Introduction to Instrumental Variables

Guillem Riambau. Econometrics, Yale NUS, Spring 2016

May 2, 2016

This is an introductory non exhaustive set of notes for Instrumental Variables (IV in the remaining). See references and textbook for more extensive and detailed notes.

1 Introduction to IVs

There are three main situations in which we will need to resort to IVs.

1. Simultaneity
2. Omitted Variable Bias
3. Measurement error

Consider the following (true) model

$$y_i = \alpha + \beta_1 x_i + \beta_2 a_i + \varepsilon_i \tag{1}$$

Where x and a jointly determine y. Assume that we cannot observe a (hence, it could be something like the individual's ability), and furthermore, that a and x are correlated. Hence the only model we can run is

$$y_i = \alpha + \beta_1 x_i + \varepsilon_i \tag{2}$$

We know that $\hat{\beta}_1$ will be biased. In particular, if the correlation between a and x is positive, it will be an upward bias.

Recall from earlier parts of the course that

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{cov(a,x)}{var(x)} \tag{3}$$

Hence, our measure is biased. What can we do? One way is to get a proxy for ability, but that's not always available. The alternative is to find what we call an instrumental variable for x. That is, a variable that explains x, isn't correlated with a, and can only

explain y through its effect on x (see examples below).

Technically, find a variable (that we will call z) such that:

1. $\rho(x, z) \neq 0$
2. $\rho(z, \eta) = 0$ where $\eta_i = \beta_2 a_i + \varepsilon_i$

The first condition means that z explains x. The second one means that z cannot explain y other than through its effect on x. This is called the exclusion restriction. We can never test the second one, since we cannot observe a in this case. Hence, economists spend a great deal of time and effort arguing why the second condition is met.

2 Three examples

2.1 Example 1: Institutions and economic success

Acemoglu, Johnson and Robinson have one of the most influential ever pieces in economics. Published in 2001, it has been cited, as of today, 8,746 times.[1] The paper is called The Colonial Origins of Comparative Development: An Empirical Investigation.

The goal of their paper is to find out whether good institutions actually lead to better economic outcomes. But we know that this is quite hard to disentangle, as good economic outcomes actually are likely to result in good institutions, too. So we have simultaneity: y affects x and x affects y. What to do?

AJR came with a quite clever way to circumvent this. They realised that Europeans, when they colonized Africa at the beginning, didn't stay in all areas. Some colonies became 'extractive colonies': these were the ones where Europeans didn't settle: they just tried to extract as many resources as possible. On the other hand, they chose to say in some others, which became 'settling colonies'. AJR argue that the kind of institutions Europeans created were very different *and have persisted until today.* To make it simple, the ones in the extractive colonies were *bad*, whereas the ones in the 'settling colonies' were good. Most importantly, AJR argue that the choice was exogenous: Europeans decided to stay in those areas in which their mortality rates were low.

This is the key: mortality rates of Europeans were caused by local diseases, mosquitos, etc. Hence, exogenous to Europeans. Moreover, it can also be argued that those local conditions have very little to do with current economic outcomes. Hence, we have an instrument: settler's mortality. Does it abide by the two conditions above stipulated?

[1] To give some perspective, other top influential articles are H.White, 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', Econometrica, 1980 (22,546 citations), or Kahneman and Tversky, 'Prospect theory: An analysis of decision under risk', Econometrica, 1979 (39,558 citations). AJR's paper was published in the American Economic Review. Note that it was published 20 years after these two key papers, which means that in terms of citations per year, it's doing quite ok, too

(i) Settler's mortality rates are correlated with quality of institutions. This is something that we can check. We just need to relate mortality rates and the quality institutions created. AJR show that there is a correlation.

(ii) Settler's mortality rates three hundred years ago cannot affect current economic outcomes - the only way they could affect them is by means of the institutions that resulted from those differences.

Hence, they use settler's mortality rates as an IV for quality of institutions. In case you're interested, they estimated large effects of institutions on income per capita.

2.2 Example 2: Policemen and crime

Steve Levitt, co-author of the Freakonomics series, also has a paper that serves as a good example of IV use: Using Electoral Cycles in Police Hiring to Estimate the Effects of Police on Crime.

Levitt tries to solve the usual dilemma: the observed correlation between crime and policemen on the streets is positive. Does it mean that the more police we have, the more crime we induce? or does it mean that we have more police in places in which there is more crime? If that is the case, how can we tell whether police actually helps to *deter* crime?

Levitt notices that the number of policemen hired increases considerably during election periods - or better said, the 12 months before an election. He also argues that election years are uncorrelated to crime - how could they be? Years are just numbers, they cannot affect crime.

In other words, the true model is

$$CRIME_{it} = \alpha + \beta_1 POLICE_{it} + X_{it}\beta_2 + \varepsilon_{it} \tag{4}$$

where t stands for period and i stands for location. X is a vector of characteristics of location i at time t.

Note that again both conditions are hold. Namely, (i) election years affect police hiring clearly (since incumbents realise that having more policemen before elections increases their popularity); and (ii) election years do not affect crime, and crime (should) not affect election years.

It is worth noting that election years *do* affect crime in an indirect way. Quoting Levitt, "the most obvious ways in which elections might systematically affect the crime rate (other than via changes in the police force) are through electoral cycles in other types of social spending, or through politically induced fluctuations in economic performance. Consequently, spending on education and public welfare programs is included in the equations, as are state unemployment rates. Having controlled for this items, it seems

plausible to argue that election timing will be otherwise unrelated to crime".

So to be precise, the true model is

$$(5)\ CRIME_{it} = \alpha + \beta_p POLICE_{it} + \beta_e EDUC_{it} + \beta_w PWELF_{it} + \beta_u UNEMP_{it} + X_{it}\gamma + \varepsilon_{it}$$

where $CRIME_{it}$ is number of crimes in i in period t, $POLICE_{it}$ is number of policment in i in period t, $EDUC_{it}$ and $PWELF_{it}$ refer to the investments in education and public welfare programmes in i in period t, $UNEMP_{it}$ refers to state unemployment levels in i in period t, and X_{it} is a vector of other controls.

Once we introduce the IV variables, this will become

$$(6)\ CRIME_{it} = \alpha + \beta_p ELECT_{it} + \beta_e EDUC_{it} + \beta_w PWELF_{it} + \beta_u UNEMP_{it} + X_{it}\gamma + \varepsilon_{it}$$

where $ELECT_{it}$ is a dummy that takes value 1 if location i had elections at time t. Since education, welfare etc, are included in the regression, $ELECT_{it}$ is uncorrelated with the errors, and hence there is no reason to suspect any bias. If they were not included, then $ELECT_{it}$ would not be doing a good job as an IV.

2.3 Example 3: Wages

Controlling for unobserved ability is another classical example of the need for IVs. Typically,

$$(7)\qquad log(\omega_i) = \alpha + \beta_1 x_i + \beta_2 a_i + \varepsilon_i$$

where x_i measures the qualifications of i(could be a dummy or a continuous variable). In the above notation, $log(\omega_i) = \alpha + \beta_1 x_i + \eta_i$

So we need to find an instrument z for a such that

$corr(z, x) \neq 0$

$corr(z, \eta) = 0$

3 Estimating $\hat{\beta}_{IV}$

Let's use the third example as the case in point. Assume the true model is as above

$$(8)\qquad log(\omega_i) = \alpha + \beta_1 x_i + \beta_2 a_i + \varepsilon_i$$

Further assume z is a good instrument for x. Then, if we have one regressor only, one candidate is for $\hat{\beta}_{IV}$ is

$$\hat{\beta}_{IV} = \frac{cov(y,z)}{cov(x,z)} \tag{9}$$

It is a good candidate because $E(\hat{\beta}_{IV}) = \beta_1$. The sketch of a proof follows:

Proof. $E(\hat{\beta}_{IV}) = \frac{cov(\alpha+\beta_1 x+\beta_2 a_i+\varepsilon_i,z)}{cov(x,z)} = \frac{cov(\alpha,z)}{cov(x,z)} + \frac{cov(\beta_1 x,z)}{cov(x,z)} + \frac{cov(\beta_2 a_i,z)}{cov(x,z)} + \frac{cov(\varepsilon_i,z)}{cov(x,z)} = 0 + \beta_1\frac{cov(x,z)}{cov(x,z)} + \beta_2\frac{cov(a_i,z)}{cov(x,z)} + \frac{cov(\varepsilon_i,z)}{cov(x,z)} = \beta_1\frac{cov(x,z)}{cov(x,z)}$ since the last two terms are 0 given the (assumption of) the exclusion restriction. ■

It can also be shown that $plim\hat{\beta}_{IV} = \beta_1$.

Note:

$$\hat{\beta}_{IV} = \frac{cov(y,z)}{cov(x,z)} = \frac{cov(y,z)/var(z)}{cov(x,z)/var(z)} = \frac{\hat{\delta}}{\hat{\gamma}}$$

where $\hat{\gamma}$ and $\hat{\delta}$ are the estimated coefficients from (10) and (11) respectively:

$$x = \alpha_1 + \gamma z + u_1 \tag{10}$$

$$y = \alpha_2 + \delta z + u_2 \tag{11}$$

Efficiency is greater with instruments that are more highly correlated with x , while still uncorrelated with the error terms.

4 Estimating $\hat{\beta}_{IV}$ when there is more than one regressor: 2SLS

2SLS stands for Two Stage Least Squares.

Some concepts:

1. Causal relation of interest: $y = \alpha + \beta_1 x + \eta$
2. First stage regression: $x = \alpha_1 + \gamma z + u_1$
3. Second stage regression: $y = \alpha + \beta_1 \hat{x} + \varepsilon$
4. Reduced form: $y = \alpha_2 + \delta z + u_2$

As the name says, generally two steps are taken:

1. Estimate $x = \alpha_1 + \gamma z + u_1$ and get the predicted values $\hat{x} = \hat{\alpha}_1 + \hat{\gamma} z$

2. Second stage regression: plug in the predicted values and estimate $y = \alpha + \beta_1 \hat{x} + \varepsilon$

Generally Stata will do both at once and compute correct standard errors.

The intuition of 2SLS is very useful: 2SLS only retains the variation in x that is generated by quasi-experimental variation (and thus hopefully exogenous).

4.1 2SLS in matrix notation

1. Run $X = Z\gamma + u_1$. Note Z may be a variable or a set of variables.
2. Get $\hat{\gamma}$ and compute $\hat{X} = Z\hat{\gamma} = Z(Z'Z)^{-1}Z'X = P_Z X$

Recall that we saw earlier in the course what we call projection matrices P_X: they are symmetric and idempotent. In this case, the projection matrix is P_Z, since we are projecting X onto the space generated by Z.

3. Plug $\hat{X} = P_Z X$ in $y = X\beta_1 + \varepsilon \Rightarrow y = \hat{X}\beta_1 + \varepsilon$
4. Run OLS of $y = \hat{X}\beta_1 + \varepsilon$ and get $\hat{\beta}_{2SLS}$:

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'y \\
&= ((P_Z X)' P_Z X)^{-1} (P_Z X)' y \\
&= (X' P_Z' P_Z X)^{-1} X P_Z' y \\
&= (X' P_Z P_Z X)^{-1} X P_Z' y \\
&= (X' P_Z X)^{-1} X P_Z' y
\end{aligned} \tag{12}
$$

although some people prefer to write it this way:

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'y \\
&= ((P_Z X)' P_Z X)^{-1} (P_Z X)' y \\
&= (X' P_Z' P_Z X)^{-1} (P_Z X)' y \\
&= (X' P_Z P_Z X)^{-1} (P_Z X)' y \\
&= (X' P_Z X)^{-1} (P_Z X)' y \\
&= ((P_Z' X)' X)^{-1} (P_Z X)' y \\
&= ((P_Z X)' X)^{-1} (P_Z X)' y \\
&= \left(\hat{X}' X\right)^{-1} \hat{X}' y
\end{aligned} \tag{13}
$$

Note that $\hat{\beta}_{2SLS} = \hat{\beta}_{OLS}$ in the special case where $z_i = x_i$ and hence $\hat{x}_i = x_i$. This is very intuitive - if we project x_i on itself, we obtain perfect predictions and the second stage of 2SLS coincides with the standard OLS regression.

The previous expressions for the 2SLS estimator remain valid when we have several explanatory variables (in the row vector x_i') and several instrumental variables (in the row

vector z_i'), in place of the scalars x_i and z_i.

Only in the (just-identifed) special case where we have the same number of instruments as we have explanatory variables (i.e. where the row vectors x_i' and z_i' have the same number of columns), we can also express the 2SLS estimator as

$$\hat{\beta}_{2SLS} = (Z'X)^{-1} Z'y$$

Proof: problem set

You should be able to show that $\hat{\beta}_{2SLS} = \hat{\beta}_{IV}$ when there is only one regressor.

4.1.1 What if some variables are exogenous but some are not?

So far, all algebra seems to be based on the fact that all explanatory variables (the xs) are endogenous. But we will find cases in which only some are, whereas some others are not (that is, they are not correlated to the error term). This is the more general case. Formally,

$$y = \alpha + X\beta + \delta W + \varepsilon$$

X is a vector of m endogenous variables (i.e. $corr(X, \varepsilon) \neq 0$). W is a vector of k exogenous variables (i.e. $corr(W, \varepsilon) = 0$). To be precise,

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \delta_1 w_1 + \delta_2 w_2 + ... + \delta_k w_k + \varepsilon$$

So all our W variables are okay already.

Say we find a set of l instruments, $l \geq m$. That is, $\tilde{Z}$ is a N by l matrix. Then a valid instrument is $Z = [\tilde{Z}W]$, so that W is an instrument of itself. In practice what we do is regress the following

$$x_i = \alpha + \gamma_1 \tilde{z}_1 + \gamma_2 \tilde{z}_2 + ... + \gamma_l \tilde{z}_l + \gamma_{l+1} w_1 + \gamma_{l+2} w_2 + ... + \gamma_{l+k} w_k + u$$
$$w_i = \alpha + \gamma_1 \tilde{z}_1 + \gamma_2 \tilde{z}_2 + ... + \gamma_l \tilde{z}_l + \gamma_{l+1} w_1 + \gamma_{l+2} w_2 + ... + \gamma_{l+k} w_k + u$$

for all x_i and all w_i (of course the αs and γs will be different for each regression, I am just simplifying notation). Why do we include the w_is? Well, it does not hurt: note that w_i is included as a regressor, so that the predicted $\hat{w}_i$ will be w_i itself.

Let $\hat{\gamma}_{w_i}$ denote the estimated coefficients from regressing w_i on $Z = [\tilde{Z}W]$ and $\hat{\gamma}_{x_i}$ denote the estimated coefficients from regressing x_i on $Z = [\tilde{Z}W]$. Then

$$
\begin{aligned}
\hat{x}_1 &= Z\hat{\gamma}_{x_1} = Z(Z'Z)^{-1}Z'x_1 = P_z x_1 \\
\hat{x}_2 &= Z\hat{\gamma}_{x_2} = Z(Z'Z)^{-1}Z'x_2 = P_z x_2 \\
&\vdots \\
\hat{x}_m &= Z\hat{\gamma}_{x_m} = Z(Z'Z)^{-1}Z'x_m = P_z x_m \\
\hat{w}_1 &= Z\hat{\gamma}_{w_1} = Z(Z'Z)^{-1}Z'w_1 = P_z w_1 = w_1 \\
&\vdots \\
\hat{w}_k &= Z\hat{\gamma}_{w_1} = Z(Z'Z)^{-1}Z'w_1 = P_z w_k = w_k
\end{aligned}
$$

so our vector of instruments (or predicted regressors is)

$$
\begin{aligned}
&[P_z x_1 \quad P_z x_2 \quad ... \quad P_z x_m \quad w_1 \quad ... \quad w_k] \\
&= [P_z x_1 \quad P_z x_2 \quad ... \quad P_z x_m \quad P_z w_1 \quad ... \quad P_z w_k] \\
&= P_z[x_1 \quad x_2 \quad ... \quad x_m \quad w_1 \quad ... \quad w_k] \\
&= P_z[XW]
\end{aligned}
$$

and now we are ready to run the second stage.

Note: the standard errors from the second stage need to be corrected. STATA and other softwares will do it for you. Why do we need correction? Intuitively, we lose a lot of variation in the xs. Once we use $\hat{x}_i$ instead of x_i, we are losing a lot of variability (note: all $\hat{x}_{ij}$s will be in the same line! where j denotes an individual in the sample.

5 Weak Instruments

Problems will arise when the instrument when this is only weakly correlated to the endogenous variable. The estimated coefficient will be biased towards the OLS coefficient. Furthermore, we do not know the correct asymptotic distribution of the parameters (for inference).

From the wikipedia page,

"Instrumental variables estimates are generally inconsistent if the instruments are correlated with the error term in the equation of interest. As Bound, Jaeger, and Baker (1995) note, another problem is caused by the selection of "weak" instruments, instruments that are poor predictors of the endogenous question predictor in the first-stage equation.[16] In this case, the prediction of the question predictor by the instrument will be poor and the predicted values will have very little variation. Consequently, they are unlikely to have much success in predicting the ultimate outcome when they are used to replace the question predictor in the second-stage equation.

In the context of the smoking and health example discussed above, tobacco taxes

are weak instruments for smoking if smoking status is largely unresponsive to changes in taxes. If higher taxes do not induce people to quit smoking (or not start smoking), then variation in tax rates tells us nothing about the effect of smoking on health. If taxes affect health through channels other than through their effect on smoking, then the instruments are invalid and the instrumental variables approach may yield misleading results. For example, places and times with relatively health-conscious populations may both implement high tobacco taxes and exhibit better health even holding smoking rates constant, so we would observe a correlation between health and tobacco taxes even if it were the case that smoking has no effect on health. In this case, we would be mistaken to infer a causal effect of smoking on health from the observed correlation between tobacco taxes and health."