Instrumental Variables Estimation

Søren Leth-Petersen

November 2016

1 Introduction

The principal goal of econometric analysis is to make causal inference, i.e. to use the econometric analysis to make statements about whether and how X causes Y. As econometricians we typically do this by formulating a linear regression equation

$$Y = \beta_0 + \beta_1 X + u \tag{1}$$

where β_1 is the primary object of interest telling us about how Y is impacted when we change X by one unit. Applying assumptions $MLR.1 - MLR.4$, i.e. assuming that we have specified the model correctly as in (1), that we have a random sample, that there is independent variation in X, and that the error term is uncorrelated with the regressors, $E(X'u) = 0$, then we can estimate β_0 and β_1 by simply applying OLS. When would that work? Take a simple example where Y is the height of an adult female and X is the height of her mother. The regression function says that mothers height X (perhaps along with other factors, which we leave out here) causes daughters' height Y. Intuitively, this makes sense because we think there is a genetic component to height, that the mother is born before the daughter (for good reasons, you might add) and that time only flows in one direction, i.e. we know that the mother has passed on her genes to the daughter. These arguments makes it plausible that $E(X'u) = 0$.

Think of the reverse case where Y is the height of the mother, and X is the height of her adult female daughter. Because we believe in the simple genetical theory outlined above we would not be willing to believe that the height of the daughter can cause the height of the mother. In fact, we are pretty sure that it is the reverse, and therefore $E(X'u) \neq 0$. It is not difficult to think of other, and perhaps more interesting examples, where it is hard to defend the assumption that $E(X'u) = 0$.

A leading case is when X is measured with error. If the measurement error is random then estimating β_1 by OLS will produce a parameter estimate that is biased towards zero. This is known as attenuation bias. But endogeneity can also arise because the individuals that we are trying to model, exhibit a particular kind of behaviour which gives rise to self-selection. For example, in a recent study Albæk et al. (2016) investigate whether peace time military service reduce criminal activity during and after military service. In this example X

is dummy variable taking the value one when a given individual in the sample has done military service and Y is criminal activity during and after military service. Why might military service impact criminal activity? For one thing, people are kept active for most of the time while doing military service, so they simply have less time to commit crime while in service. It may, however, also impact subsequent criminal activity. One of the objectives of peace time military service is to educate conscripts and to inform them about important civil values, or, in other words, to make them become good citizens. By teaching obedience and discipline military service may also provide skills that are potentially directly relevant in the labor market and, thereby, make labor market activity more attractive relative to criminal activity. However, since it is possible to join the military voluntarily in Denmark, it may not be a random selection of young people who join the military. For example, those who join might be those who have the most to gain (or equivalently, the least to loose) from joining service in terms of foregone earnings or it may simply be a group of people who have a particular preference for military service, and both factors may be related the propensity to commit crime.

How do we avoid the negative consequences of endogeneity? In the military service example one solution would be to completely randomise entry in to service. In this way it would not be possible to volunteer for service, and the randomisation would make sure that people entering service have the same characteristics as those who do not enter service, i.e. there would be no self-selection. However, complete randomisation is not possible. In fact, currently most people who enter military service do it voluntarily. So, we have to find other ways of estimating the causal effect. In this note we introduce the method of instrumental variables estimation. Instrumental variables estimation deals with situations where assumption MLR.4 cannot plausibly be defended by introducing an instrumental variable that is able to predict the endogenous variable without being correlated with the error term. Next to standard OLS regression, instrumental variable estimation is probably one of the most used tools in applied econometric analysis. The next section will outline the simple setup with one explanatory variable, which is endogenous. Section 3 will go through the more general case where there are potentially several endogenous variables and multiple instruments, including the case where there might be more instrumental variables available than endogenous variables. Throughout we will develop a Monte Carlo study to illustrate how the mechanics work and how it is implemented in Stata

2 Instrumental variables (IV) estimation

2.1 The simple case

We start out with the simplest case where there is only one explanatory variable, which is endogenous, and one instrument available. Consider a simple regression model:

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

x is suspected to be endogenous, $cov(x,u) \neq 0$, and OLS is hence going to produce biased and inconsistent estimates of β_1. The basic idea in IV estimation is to find an instrumental variable z that is able to predict the endogenous variable, $cov(z,x) \neq 0$, without predicting the error term, i.e. $cov(z,u) = 0$. In other words, the IV estimator splits the variation in x into two portions, good (read useful) and bad (read useless) variation. Bad variation is the part of the variation in x that is correlated with the error term u and hence causing bias, and good variation is the part of the variation in x that is not correlated with the error term, u and hence not causing bias. Using these assumptions the IV estimator can be derived by calculating the covariance between z and the components of (2):

$$cov(z,y) = \beta_1 cov(z,x) + \underbrace{cov(z,u)}_{=0} = \beta_1 cov(z,x) \tag{3}$$

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

So, β_1 is simply identified by the ratio of the covariance between z and y, and the covariance between z and x. Note that this is very similar to OLS. In fact if you replace z with x then you are back to OLS.
To actually make use of this insight for estimating β_1 we apply the corresponding data moments. Assume that we have available a random sample $\{y_i, x_i, z_i\}$ for $i = 1, ..., n$ observations.

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

It turns out that this estimator gives the right result in large samples, i.e. it is a consistent estimator of β_1, $plim(\widehat{\beta}_1) = \beta_1$. To see this

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\sum_{i=1}^n (z_i - \bar{z}) y_i}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\sum_{i=1}^n (z_i - \bar{z})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\sum_{i=1}^n (z_i - \bar{z}) \beta_1 x_i}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} + \frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} \\
&= \beta_1 + \frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}
\end{aligned}$$

Find the probability limit:

$$plim\, \widehat{\beta}_1 = \beta_1 + plim \frac{\frac{1}{n}\sum_{i=1}^n (z_i - \bar{z}) u_i}{\frac{1}{n}\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}$$

Recall the Law of Large Numbers saying that sample averages converge to population means

$$plim\,\frac{1}{n}\sum_{i=1}^{n} y_i = \mu$$

The IV estimator is **consistent** if $cov(z,u) = 0$ which was one of the assumptions that we started out from.

$$plim\,\widehat{\beta}_1 = \beta_1 + \frac{cov(z,u)}{cov(z,x)} = \beta_1$$

To see how this works in action we apply the estimator in a Monte Carlo study where we set up a data generating proces (DGP) for a simple linear regression model where x is measured with error.

Example 1: simulating measurement error in the simple linear regression model
We simulate data from the following DGP:

$$y_i = \beta_0 + \beta_1 x_i^* + u_i \tag{6}$$
$$x_i = x_i^* + v_i \tag{7}$$
$$z_i = x_i^* + e_i \tag{8}$$
$$\beta_0 = 1,\ \beta_1 = 2 \tag{9}$$
$$x_i^* \sim N(2,1),\ u \sim N(0,1),\ v \sim N(0,1),\ e \sim N(0,1),\ n = 200 \tag{10}$$

(6) specifies the regression function. (7) specifies that the observed variable x is related to the true but unobserved variable x^* by an standard normal random variable v. (8) specifies a variable z which is related to x^* by another standard normal random variable e. This variable is going to serve as an instrument for x in the Monte Carlo study. The idea here is that x and z are two noisy measures of the true but unobserved measure x^* where the noise component entering the two variables is independent. In this way z should satisfy the two conditions for being a good instrumental variable: $cov(z,u) = 0$ and $cov(z,x) \neq 0$. This DGP is cast to mimic the classical errors-in-variables problem outlined in Wooldridge, chapter 9. When the explanatory variable is measured with an error that is is uncorrelated with the true but unobserved variable then the OLS estimate of β_1 is going to exhibit attenuation bias according to the following expression

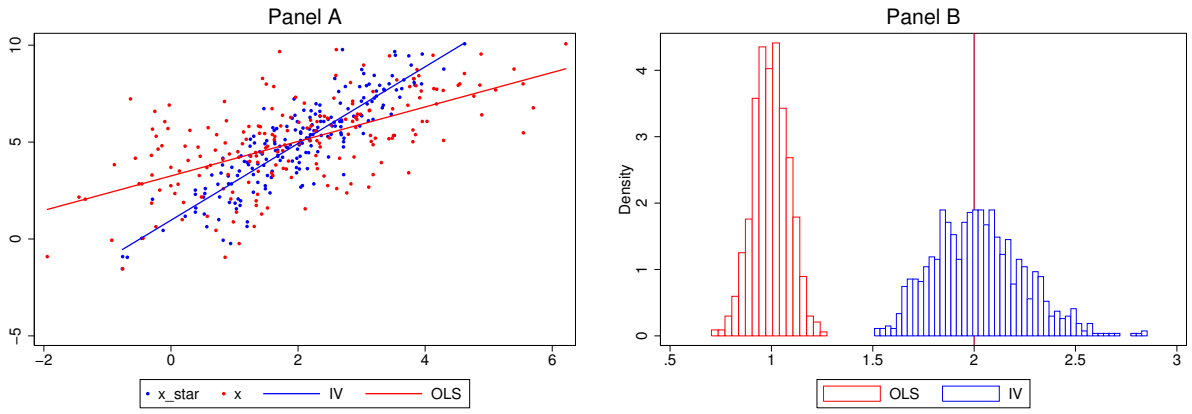
$$plim(\widehat{\beta}_1) = \beta_1(\frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_v}) \tag{11}$$

The DGP specifies $\sigma^2_{x^*} = 1$ and $\sigma^2_v = 1$, and we would thus expect that the OLS estimator on average yields and estimate $\widehat{\beta}_1^{OLS} = 1$. To implement a Monte Carlo study of the DGP in Stata we run the code in the box.
Figure 1, panel A shows the data and linear fit from one simulation of the DGP. The red

dots are data points of x and the blue dots are data points for x^*, both plotted against y. The solid red and blue lines are linear fits made one the two data clouds. The measurement error in x shows by the fact that the red dots span -1.95; 6.22 on the x-axis whereas the blue dots span -0.77;4.62. In other words the spread of x is bigger than the spread of x^* Accordingly the red regression line is less steep than the blue line representing the true slope. This illustrates the attenuation bias described by the (11).

```
program olsdata, rclass
        drop _all
        set obs 200                                  /* NUMBER OF OBS. */

        gen x_star=2+1*rnormal()                     /* DGP */
        gen x=x_star+1*rnormal()
        gen z=x_star+1*rnormal()
        gen u=rnormal()
        gen y=1+2*x_star+u

        regress y x                                  /* OLS ESTIMATES */
        return scalar b_ols=_b[x]

        ivregress 2sls y (x=z)                       /* IV ESTIMATES */
        return scalar b_iv=_b[x]
end

simulate beta_ols=r(b_ols) beta_iv=r(b_iv) ///  /* SIMULATE */
        , seed(117) reps(1000) nodots:olsdata
```

Figure 1: Results from Monte Carlo study



Panel B shows the distribution of $\widehat{\beta}_1^{OLS}$ and $\widehat{\beta}_1^{IV}$. $\widehat{\beta}_1^{OLS}$ is centred at 1, exactly as predicted by (11), and $\widehat{\beta}_1^{IV}$ is centred at 2, which is the true value of the parameter. This derives from the fact that z satisfies the two conditions for being a valid instrument. Another point to note is that the distribution of $\widehat{\beta}_1^{OLS}$ estimates is much more compressed than the distribution of $\widehat{\beta}_1^{IV}$ estimates. This illustrates a general feature of instrumental variables estimators, namely that they are generally ineffecient. We shall return to this later.

2.2 Adding control variables

The simple setup introduced in the previous section is attractive for introducing the principle idea of instrumental variables estimation. In practice, however, we rarely find use of a model with only one explanatory variable. Typically we would want to also control for a number of exogenous control variables. A more realistic set up is therefore

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \tag{12}$$

y is the dependent variable, x_1 is an explanatory variable that is endogenous, i.e. $cov(x_1, u) \neq 0$, and $x_2, ...x_k$ are exogenous variables. Estimating (12) by OLS will yield biased and inconsistent estimates of all the parameters, not just β_1. However we can straightforwardly apply IV estimation to equation (12). As before, we assume that we have available an instrumental variable z. The identifying assumptions are: $cov(z, x_1) \neq 0$ $cov(z, u) = 0$, $cov(x_j, u) = 0, j = 2, ..., k$. With this setup in hand we can actually make real empirical analysis, which the following example may illustrate.

Example 2: Peace time military service
Albæk et al. (2016) investigate whether peace time military service reduce criminal activity during and after military service. They estimate the following type of regression function:

$$crime = \beta_0 + \beta_1 D^{military} + \beta_2 x_2 + ... + \beta_k x_k + u \tag{13}$$

crime measures whether the individual under observation has committed crime (and has been caught!) during or after military service. $D^{military}$ is a dummy variable that takes the value one if a person has done military service. $x_2, ..., x_k$ are a set of control variables. The challenge is that those who enter military service are not selected randomly, since it is possible to volunteer. This implies that $D^{military}$ is potentially correlated with the error term, i.e. $cov(D^{military}, u) \neq 0$. However, In Denmark there is a draft lottery for military service. It literally works as a lottery where all young men (who are fit for service) draw a lottery number from a pool of numbers that is as big as the draft cohort. Upon having completed the draft lottery the military determines how many men are needed for service, picks the volunteers and then fills up with non-volunteers who have drawn the lowest lottery numbers. The group of people who end up doing military service thus consists of both volunteers and draftees. The potential endogeneity arises from the volunteers, and we would therefore like to single out only the part of the variation in $D^{military}$ that is due to the lottery (call the lottery number z), since this is surely unrelated to criminal activity, i.e. $cov(z, u) = 0$,

while at the same time able to predict entry into service $cov(D^{military}, z) \neq 0$. Albæk et al. (2016) use data that cover an extract of the 1964 birth cohort of Danish men. The data set includes information about convictions, schooling, labor market attachment, earnings, and family background. The information in the data set makes it possible to identify pre-conscription convictions, and this is used for identifying youth offenders and to estimate effects separately for this group. There are 951 youth offenders and 10,953 non-offending youths in the data set. An extract of the regression corresponding to (13) is presented in table 1 below, where separate regressions are presented for youth offenders and non-offending youths. The dependent variable is accumulated crimes over the period 1982-1990, and the key explanatory variable is a dummy variable indicating whether the person has done military service. For each group both OLS and IV regressions using the draft lottery number as an instrument for military service are presented.

Table 1: The effect of military service on property crime

	Youth offenders		Non-offending youths	
	OLS	2SLS	OLS	2SLS
Military enrollment	-0.285***	-0.438**	0.014	0.016
	(-2.78)	(-2.05)	(1.08)	(0.64)
No. of observations	951	951	10,593	10,593

Note: this table is an extract of table 5 in Albæk et al. (2016). Regressions include a list of control variables with information about draft age, family background, education, body height and weight, IQ test score, location of the home. t-values in parentheses.

Overall the results indicate that youth offenders reduce crime as a consequence of being enrolled in the military. Specifically, the OLS estimates indicate that one crime less is committed for every 3.5 youth offenders who has served. The IV estimates suggest a reduction of one crime for every 2.3 youth offender who has served, i.e. a bigger effect. Both effects are significant, and the fact that the magnitude of the estimate changes when applying IV estimation rather than OLS suggests that there is not random selection into military service. However, we do note, that the standard errors are so big that we cannot distinguish the two estimates in practice. The estimated effects among non-offending youths are insignificant in both cases. Based on the hypothesis that military service provides disciplin and education to become a better citizen this result is not surprising, because non-offending youths did not have a criminal record to improve on.

3 Two stage least squares

The simple IV estimator can be implemented when there are exactly as many instrumental variables available as there are endogenous explanatory variables. This is called the exactly

identified case. Sometimes, however, there there are more instrumental variables available than there are endogenous explanatory variables. In that case it is necessary to turn to a more general methodology, which is called Two Stage Least Squares (2SLS). We will start out presenting the 2SLS within the simple framework outlined in section 2. Then we will generalise it to allow for the case where there are multiple endogenous explanatory variables and potentially more instrumental variables than endogenous explanatory variables. This is called the overidentified case.

3.1 The simple case

The simple IV estimator from section 2 can be implemented in two steps:

1. Estimate $x_i = \pi_0 + \pi_1 z_i + e_i$ and compute $\widehat{x}_i = \widehat{\pi}_0 + \widehat{\pi}_1 z_i$
2. Estimate $y_i = \beta_0 + \beta_1 \widehat{x}_i + u_i$

In the first step x_i is regressed on the instrumental variable z_i, and $\widehat{x}_i$ is predicted out of this first regression. In the second step, y_i is regressed on $\widehat{x}_i$. This procedure provides a consistent estimate of β_1. The intuition is straightforward: z is exogenous and is used to split the variation in x into a "good" part that is not correlated with u and a "bad" part. That is: $x_i = \widehat{x}_i + \widehat{e}_i$ where $cov(\widehat{x}_i, u) = 0$ and $cov(\widehat{e}_i, u) \neq 0$. In the second step $\widehat{x}_i$ is used as a regressor. Because this part of x_i is exogenous we can estimate β_1 consistently in the second step.

3.2 The general case

We will now generalize the 2SLS method to the case where the model of interest potentially has more than one endogenous explanatory variable, more instrumental variables than there are endogenous explanatory variables, as well as several exogenous control variables. To do this we will turn to using matrix notation. We assume that we have available a random sample consisting of n observations. The regression model of interest is given by

$$y = X\beta + u \tag{14}$$

y and u are $(n \times 1)$ column vectors, X is a $(n \times k)$ where k indicate the number of explanatory variables including a constant term. We assume that $l < k$ of the variables in X are endogenous, and we arrange X so that these are positioned in the last l columns. As a consequence the first $k - l$ columns of X are exogenous variables. We also have available g instrumental variables. Given that there are l endogenous variables we assume that there are at least as many (and possibly more) instrumental variables as there are endognous regressors, i.e. $g \geq l$. We arrange the instrumental variables in a matrix Z where the first $k - l$ columns are the same as the first $k - l$ columns of X and the last g columns are filled in by the instrumental variables.

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_{k-l} & x_{k-l+1} & \dots & x_k \end{bmatrix} \tag{15}$$

$$Z = \begin{bmatrix} x_1 & x_2 & \dots & x_{k-l} & z_1 & \dots & z_g \end{bmatrix} \tag{16}$$

By assumption we have that all of the variables in Z are uncorrelated with the error term, i.e.

$$plim \frac{1}{n} Z'u = 0 \tag{17}$$

We also assume that

$$plim \frac{1}{n} Z'X = \Sigma_{ZX} \tag{18}$$

$$plim \frac{1}{n} Z'Z = \Sigma_{ZZ} \tag{19}$$

have full column rank so that they are nonsingular implying that X and Z are correlated.

Now consider the 2SLS approach within this more general setup. In the first step we regress the k variables X on the $k - l + g$ exogenous variables Z using OLS:

$$X = Z\Pi + E \tag{20}$$

where Π is a $g \times g$ matrix of parameters and E is an $n \times k$ matrix of residuals. These regressions are called *first stage regressions*. From these we obtain predictions of the k variables in X. The OLS parameters are given by

$$\widehat{\Pi} = (Z'Z)^{-1} Z'X \tag{21}$$

and the predicted values for the X variables are then

$$\widehat{X} = Z\widehat{\Pi} = Z(Z'Z)^{-1} Z'X = P_Z X \tag{22}$$

where $P_Z = Z(Z'Z)^{-1}Z'$, which is called the projection matrix, is symmetric, $P_Z = P_Z'$, and idempotent, $P_Z P_Z = P_Z$. Because $plim \frac{1}{n} Z'u = 0$ the predicted X's are also exogenous and can be used as regressors in a second step:

$$\widehat{\beta} = (\widehat{X}'\widehat{X})^{-1} \widehat{X}'y \tag{23}$$

$$= (X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X)^{-1} X'Z(Z'Z)^{-1}Z'y \tag{24}$$

$$= (X'P_Z P_Z X)^{-1} X'P_Z y \tag{25}$$

$$= (X'P_Z X)^{-1} X'P_Z y \tag{26}$$

In the special case where there are exactly as many instruments as endogenous explanatory variables, i.e. $g = l$, we can reduce this expression further. When $g = l$ then $Z'X$ is a $k \times k$ square matrix that can be inverted, and we can also make use of the following property for square matrices $(ABC)^{-1} = (A)^{-1}(B)^{-1}(C)^{-1}$.

$$\begin{aligned}
\widehat{\beta} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y && (27)\\
&= (X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y && (28)\\
&= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y && (29)\\
&= (Z'X)^{-1}Z'y && (30)
\end{aligned}$$

which is the analogue of the IV formula from (4).

3.3 Inference

We have now derived the 2SLS estimator for β. In order to be able to test hypotheses we also need formulas for calculating the variance. We do this both for the case where errors are homoskedastic and for the case where they are heteroskedastic.

3.3.1 Standard errors

The 2SLS estimator is given by:

$$\begin{aligned}
\widehat{\beta} &= (X'P_ZX)^{-1}X'P_Zy && (31)\\
&= \beta + (X'P_ZX)^{-1}X'P_Zu && (32)
\end{aligned}$$

The asymptotic covariance matrix is given by

$$\begin{aligned}
plim(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)' &= (X'P_ZX)^{-1}X'P_Zuu'P_ZX(X'P_ZX)^{-1} && (33)\\
&= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'uu'\widehat{X}(\widehat{X}'\widehat{X})^{-1} && (34)
\end{aligned}$$

An estimator of the asymptotic variance of the IV estimator in the case of heteroskedastic errors is:

$$\widehat{Var}(\widehat{\beta}) = (\widehat{X}'\widehat{X})^{-1}\sum_{i=1}^{n}\widehat{u}_i^2\widehat{x}_i'\widehat{x}_i(\widehat{X}'\widehat{X})^{-1} \quad (35)$$

Because we know that the data are sampled so that each unit, i, is drawn randomly it suffices to calculate $\widehat{u}_i^2\widehat{x}_i'\widehat{x}_i$ for each individual and then sum over the $i = 1, ..., n$ observations. The square root of the diagonal elements of the matrix expression in (34) are the heteroskedasticity-robust standard errors for the IV estimator.

Stata calculates robust standard errors if you add the option 'robust'

```
ivregress 2sls y (x=z), robust
```

If we are willing to assume homoskedasticity then the expression becomes simpler:

$$\widehat{Var}(\widehat{\beta}) = \widehat{s}^2(X'P_ZX)^{-1} \tag{36}$$

where $s = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2$. Expression (35) is a $k \times k$ matrix. Standard errors of β are given by the square root of the diagonal of this matrix.

3.3.2 Test for overidentifying restrictions

When there are more instruments available than endogenous variables then it is possible to test the assumption that It is possible to test the assumption that $plim\frac{1}{n}Z'u = 0$, i.e. that the instruments are uncorrelated with the error term from the second stage regression. When more instrumental variables than endogenous regressors are available then there are $(g-l)$ testable restrictions. Intuitively it is possible to think of it in the following way: the first l instrumental variables makes it possible to obtain a consistent estimate of $\widehat{\beta}$. Given this it is possible to calculate $\widehat{u}$ and then to check whether $\widehat{u}$ is correlated with the remaining $(g-l)$ instrumental variables (assuming that the first l instrumental variables are uncorrelated with u.

The test is conducted by running an auxiliary regression

$$\widehat{u} = Z\phi + v \tag{37}$$

The null hypothesis is $H_0 : \phi = 0$, ie. that the exogenous variables are uncorrelated with the residuals from the main equation.The test statistic is nR^2, and it is is χ^2-distributed with $(g-l)$ df under H_0. Unfortunately, the test for overidentifying restrictions has low power in small samples, i.e. it is not very good at identifying instrumental variables that are correlated with the error term in the equation of interest unless the data set is large. We will return to this point in a simulation example below.

When $l = g$, i.e. when there are exactly as many instrumental variables available as endogenous regressors then it is not possible to test the hypothesis that the instruments are uncorrelated with the error term, since in that case $R^2 = 0$ from the regression (37). To see this write up the OLS estimator for ϕ and substitute in for $\widehat{u}$

$$\begin{aligned}
\widehat{\phi} &= (Z'Z)^{-1}Z'\widehat{u} \\
&= (Z'Z)^{-1}Z'(y - X\widehat{\beta}) \\
&= (Z'Z)^{-1}Z'(y - X(Z'X)^{-1}Z'y) \\
&= (Z'Z)^{-1}(Z'y - Z'X(Z'X)^{-1}Z'y) \\
&= (Z'Z)^{-1}(Z'y - Z'y) \\
&= 0
\end{aligned}$$

3.3.3 Test for exogeneity

The reason for applying the IV/2SLS estimator is that one or more of the explanatory variables are endogenous. Typically, we hypothesise about this, but at the end of the day we do not know it. Nevertheless, it is important to know about it, because if it turns out that we can just as well treat the regressors as exogenous then it is better to apply OLS than IV/2SLS since OLS is BLUE, i.e. it is the most efficient linear estimator and hence produce the smallest standard errors on the estimates. Fortunately, a test for exogeneity of the regressors in the main equation has been developed.
Return to the system of k first stage regressions in (20) of which the last l equations are the first stage regressions for $x_{k-l+1} \dots x_k$. Having estimated these first stage equations it is possible to obtain the residuals $\hat{E} = \hat{E}_{k-l+1} \dots \hat{E}_k$. Intuitively, $\hat{E}$ collects all the bad variation in X, i.e. the part of the variation in X that is correlated with the error term in the main equation, u. The test for exogeneity of the regressors in the main equation simply estimates the main equation by OLS while adding the residuals obtained from the first stage as regressors:

$$y = X\beta + \hat{E}\rho + \epsilon$$

Inserting $\hat{E}$ is an attempt at making the part of u, which created the problem of correlation between X and u, observable. If $\hat{\rho}$ turns out to be significant then that is evidence that $\hat{E}$ is indeed correlated with u. More precisely, the null hypothesis is $H_0 : \rho = 0$, i.e. that X is exogenous. The test is conducted as a t-test if only one of the $X's$ is suspected to be endogenous and as an F-test if more regressors are suspected endogenous.

Example 3: simulating the 2SLS estimator and the test statistics for exogeneity and overidentifying restrictions
In this example we will simulate data for a regression equation where the explanatory variable is endogenous and illustrate how the 2SLS estimation works. The example will show how the 2SLS estimator is inefficient, i.e. it has bigger variance than OLS. Finally, the example will show the test for overidentifying restrictions has low power in small samples, i.e. it is not very good at rejecting instrumental variables that are in fact correlated with the error term in the equation of interest.

Consider the following data generating proces (DGP)

$$z_{1i} = \chi_{1i} \tag{38}$$

$$z_{2i} = \theta_1 z_{1i} + \theta_2 u_i + \chi_{2i} \tag{39}$$

$$x_i = \beta_0 u_i + \beta_1 z_{1i} + \beta_2 z_{2i} + \epsilon_i \tag{40}$$

$$y_i = \delta_0 + \delta_1 x_i + u_i \tag{41}$$

$$\delta_0 = 2, \delta_1 = 0.75 \tag{42}$$

$$\theta_1 = 0.25, \theta_2 = 0.2, \tag{43}$$

$$\beta_0 = -1, \beta_1 = 1, \beta_2 = 0.5 \tag{44}$$

$$\chi_{1i} \sim U(-0.5, 0.5),\ \chi_{2i} \sim N(0,1),\ u_i \sim U(-1,1), \epsilon \sim N(0,1) \tag{45}$$

(38) specifies an instrumental variable z_1. (39) specifies another, z_2, which is correlated with z_1 but also has independent variation determined by u an χ_2. (40) specifies the first stage regression. x is a function of the instrumental variables z_1, and z_2 , but it is also a function of u, which will also be the error term in the equation of interest. ϵ is just an independent error term in the first stage regression. Finally, (41) specifies the regression equation of interest. y is simply a function of x and the error term u. The fact that u appears in both (40) and (41) means that x is endogenous in (41). (42)-(44) specifies the parameter values, and (45) specifies the distributions from which we draw the random variables. Because u appears in both (40) and (41) and $\beta_0 = -1$ we would expect the OLS to produce a downwards biased estimate of δ_1. Note also, that u appears in the specification of z_2. This means that z_2 will be correlated with the error term in the equation of interest (41). We will simulate the test statistic for the overidentification test in order to see how well it is able to detect this violation of the basic assumptions for identification.
The Monte Carlo simulation is implemented by running the code in the box below. We calculate the OLS and 2SLS estimate of δ_1 as well as the test for overidentifying restrictions and the test for exogeneity and plot histograms for 10,000 simulations.

```
program ivdata, rclass
        drop _all

        *SET NUMBER OF CURRENT OBSERVATIONS
        set obs 100                /* Run for obs=100, 1000, 2000  */

        *DATA GENERATING PROCESS
        generate u = 2*(runiform()-0.5)
        generate z1 = (runiform()-0.5)
        generate z2 = 0.25*z1 + rnormal()+ 0.2*u
        generate x = -1*u + 1*z1 + 0.5*z2 + rnormal()
        generate y = 2 + 0.75*x + u

        *CALCULATE OLS ESTIMATES
        regress y x
        return scalar b1_ols=_b[x]

        *CALCULATE IV/2SLS ESTIMATE
        ivregress 2sls y (x=z1 z2)
        return scalar b1_iv=_b[x]
        predict uhat, residual

        * CALCULATE OI TEST
        regress uhat z1 z2
        return scalar OI=e(N)*e(r2)

        * CALCULATE TEST FOR EXOGENEITY
        regress x z1 z2
        predict ehat, residual

        regress y x ehat
        return scalar t_ehat=_b[ehat]/_se[ehat]

end

*SIMULATE PROGRAM 10000 TIMES
simulate ols=r(b1_ols) iv1=r(b1_iv) OI=r(OI)  t=r(t_ehat), seed(117) ///
        reps(10000) nodots:ivdata
```

We start out by simulating using 200 observations 10,000 times. In a first simulation we set $\theta_2 = 0$, so that both instrumental variables z_1 and z_2 satisfy the conditions for being valid instruments, but otherwise keep the setup as outlined in (38)-(45). In this way we will be

able to see that the OLS estimator is biased while the IV/2SLS estimator on average is able to produce the the right value of δ_1. Figure 2, Panel A shows histgrams of the distribution of the OLS and the 2SLS estimates. The modal point for the distribution of the OLS estimates is clearly smaller than the true parameter value as indicated by the vertical red line. This is makes sense, because x is negatively related to u because $\beta_0 = -1$. The distribution of the 2SLS estimates is centred at the true value. The figure also illustrates that the OLS estimator has a smaller variance than the IV/2SLS estimator.

Figure 2: Results from Monte Carlo study with 200 observations. z_1 and z_2 are valid instruments

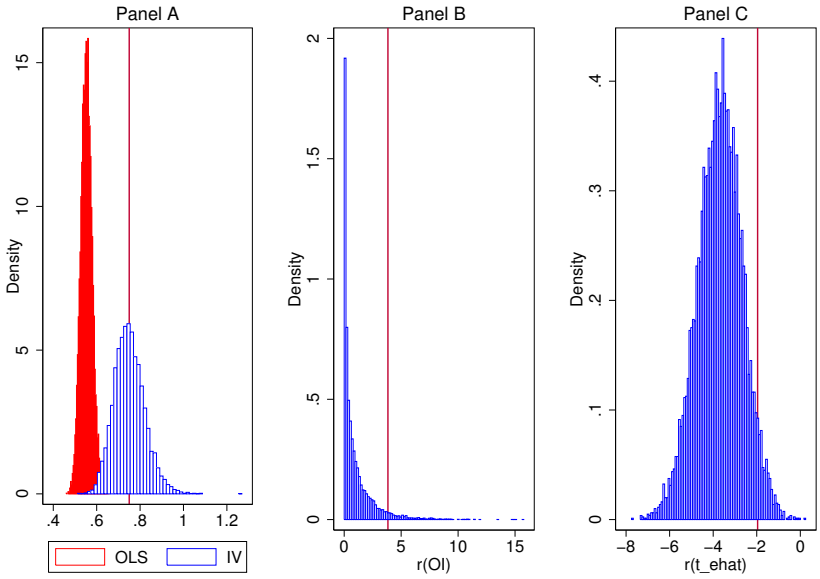


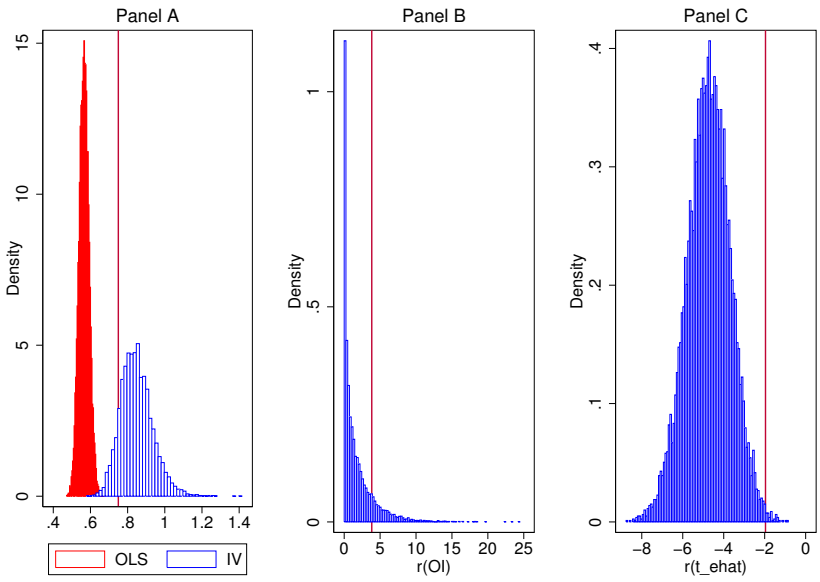
Figure 2, Panel B plots a histogram of the OI test statistic. The histogram shows that the major fraction of the calculated test statstics from the simulation falls within the acceptance region ($OI < 3.84$), just as it should, since in this first simulation both z_1 and z_2 are valid instrumental variables. Only 5 percent of the OI statistics reject the null hypothesis. Panel C plots a histogram of the t-test statistic for the exogeneity test. 95 percent of the simulated test statstics fall within the rejection region ($t < -1.96$), and this is also as expected, because x is endogenous because $\beta_0 = -1$. The exogeneity test seem to do good even in small samples.

Next, we simulate data exactly as is specified by (38)-(45), ie. $\theta_2 = 0.2$. Doing this means that z_2 is no longer a valid instrumental variable because it is correlated with u, the error term entering the equation of interest (41). We will now simulate data sets with 200 observations 10,000 times and see whether the test for overidentifying restrictions is able to reject the the hypothesis that z_1 and z_2 are valid instruments.

Figure 3, Panel A shows the distribution of OLS and IV/2SLS estimates of δ_1. OLS is still

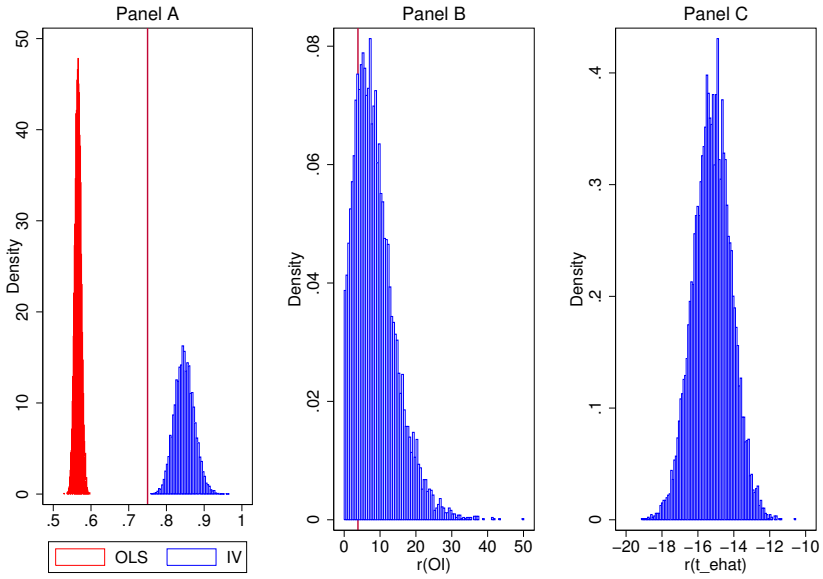
biased as before, but now the distribution of the 2SLS estimates is also not centered at the true value, albeit the distribution suggests that the 2SLS estimator is only slightly biased. Figure 3, Panel B plots a histogram of the OI test statistic. The histogram shows that the major fraction of the calculated test statstics from the simulation falls within the acceptance region ($OI < 3.84$). In fact only 14 percent of the OI statistics reject the null hypothesis. The test for overidentifying restrictions thus appears to do a bad job at detecting instruments that are not valid in small samples. Panel C plots a histogram of the t-statistic for the exogeneity test. It is not quite clear what to expect concerning the outcome of the exogeneity test since one of the instrumental variables are not valid. However, the histogram shows that the overwhelming fraction, 99 percent to be precise, of the calculated test statstics from the simulation falls within the rejection region ($t < -1.96$).

Figure 3: Results from Monte Carlo study with 200 observations



In a final experiment we keep the specification as in (38)-(45) but increase the number of observations in each simulation to 2000 (as opposed to 200 observations in the previous case). Figure 4, Panel A shows that both the OLS estimator and the 2SLS estimator is now estimated with less variance. It is now clear that they are both biased, albeit the 2SLS not as much as the OLS estimator. In panel B the distribution of test statistics from the test for overidentifying restrictions is plotted. The red vertical line shows the critical valuue, and it is now clear that the null hypothesis is rejected in most cases. The exact fraction being rejected is 80 percent. Panel C shows that in all cases the exogeneity test statistic is larger than the critical value, and the test thus rejects exogenity of x in all cases.

Figure 4: Results from Monte Carlo study with 2000 observations



In summary, the 2SLS estimator is inefficient and is mainly informative about the true value in big samples. The test for exogeneity is generally good at rejecting the null hypothesis of exogeneity when it is not true. The test for overidentifying restriction, on the other hand, is only informative in big samples.