women case and father education instrument, now the IV estimated coefficient on education is larger than the OLS. (this time we have lots of variables, so even if the coeff on ability is positive and ability is positively correlated with education, this does not imply an upward bias of the OLS coeff, too many variables and correlation at play)

(iii) IV standard error are much much larger than OLS ones (often 5-6 times OLS), and we have 3000 observations!!!

Two Stage Least Squares Estimators

We have seen that several iunstruments can be used for the the variable education in the wage equation. So far we have seen father education and proximity to college, but many others have been used in the vast empirical literature on wage equations: e.g. number of siblings (the more the children the less the education), dummy for being born in first quarter (reach earlier the age of compulsory age for schooling, in general 16, with less time at school), etc.

Thus, a question arises: why do not use more than one instrument for a given endogenous regressors. Intuitively, if instruments are valid, the more we use the better (at least in terms of efficiency).

Consider again,

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + u_i$$

where $x_{3,i}$ is exogeneous and $x_{2,i}$ is endogeneous. Suppose, that we have two valid instruments for $x_{2,i}$, namely $z_{2,i}$ and $z_{3,i}$, so that $Cov(x_2, z_2) \neq 0$, $Cov(x_2, z_3) \neq 0$, and $Cov(u, z_2) = Cov(u, z_3) = 0$. Now, we regress $x_{2,i}$ on constant, $z_{2,i}$ and $z_{3,i}$ and we obtain the predicted value $\widehat{x}_{2,i}$ (note that at least one of the coefficients in the previous regression has to be significatively different from zero). Now, we run the OLS regression

$$y_i = y_i = \beta_1 + \beta_2 \widehat{x}_{2,i} + \beta_3 x_{3,i} + u_i.$$

The resulting estimator is called *two-step least square estimators TSLS* and denoted $\widehat{\beta}_{2sls}$.

In general in one equation we may have more than one regressor which is correlated with the error. Consider the model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

and you suspect that at least one regressors, but possibly more than one, is correlated with the error. Suppose $\mathbf{X}$ is $n \times k$ and $\mathbf{Z}$ is $n \times p$, $p \geq k$ ($\mathbf{Z}$ in another vector of variables that we call instruments). The condition $p \geq k$, i.e. at least as many instruments as regressors, is called *order condition.* Of course, as instrument for intercept we use a vector of ones and for exogeneous regressors we use as instrument the regressor itself. We proceed as follows:

STEP 1: Regress $\mathbf{X}$ on $\mathbf{Z}$, i.e. we estimate the following model,

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Pi} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is $n \times k$, $\mathbf{Z}$ is $n \times P$, $\boldsymbol{\Pi}$ is $p \times k$ and $\boldsymbol{\epsilon}$ is $n \times k$. When $k > 1$, this is slightly different from the usual regression in which the dependent variable is a scalar, but do not worry just proceed in the usual way. The OLS estimator for $\boldsymbol{\Pi}$, call it $\widehat{\boldsymbol{\Pi}}$, is

$$\widehat{\boldsymbol{\Pi}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}, \tag{8}$$

and note that $\widehat{\boldsymbol{\Pi}}$ is $p \times k$. Equation (8) is known as *reduced form equation.* Now we are ready for the second step.

STEP 2: we estimate the following model,

$$\mathbf{y} = (\mathbf{Z}\widehat{\mathbf{\Pi}})\boldsymbol{\beta} + \mathbf{v},$$

note that in the second stage we regress $\mathbf{y}$ on $\mathbf{Z}\widehat{\mathbf{\Pi}}$, which is the projection of $\mathbf{X}$ on $\mathbf{Z}$. The OLS estimator for $\boldsymbol{\beta}$ from the second stage is called 2SLS estimator (which is an IV estimator). Note that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{2sls} &= (\widehat{\mathbf{\Pi}}'\mathbf{Z}'\mathbf{Z}\widehat{\mathbf{\Pi}})^{-1}\widehat{\mathbf{\Pi}}'\mathbf{Z}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}
\end{aligned}$$

In the context of 2SLS, the equivalent condition of correlation between instruments and regressor is:

$$p\lim\left(\mathbf{X}'\mathbf{Z}/\mathbf{n}\right) = \mathbf{Q}_{zx}$$

where $\mathbf{Q}_{xz}$, which is is a $k \times p$ matrix, has rank k. This condition is known as *rank conditions.*

Under the following assumptions:

IV1: linearity

IV2: $(y_i, X_i, Z_i)'$ are identically and independently distributed (iid), with y scalar and X_i $1 \times k$, Z_i $1 \times p$, $p \geq k$.

IV3: $\mathbf{E}(\mathbf{X}'\mathbf{X}/\mathbf{n}) = \mathbf{Q}_{XX}$, which is positive definite

IV4: $\mathbf{E}(\mathbf{Z}'\mathbf{X}/\mathbf{n}) = \mathbf{Q}_{ZX}$, which is of rank k.

IV5: $\mathbf{E}(\mathbf{Z}'\boldsymbol{\epsilon}/\mathbf{n}) = \mathbf{0}$

IV6: $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{Z}) = \sigma^2\mathbf{I}_n$

$$p\lim\widehat{\boldsymbol{\beta}}_{2sls} = \boldsymbol{\beta}$$

and

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_{2sls} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \sigma^2(\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx})^{-1})$$

and

$$\left(s^{-2}\mathbf{X}'\mathbf{Z}/\mathbf{n}(\mathbf{Z}'\mathbf{Z}/\mathbf{n})^{-1}\mathbf{Z}'\mathbf{X}/\mathbf{n}\right)^{1/2} n^{1/2}(\mathbf{b}_{iv} - \boldsymbol{\beta}) \xrightarrow{d} N(0, I_k)$$

with $s^2 = \mathbf{ee}/n - k$.

We have so far considered the general case in which $p \geq k$, in particular when $p > k$ (more instruments than random variables) we say that the model is overidentified, while when $p = k$ (as many instruments as regressors) then we say that the model is exactly identified. The 2SLS estimator simplifies in the the exact identification case, i.e. when $p = k$ to

$$\widehat{\boldsymbol{\beta}}_{2sls2sls} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

Choosing the Instruments

We have seen that instruments should satisfy two properties (i) they have to be highly correlated with the original regressors (ii) they have to be uncorrelated with the errors.
Another issue is: how many instrument should we use? In principle, as many as possible, as by increasing the number of instruments we make the IV estimator more efficient (i.e. having smaller variance). Why? Remember the logic of $2SLS$. Suppose we start with the linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$. In the first step we regress $\mathbf{X}$ on $\mathbf{Z}$ and the resulting coefficient is $\widehat{\mathbf{\Pi}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. In the second step we regress $\mathbf{y}$ on $\mathbf{Z}\widehat{\mathbf{\Pi}}$, i.e. we estimate by OLS the following model,

$$\mathbf{y} = \mathbf{Z}\widehat{\mathbf{\Pi}}\boldsymbol{\beta} + (\mathbf{X} - \mathbf{Z}\widehat{\mathbf{\Pi}})\boldsymbol{\beta} + \mathbf{u} = \mathbf{Z}\widehat{\mathbf{\Pi}}\boldsymbol{\beta} + \mathbf{v},$$

where $\mathbf{v} =(\mathbf{X} - \mathbf{Z}\widehat{\mathbf{\Pi}})\boldsymbol{\beta} + \mathbf{u}$. Now,

$$\widehat{\boldsymbol{\beta}}_{2sls} = (\widehat{\mathbf{\Pi}}'\mathbf{Z}'\mathbf{Z}\widehat{\mathbf{\Pi}})^{-1}\widehat{\mathbf{\Pi}}'\mathbf{Z}'\mathbf{y}$$

We now want to see two very important facts: (a) The variance of the $2SLS$ estimator decreases as the number of instruments increases, (b) if the error are indeed uncorrelated with the residuals the OLS estimator is more efficient than the $2SLS$ estimator, in the sense of having a smaller variance. About fact (a) is easy to see:

$$E(\mathbf{v}'\mathbf{v}) =E(\mathbf{u}'\mathbf{u})+E(\boldsymbol{\beta}'(\mathbf{X} - \mathbf{Z}\widehat{\mathbf{\Pi}})\prime(\mathbf{X} - \mathbf{Z}\widehat{\mathbf{\Pi}})\boldsymbol{\beta}),$$

as the covariance term is zero. Now the second term can only decreases as the number of instruments increases, in fact the higher is the number of instruments the smaller is the square error from the first stage. As for fact (b), we have shown it for the simple model and only one instrument. Though, it is possible to show that, under conditional homoskedasticity,

$$avar\left(n^{1/2}(\widehat{\boldsymbol{\beta}}_{2sls} - \boldsymbol{\beta})\right) - avar\left(n^{1/2}(\widehat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})\right)$$

is a positive definte matrix. Thus, OLS is more efficient than 2SLS.

Thus given fact (a), we want to use as many instruments as we can, provided they are all uncorrelated with the error, given fact (b), if the regressor are uncorrelated with the error we want to use OLS and not $2SLS$. The first issue is: how many instrument should we use? In principle, as many as possible, as by increasing the number of instruments we make the 2SLS estimator more efficient (i.e. having smaller variance). However, we should be sure that all the instruments we are using satisfying the two properties above.

If $k = p$, i.e. we have as many instruments as original regressors, than WE CANNOT estimate whether the instruments are uncorrelated with the errors or not. Why? When $p = k$,

$$\mathbf{b}_{iv}= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

Now the IV residuals is

$$\mathbf{e}_{iv}= \mathbf{y} - \mathbf{X}\mathbf{b}_{iv}= \mathbf{y} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

Thus

$$\begin{aligned}\mathbf{Z}'\mathbf{e}_{iv} &= \mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \\ &= \mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{y} = \mathbf{0}\end{aligned}$$

Thus by construction the the IV residuals are orthogonal of the instruments whenever $k = p$. How can we proceed???

We can start running IV using as many instruments as the number of original regressors. Let $\mathbf{e}_{iv}$ be defined as above, we can then regress $\mathbf{e}_{iv}$ on $\mathbf{Z}$ (where $\mathbf{Z}$ is $n \times k$) and on additional instruments say $\mathbf{Z}_2$, where $\mathbf{Z}_2$ is $n \times q$. Take the R^2 from the regression of $\mathbf{e}_{iv}$ on $\mathbf{Z}$ and $\mathbf{Z}_2$. Under the null hypothesis that additional instrument $\mathbf{Z}_2$ are not correlated with the error, $nR^2 \overset{d}{\to} \chi^2_q$. Thus, if we get a value below the 95% percentile of a chi-square with q degree of freedom we decide that the additional instruments are not correlated with the errors, otherwise we decide to reject the null and conclude that (some of) the additional instruments is correlated with the errors. IMPORTANT: if we do not reject the null that the additional instruments are uncorrelated with the errors, it may still be possible that the original k instruments were indeed correlated with the errors. Thus in practice, it may be a good idea run this test rotating the initial and additional instruments.

The Hausman-Wu Test

We have seen that error in variables, omitted variables, endogeneity induce correlation between errors and regressors.

We have seen that if regressors and error are correlated, then the OLS estimator is no longer consistent for the parameter of interest. However, if we can find proper instruments (i.e. the instruments are correlated with the regresors and uncorrelated with the errors), then IV (instrumental variable) estimators are consistent and asymptotically normal. Thus, whenever we are uncertain about whether errors and regressors are correlated, we may think it is always a good idea to use IV. This is not the case, as, if $E(\mathbf{X}'_i\epsilon_i) = 0$, then OLS are more efficient than IV estimators, in the sense of having a smaller variance and so they are more precise. Thus we are interested in testing

$$H_0 : E(\mathbf{X}'_i\epsilon_i) = 0 \text{ versus } H_A : E(\mathbf{X}'_i\epsilon_i) \neq 0$$

WARNING: do not even think to construct a statistic based on $n^{-1}\sum \mathbf{X}'_i e_i$, where e_i are OLS residuals, as this is NUMERICALLY EQUAL to ZERO. We first begin with the easier case in which few of the regressor are correlated with the errors, but the remaining are not correlated. We then proceed to the more complex case in which all regressors can be potentially correlated with the errors. Suppose that we suspect that only a subset of the $\mathbf{X}$ are correlated with the errors, say we suspect that $E(X_{ij}\epsilon_i) \neq 0$, for say $j = k_{1+1}, \ldots k$. Often, common sense and economic theory allow us to decide a priori whether certain regressors are not correlated with the errors, while other may be instead correlated. If we "know" that $E(X_{ij}\epsilon_i) = 0$ for $i = 1, ..., k_1$ we can use $X_1, ..., X_{k_1}$ as instruments for themselves (these of course are the best instruments!). On the other hand we need to choose r, $r \geq (k - k_1)$ instruments for $X_{k_{1+1}}, ..., X_k$, call these instruments $Z_1, ..., Z_r$. Summarizing the vector of instruments is given by

$$\mathbf{Z}_i = (1, X_{2i}, \ldots X_{i,k1}, Z_{i1}, \ldots Z_{ir}),\ p = k_1 + r > k$$

i.e. $r > k - k_1$. (Note that X_{1i} is the intercept). The test described below is know as Hausman-Wu test.

For notational simplicity let $\mathbf{X}_i^{*}= (X_{i,k1+1}, \ldots X_{ik})'$. We regress $\mathbf{X}^*$ on $\mathbf{Z}$ and the fitted value from such regression are then,

$$\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}^*$$

Then we regress $\mathbf{y}$ on $\mathbf{X}$ and on $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}^*$, let $\widehat{\boldsymbol{\delta}}$ be the OLS coefficient on $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}^*$, we can then perform a F test for the null that $\boldsymbol{\delta} = \mathbf{0}$ versus the alternative $\boldsymbol{\delta} \neq \mathbf{0}$. Under the null, $(k - k_1)F$ is asymptotically distributed as a $\chi^2_{k-k_1}$, while under the alternative diverges to plus infinity. If we reject (do not reject) the null of $\boldsymbol{\delta} = \mathbf{0}$, then we reject (do not reject) the null that $E(\mathbf{X}'\boldsymbol{\epsilon}) = \mathbf{0}$.

Testing the effect of Education on Earnings

The study on the effect of education on earning has received lot of attention over the last years. Stylized fact: the wage gap between educated and uneducated or skilled and unskillied workers has been increased a lot in the 90s (while for example the wage gap between genders has decreased). Thus it seems that education has high impact on earnings. Also, for policy purposes (fees, students loan, investment in education), it is important to know the returns of education. A traditional model used in labour economics to study education is the following:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 age + \beta_3 age^2 + \beta_4 education \\ &\quad + \beta_5 X_i + error \end{aligned}$$

where X_i includes set of dummy, like gender, industry, geographical area etc. Typically education is measured as years of schooling. It is not difficult to image the years of schooling measures education only up to an error. Also, it is possible that people who got a higher education is "smarter". Second, and more surprinsingly, it seems that the reported years of schooling is also subjected to a serious measurement errors (incorrect self report years). All these reason raises a serious doubt about the fact that errors and regressor (schooling) are uncorrelated. Ashenfelter and Krueger (1994) had the idea of using data on couple of twins. First twins tend to have same characteristic (same bckground, same original income etc..) so that we can sort of control for omitting social economic variables. Second, they use the sibling's report of years of schooling of the other sibling as an instrument for the latter. Let y_{ij} be the earning of sibling j, $j = 1, 2$, in twins i and $S_{ij}(j)$ the self report of yers of schooling of sibling j in twins i. (The data set consists on observations of several couples of twins). **OLS**

$$\begin{aligned} y_{ij} &= \beta_1 + 0.088 age_i - 0.087 age_i^2 + 0.084 S_{ji}(j) \\ &\quad +0.2 sex - 0.41 race + residual \end{aligned}$$

IV (as instrument for years of schooling of one sibling we use the years of schooling for that schooling reported by the other sibling).

$$\begin{aligned} y_{ij} &= \beta_1 + 0.088 age_i - 0.087 age_i^2 + 0.116 S_{ji}(j) \\ &\quad +0.21 sex - 0.42 race + residual \end{aligned}$$

It is immediate to note, that the coefficient on schooling from the IV estimator is somewhat different from the OLS. Next time we shall study a formal test for comparing IV and OLS estimator.

Hausman Test

We have considered last time the Hausman-Wu test for the null hypothesis that some (but not all!) of the regressors are uncorrelated with the errors, against the alternative that they are instead correlated. However, the Hausman-Wu test is valid under the maintained hypothesis that some of the regressors are uncorrelated with the errors. As you see in the homework, we CANNOT use the Hausman-Wu for testing the null that ALL regressor are uncorrelated with the errors versus the alternative that at least one regressor is instead correlated with the error. We now want to construct a test for,

$$H_0 : E(\mathbf{X}_i'\epsilon_i) = 0 \text{ versus } H_A : E(\mathbf{X}_i'\epsilon_i) \neq 0$$

The idea underlying the Hausman test is the following: under the null hypothesis, both the OLS and the IV (or 2SLS) estimators are consistent for the true parameter, thus we expect that as the sample gets large the OLS and the IV estimators will be close each other. On the other hand, under the alternative, the OLS estimator is not consistent for the true parameters, while the IV (or 2SLS) it is. Therefore, we should expect that under the alternative, the two are far away each other. The key ingredient for the Hausman test statistic is then $(\mathbf{b}_{iv} - \mathbf{b}_{ols})$. Now,

$$(\mathbf{b}_{iv} - \mathbf{b}_{ols}) = (\mathbf{b}_{iv} - \boldsymbol{\beta}) - (\mathbf{b}_{ols} - \boldsymbol{\beta})$$

and we know that under the null,

$$\begin{aligned} p \lim_{n\to\infty} (\mathbf{b}_{iv} - \mathbf{b}_{ols}) &= p \lim_{n\to\infty} (\mathbf{b}_{iv} - \boldsymbol{\beta}) - \\ p \lim_{n\to\infty} (\mathbf{b}_{ols} - \boldsymbol{\beta}) &= 0 - 0 = 0 \end{aligned}$$

on the other hand under the alternative

$$\begin{aligned} p \lim_{n\to\infty} (\mathbf{b}_{iv} - \mathbf{b}_{ols}) &= p \lim_{n\to\infty} (\mathbf{b}_{iv} - \boldsymbol{\beta}) - p \lim_{n\to\infty} (\mathbf{b}_{ols} \\ - \boldsymbol{\beta}) &= 0 + \text{ something different from } 0 \end{aligned}$$

Thus we can base our statistic on $(\mathbf{b}_{iv} - \mathbf{b}_{ols})$. However, if we want to have a test with a given type I error and a type II error approaching zero as the sample size gets large, we need to construct a statistic which has a well defined limiting distribution (under the null), and diverges under the alternative. In order to obtain that, we need to scale $(\mathbf{b}_{iv} - \mathbf{b}_{ols})$ by $n^{1/2}$ and divide it by the proper variance. Let's consider,

$$(Var(n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols}))^{-1/2} n^{1/2}((\mathbf{b}_{iv} - \mathbf{b}_{ols})$$

Under the set of $L1 - L6$ and $IV1 - IV7$ assumptions,

$$\begin{aligned}&(Var(n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols}))^{-1/2} n^{1/2}((\mathbf{b}_{iv} - \mathbf{b}_{ols})\\ &\xrightarrow{d} N(0, I_k)\end{aligned}$$

Thus we need to find the exact expression for $Var(n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols}))$.

$$Var(n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols})) = Var(n^{1/2}((\mathbf{b}_{iv} - \boldsymbol{\beta}) - (\mathbf{b}_{ols} - \boldsymbol{\beta})))$$

$$\begin{aligned} = \;& Var(n^{1/2}(\mathbf{b}_{iv} - \boldsymbol{\beta})) + Var(n^{1/2}(\mathbf{b}_{ols} - \boldsymbol{\beta}))\\ & -2Cov(n^{1/2}(\mathbf{b}_{iv} - \boldsymbol{\beta})), n^{1/2}(\mathbf{b}_{ols} - \boldsymbol{\beta}))\end{aligned}$$

We have seen that

$$\lim_{n\to\infty} Var(n^{1/2}(\mathbf{b}_{iv} - \boldsymbol{\beta})) = \boldsymbol{\sigma}^2(\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX})^{-1}$$

$$\lim_{n\to\infty} Var(n^{1/2}(\mathbf{b}_{ols} - \boldsymbol{\beta})) = \boldsymbol{\sigma}^2\mathbf{Q}_{XX}^{-1}$$

where $\mathbf{Q}_{XX} = p\lim_{n\to\infty}\frac{\mathbf{X}'\mathbf{X}}{n}$, $\mathbf{Q}_{ZZ} = p\lim_{n\to\infty}\frac{\mathbf{Z}'\mathbf{Z}}{n}$, $\mathbf{Q}_{XZ} = p\lim_{n\to\infty}\frac{\mathbf{X}'\mathbf{Z}}{n}$. Now, the difficult part is to compute the covariance between the two estimators (don't even hope it is zero!).

$$\lim_{n\to\infty} Cov(n^{1/2}(\mathbf{b}_{iv} - \boldsymbol{\beta})), n^{1/2}(\mathbf{b}_{ols} - \boldsymbol{\beta}))$$

$$= (\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX})^{-1}\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}E(\epsilon_i^2\mathbf{Z}_i'\mathbf{X}_i)\mathbf{Q}_{XX}^{-1}$$

$$\begin{aligned} =\;& \sigma^2(\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX})^{-1}\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX}\mathbf{Q}_{XX}^{-1}\\ =\;& \sigma^2\mathbf{Q}_{XX}^{-1} = \lim_{n\to\infty} Var(n^{1/2}(\mathbf{b}_{ols} - \boldsymbol{\beta}))\end{aligned}$$

Thus,

$$\begin{aligned} Var(n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols})) \;=\;& Var(n^{1/2}(\mathbf{b}_{iv} - \boldsymbol{\beta}))\\ & -Var(n^{1/2}(\mathbf{b}_{ols} - \boldsymbol{\beta}))\end{aligned}$$

More formally we can construct an Hausman test, where

$$\begin{aligned} H \;=\;& n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols})'(\widehat{Var}(n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols}))^{-1}\\ & \times n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols})\end{aligned}$$

where $\widehat{Var}(n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols}))$ is the estimated variance of $n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols})$. From what we have seen above,

$$\widehat{Var}(n^{1/2}(\mathbf{b}_{iv} - \mathbf{b}_{ols})) =$$

$$s^2((\mathbf{X'Z/n}(\mathbf{Z'Z/n})^{-1}\mathbf{Z'X/n})^{-1} - (\mathbf{X'X/n})^{-1})$$

with $s^2 = \mathbf{e'e}/n$, where $\mathbf{e}$ are either the OLS or the IV residuals. Thus we can rewrite H as:

$$\begin{aligned} H &= n(\mathbf{b}_{iv} - \mathbf{b}_{ols})' \Big(s^2((\mathbf{X'Z/n}(\mathbf{Z'Z/n})^{-1}\mathbf{Z'X/n})^{-1} \\ &\quad -(\mathbf{X'X/n})^{-1}) \Big)^{-1} (\mathbf{b}_{iv} - \mathbf{b}_{ols}) \end{aligned}$$

We have that: (a) Under H_0, $H \xrightarrow{d} \chi^2_k$, and (b) under H_A, H diverges to plus infinity.