Econometrics II

Lecture 4: Instrumental Variables Part I

Måns Söderbom*

12 April 2011

*mans.soderbom@economics.gu.se. www.economics.gu.se/soderbom. www.soderbom.net

1. Introduction

Recall from lecture 3 that, under the conditional independence assumption (CIA), we can correct for selection bias by conditioning on a set of observable variables. This approach amounts to moving the unobservable variable from the residual to the specification itself.

The instrumental variable approach, in contrast, leaves the unobservable factor in the residual of the structural equation, instead modifying the set of moment conditions used to estimate the parameters.

Outline of today's lecture:

- Recap & motivation of instrumental variable estimation
- Identification & definition of the just identified model
- Two-stage least squares (2SLS). Overidentified models.
- Generalized method of moments (GMM)
- Inference & specification tests
- IV estimation in practice - problems posed by weak & invalid instruments.

References:

Angrist and Pischke, Chapter 4.1-4.2 (4.1.3 is **optional)**; Greene, 12.3-4.

Murray, Michael P.(2006) "Avoiding Invalid Instruments and Coping with Weak Instruments," Journal of Economic Perspectives, 2006, vol. 20, issue 4, pages 111-132.

2. IV and Causality

Angrist and Pischke (AP) tell the IV story in two iterations:

1. In a restricted model with constant effects
2. In an unrestricted framework with heterogeneous potential outcomes

The second interpretation of the IV estimator has become very popular in the recent literature. However, exploring the mechanics of the IV estimator is easier if we impose homogeneous effects, so that's where we will begin.

We focus on the following potential outcomes model for earnings (introduced in the previous lecture):

$$\begin{aligned} Y_{si} &\equiv f_i(S), \\ f_i(S) &= \alpha + \rho S + \eta_i, \\ \eta_i &= A_i'\gamma + v_i, \end{aligned}$$

where we shall refer to A_i as a vector of 'ability' variables, and γ is a vector of population regression coefficients so that v_i and A_i are uncorrelated by construction.

For now, we also assume

$$E(v_i S_i) = 0,$$

implying that the variables A_i are the only reason why η_i and S are correlated. Hence, if A_i were observable, we would estimate the following "long" regression using OLS:

$$Y_i = \alpha + \rho S_i + A_i'\gamma + v_i.$$

Now suppose that A_i is **unobserved**. This implies our estimable equation takes the following form:

$$Y_i = \alpha + \rho S_i + \eta_i.$$

where $\eta_i = A_i'\gamma + v_i$ is the compound error term. Can we still estimate the parameter of interest ρ? Clearly OLS won't work now, since there will be omitted variables bias. But if we have access to an instrument z_i that is:

1. **Correlated** with the causal variable of interest (instrument relevance), $E(z_i S_i) \neq 0$; and

2. **Uncorrelated** with unobservable determinants of the dependent variable (instrument validity; or exclusion restriction), $E(z_i \eta_i) = 0$

then we can use instrumental variables techniques to estimate ρ. The latter condition can be used to derive an expression for ρ. Using matrix notation as follows

$$\begin{aligned} \mathbf{x}_i &= \begin{bmatrix} 1 & S_i \end{bmatrix} \\ \mathbf{z}_i &= \begin{bmatrix} 1 & z \end{bmatrix}, \\ \boldsymbol{\beta} &= \begin{bmatrix} \alpha \\ \rho \end{bmatrix} \end{aligned}$$

we write the causal relation of interest as

$$Y_i = \mathbf{x}_i \boldsymbol{\beta} + \eta_i,$$

and the moment conditions (or orthogonality conditions) as

$$E(\mathbf{z}_i' \eta_i) = \mathbf{0}.$$

Combining these two equations, we get

$$\begin{aligned} E(\mathbf{z}_i' \eta_i) &= \mathbf{0} \\ E(\mathbf{z}_i' (Y_i - \mathbf{x}_i \boldsymbol{\beta})) &= \mathbf{0} \\ E(\mathbf{z}_i' \mathbf{x}_i) \boldsymbol{\beta} &= E(\mathbf{z}_i' Y_i), \end{aligned}$$

which is a system of two linear equations. Assuming we can invert $E(\mathbf{z}'\mathbf{x})$, we can thus solve for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = [E(\mathbf{z}_i' \mathbf{x}_i)]^{-1} E(\mathbf{z}_i' Y_i). \tag{2.1}$$

This solves for our two unknown parameters $\boldsymbol{\beta}' = (\alpha, \rho)'$ from two linear equations, hence this model is **exactly identified.**[1] It follows from (2.1) that

$$\rho = \frac{Cov\left(Y_i, z_i\right)}{Cov\left(S_i, z_i\right)} = \frac{Cov\left(Y_i, z_i\right)/V\left(z_i\right)}{Cov\left(S_i, z_i\right)/V\left(z_i\right)}. \tag{2.2}$$

This shows that ρ is the ratio of the population regression of Y_i on z_i (called the reduced form) to the population regression of S_i on z_i (called the first stage); that is:

$$\begin{aligned} S_i &= \pi_0 + \pi_1 z_i + e_i, \\ Y_i &= \theta_0 + \theta_1 z_i + u_i, \end{aligned}$$

implies $\rho = \theta_1/\pi_1$. The IV *estimator* of ρ is the sample analog of (2.2).

Note the following:

- It is now obvious why the instrument must be relevant, i.e. correlated with the causal variable of interest. The relevance condition can be tested, for example by computing the t-statistic associated with $\hat{\pi}_1$ in the first stage regression. If the first stage is only marginally significantly different from zero, the IV estimates are unlikely to be informative - more about this later.

- The validity of the exclusion restriction, however, cannot be tested, because the condition involves the unobservable residual. Therefore, this condition has to be taken on faith, which is why relating it to economic theory is very important for the analysis to be convincing. We return to this at the end of this lecture, drawing on Michael Murray's (2006) survey paper.

A corollary of the exclusion restriction and instrument relevance is that the instruments cannot be explanatory variables in the original equation.

- Hence, if z_i is a valid and informative instrument, z_i impacts on Y_i *but only indirectly, through the variable* S_i.

[1] For the matrix $E\left(\mathbf{z}'\mathbf{x}\right)$ to be invertible it must have full rank, i.e. rank $E\left(\boldsymbol{z}'\boldsymbol{x}\right) = 2$ in our case.

- In what sense is an instrument very different from a proxy variable?

Finding an IV: Illustration Good instruments come from a combination of

- institutional knowledge(e.g. costs of schooling may vary across regions due to different regional policies)
- ideas about the processes determining the variable of interest

Keeping this in mind, let's have a closer look at the study by **Angrist and Krueger** (1991; QJE). Recall that these authors exploit variation - supposedly *exogenous* variation - in years of schooling driven by compulsory schooling in order to identify the returns to education for U.S. men.

- Most states require students to enter school in the calendar year in which they turn 6. Hence, those born late in the year are young for their grade.
- Students are required to remain in school only until their 16th birthday.
- So someone born early in the year may drop out in grade G while someone born late may not drop out until grade $G + 1$.
- Hence the 'natural experiment': children are compelled to attend school for different lengths of time, depending on their birthdays.
- Have a look at Figure 4.1.1 in AP: Panel (A) shows that those born early tend to have less education, on average, than those born late - consistent with the observation that those born early can exit at a lower level (G vs. $G + 1$; see above).
- Note that Panel A is basically the first stage regression: the graph is interpretable as predicted schooling based on a regression where years of schooling is the dependent variable, and year-of-birth and quarter-of-birth (and their interactions) are the explanatory variables
- Panel B shows how the average weekly wage varies with quarter of birth and year of birth. This is the reduced form regression, shown graphically. Findings:

  - Older men have higher earnings (due to experience)
  - Men born in early quarters have lower earnings.

- The latter result, combined with the patterns in the first stage regression, is consistent with the 'story' about how quarter of birth impacts on earnings. Importantly, because quarter of birth is likely unrelated to innate ability, it may be a valid instrument.

Mathematical representation of the story above:

$$\begin{aligned} S_i &= X_i'\pi_{10} + \pi_{11} z_i + \xi_{1i} \\ Y_i &= X_i'\pi_{20} + \pi_{21} z_i + \xi_{2i}. \end{aligned}$$

Now interpret the relationships just discussed within this framework:

- What are the endogenous variables?
- What are the exogenous variables?
- What are the exogenous covariates?
- What's the expression for the IV estimate of the return to education (ρ)?

2.1. Two-Stage Least Squares

The reduced-form equation

$$Y_i = X_i'\pi_{20} + \pi_{21} z_i + \xi_{2i}$$

can be derived by substituting the first-stage regression

$$S_i = X_i'\pi_{10} + \pi_{11} z_i + \xi_{1i}$$

into the causal relation of interest

$$\begin{aligned} Y_i &= X_i'\alpha + \rho S_i + \eta_i \\ Y_i &= X_i'\alpha + \rho\left(X_i'\pi_{10} + \pi_{11}z_i + \xi_{1i}\right) + \eta_i \\ Y_i &= X_i'\pi_{20} + \pi_{21}z_i + \xi_{2i}. \end{aligned}$$

Alternatively, the causal relation of interest be written as

$$Y_i = X_i'\alpha + \rho\left[X_i'\pi_{10} + \pi_{11}z_i\right] + \left\{\rho\xi_{1i} + \eta_i\right\},$$

where the term inside [.] is the population fitted value from the first stage regression. Since X_i' and z_i are exogenous, they are uncorrelated with the equation residual $\{\rho\xi_{1i} + \eta_i\}$.

In practice, we almost always work with data from samples. Applying OLS to the first-stage regression results in fitted values that are consistently estimated:

$$\hat{S}_i = X_i'\hat{\pi}_{10} + \hat{\pi}_{11}z_i + \xi_{1i}.$$

And the coefficient on $\hat{S}_i$ in the regression of Y_i on X_i and $\hat{S}_i$ is called the two-stage least squares (2SLS) estimator of ρ.:

$$Y_i = X_i'\alpha + \rho\hat{S}_i + \left\{\eta_i + \rho S_i - \rho\hat{S}_i\right\}.$$

Note that literally proceeding in two steps will give us the 2SLS estimate of ρ, however the standard errors reported from an OLS regression for the second stage will be wrong. So in practice we always use some software routine like **ivreg** (Stata) to obtain 2SLS estimates and correct standard errors.

Multiple instruments.

- We considered above the simple IV estimator with one endogenous explanatory variable, and one instrument. This is a case of **exact identification**. Similarly, if you have two endogenous explana-

tory variables and two instruments, the model is again exactly identified.

- If you have less instruments than endogenous regressors, the model is **underidentified**. This means you will not be able to estimate the parameter(s) of interest.

- If you have more instruments than endogenous regressors, the model is **overidentified**.

- In practice it is often a good idea to have more instruments than strictly needed, because the additional instruments can be used to increase the precision of the estimates, and to construct tests for the validity of the overidentifying restrictions (which sheds some light on the validity of the instruments).

- But be careful! While you can add instruments appealing to this argument, a certain amount of moderation is needed here. More on this below.

- Consider again the Angrist-Krueger study. Suppose we have 3 instrumental variables for S_i: z_{1i}, z_{2i}, z_{3i}. These would be dummies for first-, second-, and third-quarter births in this context, all of which are assumed uncorrelated with the residual η_i. The first-stage equation becomes:

$$S_i = X_i'\pi_{10} + \pi_{11}z_{1i} + \pi_{12}z_{2i} + \pi_{13}z_{3i} + \xi_{1i}, \tag{2.3}$$

  while the second stage is the same as previously; i.e.

$$Y_i = X_i'\alpha + \rho\hat{S}_i + \left\{\eta_i + \rho S_i - \rho\hat{S}_i\right\},$$

  but with the predictions now based on eq (2.3).Notice that we are now using **all** the instruments simultaneously in the first stage regression. By definition, the OLS estimator of the first stage regression will construct the **linear combination** of the instruments most highly correlated with S_i. By assumption all the instruments are exogenous, hence this procedure retains more exogenous variation in S_i than would be the case for **any** other linear combination of the instruments.

- Another way of saying this is that the instruments produce exogenous variation in predicted S_i, and OLS estimation in the first stage ensures there is as much such variation as possible. With fewer instruments there would be less exogenous variation in this variable, hence such estimators would not be efficient.

- What is the **relevance condition**, in this case where there are more instruments than endogenous regressors? In the current example, where we only have one endogenous regressor, it is easy to see that at least one of θ_j in the first stage has to be nonzero for the model to be identified.

You would be forgiven for thinking that, in practical applications, we should then use as many instruments as possible. After all, we said that including more instruments improves efficiency of the 2SLS estimator.

However, it is now well known that having a very large number of instruments, relative to the sample size, results in potentially serious bias, especially if some/many/all of the instruments are only weakly correlated with the endogenous explanatory variables. As we shall see below, using too many (weak) instruments tends to bias the 2SLS estimator towards the OLS estimator - i.e. the estimator we're trying to move away from! (What would happen if your number of instruments is equal to the number of observations?)

The advice on how to proceed in practice is to use a moderately overidentified model, trading off less efficiency for less bias. More on this below.

Now have a look at the Angrist-Krueger results, reproduced in Table 4.1.1. Note the following:

- 2SLS estimates are mostly **larger** than the corresponding OLS estimates, which suggests omitted ability is not a serious problem for the OLS estimator

- There are many instruments in column (7). The idea is to improve precision (indeed standard errors fall somewhat), but this may also lead to bias as discussed above.

General expression for the 2SLS estimator (Greene, 12.3.3) Define $\boldsymbol{Z}, \boldsymbol{X}, Y$ to be data matrices. Suppose there are K explanatory variables in the $\boldsymbol{X}$ matrix (including a constant), and L instruments

in the $\boldsymbol{Z}$ matrix (including a constant). The dimensions of the data matrices are thus as follows: $\boldsymbol{Z}$ is $N \times L$, $\boldsymbol{X}$ is $N \times K$, and Y is $N \times 1$).

Recall that, for the 2SLS estimator, we have

$$\hat{\boldsymbol{\beta}}^{2SLS} = \left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}\hat{\boldsymbol{X}}'Y, \tag{2.4}$$

which is the formula for the OLS estimator where we use predicted instead of actual values of the explanatory variables (for the exogenous variables in X, predicted and actual values coincide, of course). Now write the 2SLS estimator in terms of the raw data vectors $\boldsymbol{Z}$ and $\boldsymbol{X}$. Notice first that

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X},$$

(this is simply using the OLS formula for the K dependent variables in the first stage - i.e. the K explanatory variables in the second stage). I can now plug this into (2.4):

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{2SLS} &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'Y \\
\hat{\boldsymbol{\beta}}^{2SLS} &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'Y.
\end{aligned}$$

A common way of writing this is as

$$\hat{\boldsymbol{\beta}}^{2SLS} = \left(\boldsymbol{X}'\boldsymbol{P}_z\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_z Y,$$

where $\boldsymbol{P}_z = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$ is known as the **projection matrix**.

2.2. The Wald Estimator

The simplest IV estimator uses a single dummy instrument to estimate a model with one endogenous regressor and no covariates. The causal relation of interest is thus specified as

$$Y_i = \alpha + \rho S_i + \eta_i,$$

where η_i may or may not be correlated with S_i. Given that the instrument is a dummy variable, we have

$$\begin{aligned}
Cov\left(Y_i, z_i\right) &= E\left(Y_i z_i\right) - E\left(Y_i\right) E\left(z_i\right) \\
Cov\left(Y_i, z_i\right) &= pE\left(Y_i | z_i = 1\right) - E\left(Y_i\right) p \\
Cov\left(Y_i, z_i\right) &= pE\left(Y_i | z_i = 1\right) - \left(pE\left(Y_i | z_i = 1\right) + \left(1 - p\right) E\left(Y_i | z_i = 0\right)\right) p \\
Cov\left(Y_i, z_i\right) &= pE\left(Y_i | z_i = 1\right)\left[1 - p\right] - \left(1 - p\right) E\left(Y_i | z_i = 0\right) p \\
Cov\left(Y_i, z_i\right) &= \left\{E\left(Y_i | z_i = 1\right) - E\left(Y_i | z_i = 0\right)\right\} p\left[1 - p\right].
\end{aligned}$$

Along similar lines we can show that,

$$Cov\left(S_i, z_i\right) = \left\{E\left(S_i | z_i = 1\right) - E\left(S_i | z_i = 0\right)\right\} p\left[1 - p\right].$$

Recall $\rho = \frac{Cov(Y_i, z_i)}{Cov(S_i, z_i)}$; it follows that for the present model we have,

$$\rho = \frac{E\left(Y_i | z_i = 1\right) - E\left(Y_i | z_i = 0\right)}{E\left(S_i | z_i = 1\right) - E\left(S_i | z_i = 0\right)}.$$

This beautiful equation is the population analog of the **Wald estimator.** Interpretation is straightforward:

- The key assumption underlying the IV estimator is that the **only reason** for any relation between the dependent variable and the instrument (the numerator) is the effect of the instrument on the causal variable of interest (the denominator).

- In the context of a dummy instrument, it is therefore natural to divide the reduced form difference in means by the corresponding first-stage difference in means.

Perhaps the following decomposition is helpful for the intuition here:

$$\frac{\Delta Y}{\Delta X} = \frac{\Delta Y}{\Delta Z}\left(\frac{\Delta X}{\Delta Z}\right)^{-1} = \frac{\Delta Y/\Delta Z}{\Delta X/\Delta Z}.$$

3. Variance of the 2SLS estimator

Note: This draws on Greene, section 12.3.

Recall the general definition of the 2SLS-estimator:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{2SLS} &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'Y \\ \hat{\boldsymbol{\beta}}^{2SLS} &= \left(\boldsymbol{X}'\boldsymbol{P}_z\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_z Y\end{aligned}$$

where $\boldsymbol{P}_z = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$ is the projection matrix. Under homoskedasticity (constant variance of the error term), the covariance matrix has the same form as OLS, but in terms of predicted values:

$$\hat{Avar}\left(\hat{\boldsymbol{\beta}}^{2SLS}\right) = \hat{\sigma}^2\left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}.$$

Recall:

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}$$

(OLS formula applied to the first stage), thus

$$\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}} = \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-\mathbf{1}}\boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-\mathbf{1}}\boldsymbol{Z}'\boldsymbol{X},$$

i.e.

$$\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}} = \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-\mathbf{1}}\boldsymbol{Z}'\boldsymbol{X}$$

hence

$$Av\hat{a}r\left(\hat{\boldsymbol{\beta}}^{2SLS}\right)=\hat{\sigma}^{2}\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}, \tag{3.1}$$

where $Av\hat{a}r$ means 'asymptotic variance',

$$\hat{\sigma}^{2}=\left(N-K\right)^{-1}\hat{u}'\hat{u},$$

and

$$\hat{u}=Y-\boldsymbol{X}\,\hat{\boldsymbol{\beta}}^{2SLS},$$

is the $N\times 1$ column vector of estimated residuals. Notice that these residuals are **not** the residuals from the second-stage OLS regression of the dependent variable $\boldsymbol{Y}$ on the predicted variables of $\boldsymbol{X}$.

You might not think the variance formula above terribly enlightening. Some intuition can be gained by returning to the single-regressor single-instrument model

$$\begin{aligned} y &= \beta_1+\beta_2 x_2+u, \\ x_2 &= \delta_1+\delta_2 z_2+r. \end{aligned}$$

The variance of $\hat{\beta}_2^{IV}$ then simplifies to

$$\begin{aligned} Av\hat{a}r\left(\hat{\beta}_2^{IV}\right) &= \hat{\sigma}^{2}\left(\frac{\sum_i\left(\tilde{z}_{2i}\right)^2}{\sum_i\left(\tilde{x}_{2i}\tilde{z}_{2i}\right)^2}\right) \\ Av\hat{a}r\left(\hat{\beta}_2^{IV}\right) &= \hat{\sigma}^{2}\frac{1}{N}\sum_i\frac{\left(\tilde{z}_{2i}\right)^2}{N}\left(\frac{N}{\sum_i\left(\tilde{x}_{2i}\tilde{z}_{2i}\right)}\right)^2 \\ Av\hat{a}r\left(\hat{\beta}_2^{IV}\right) &= \hat{\sigma}^{2}\frac{1}{N}\frac{\sigma_z^2}{cov\left(x_{2i},z_{2i}\right)^2} \\ Av\hat{a}r\left(\hat{\beta}_2^{IV}\right) &= \hat{\sigma}^{2}\frac{1}{N\rho_{xz}^2\sigma_x^2}, \end{aligned}$$

where I have sample-demeaned the variables to eliminate the constants, and $\rho_{xz}=cov\left(z_{2i},x_{2i}\right)/\left(\sigma_z\sigma_x\right)$ is the correlation between x_2 and z_2.

Now notice the following:

- Just like the OLS estimator, the variance of the IV estimator decreases to zero at a rate of $(1/N)$.
- Just like the OLS estimator, the variance of the IV estimator falls, as the variance of the explanatory variable increases; and increases as the variance of the residual increases.
- It is now obvious why the assumption that the instrument is correlated with the explanatory variable is crucial: as ρ_{xz} tends to zero, the variance will tend to infinity.
- It's also obvious why your standard errors rise as a result of using instruments (compared to OLS) - since OLS amounts to using x as an instrument for itself, thus resulting in $\rho_{xz}^2 = 1$; whenever x and z are not perfectly correlated, the variance will be higher.

Heteroskedasticity-Robust Inference for 2SLS. If the error term is heteroskedastic, issues similar to those for OLS emerge for 2SLS:

- The 2SLS estimator is no longer asymptotically efficient (but it remains consistent),
- The variance formula (3.1) is no longer valid.

The two most common ways of guarding against heteroskedasticity are:

1. Use a heteroskedasticity-robust estimator of the variance matrix:

$$\hat{Avar}_{ROBUST}\left(\hat{\boldsymbol{\beta}}^{2SLS}\right) = \left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\hat{\boldsymbol{x}}_i'\hat{\boldsymbol{x}}_i\right)\left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}.$$

   Notice how similar this is to the robust variance estimator for OLS. Stata reports standard errors based on this estimator if you add 'robust' as an option in ivreg2.

4. Testing for exogeneity and validity of overidentifying restrictions

Whenever we use instrumental variables techniques we should carry out tests for **exogeneity** and for the **validity of the overidentifying restrictions**.

4.1. Testing for exogeneity: 2SLS

- Note: this is not discussed in great detail in AP, so I draw on Greene, Section 12.4. The exposition is also inspired by Chapter 6.2.1 in Wooldridge (2003; Econometric Analysis of Cross Section and Panel Data)

The main reason for using 2SLS or GMM is that we suspect that one or several of the explanatory variables are endogenous. If endogeneity is in fact **not** a problem, your instrumental variable estimator will be consistent (provided, of course, that the instruments are valid and relevant), but inefficient (i.e. higher variance than for OLS, given that OLS is valid). Therefore it is good practice to test for exogeneity. If we can accept the hypothesis that the explanatory variables are uncorrelated with the residual we are better off relying on OLS.

Consider the model

$$y_1 = \boldsymbol{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1,$$

where $\boldsymbol{z}_1$ is a $(1 \times L_1)$ vector of exogenous variables (including a constant), $\boldsymbol{\delta}_1$ is $(L_1 \times 1)$, and u_1 is the error term. The variable y_2 is potentially endogenous. I further assume that a set of (valid and relevant) instruments are available, so that

$$E(\boldsymbol{z}'u) = 0$$

holds by assumption, where $\boldsymbol{z}$ contains all the exogenous explanatory variables in the structural equation $\boldsymbol{z}_1$ **and** at least one instrument.

We are not sure if y_2 is endogenous or exogenous. If it is endogenous, we have

$$E(\boldsymbol{y}_2'u) \neq 0,$$

and I would identify the model relying on $E(\boldsymbol{z}'u) = 0$ only. However, if y_2 is really exogenous, then one additional moment condition becomes available:

$$E(\boldsymbol{y}_2'u) = 0.$$

In that case OLS will be fine. The null hypothesis, then, is that y_2 is exogenous.

$$H_0 : E(\boldsymbol{y}_2'u) = 0.$$

There are several ways of carrying out a test like this in practice.

4.1.1. The original Hausman (1978) test

Hausman (1978) proposed a test for exogeneity based on a comparison of the OLS and 2SLS estimators of $\boldsymbol{\beta}_1 = \left(\boldsymbol{\delta}_1', \alpha_1\right)'$. The general idea is very intuitive: if y_2 is in fact exogenous, then OLS and 2SLS estimators should differ only because of **sampling error** - i.e. they should not give significantly different results. Hausman showed that, under the null hypothesis, the test statistic

$$H = \left(\hat{\boldsymbol{\beta}}_1^{OLS} - \hat{\boldsymbol{\beta}}_1^{2SLS}\right)' \left[\hat{Avar}\left(\hat{\boldsymbol{\beta}}^{2SLS} - \hat{\boldsymbol{\beta}}_1^{OLS}\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}}_1^{OLS} - \hat{\boldsymbol{\beta}}_1^{2SLS}\right)$$

follows a Chi-squared distribution where the number of degrees of freedom equals the number of explanatory variables in the model. Notice the quadratic form of this expression. A complication here is posed by the calculation of $\hat{Avar}\left(\hat{\boldsymbol{\beta}}_1^{2SLS} - \hat{\boldsymbol{\beta}}_1^{OLS}\right)$. Hausman showed, however, that, asymptotically,

$$\hat{Avar}\left(\hat{\boldsymbol{\beta}}_1^{2SLS} - \hat{\boldsymbol{\beta}}_1^{OLS}\right) = \hat{Avar}\left(\hat{\boldsymbol{\beta}}_1^{2SLS}\right) - \hat{Avar}\left(\hat{\boldsymbol{\beta}}_1^{OLS}\right),$$

which is very useful. Hence, in practice the Hausman statistic is given by

$$H = \left(\hat{\boldsymbol{\beta}}_1^{OLS} - \hat{\boldsymbol{\beta}}_1^{2SLS}\right)' \left[\hat{Avar}\left(\hat{\boldsymbol{\beta}}_1^{2SLS}\right) - \hat{Avar}\left(\hat{\boldsymbol{\beta}}_1^{OLS}\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}}_1^{OLS} - \hat{\boldsymbol{\beta}}_1^{2SLS}\right).$$

Unfortunately, this particular test often proves problematic to use. The main problem is that, in small samples, there is no guarantee that $Av\hat{a}r\left(\hat{\boldsymbol{\beta}}_1^{2SLS}\right) > Av\hat{a}r\left(\hat{\boldsymbol{\beta}}_1^{OLS}\right)$. Clearly, if that happens we obtain a negative test statistic, which is hard to interpret given that H is non-negative in theory (follows a Chi-squared distribution under the null).

4.1.2. A regression-based Hausman test

Hausman has also derived a regression-based form of the test just outlined, which is less awkward to use in practice. This test, which is asymptotically equivalent to the original form of the Hausman test, is very general and very easy to implement in practice. To motivate this test, consider the reduced form equation (first stage):

$$y_2 = \boldsymbol{z}\boldsymbol{\pi} + v_2,$$

where $\boldsymbol{z}$ is uncorrelated with v_2 by definition; and the structural equation

$$y_1 = \boldsymbol{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1,$$

where u_1 is uncorrelated with $\boldsymbol{z}$, by assumption. Now think about the implications of y_2 being either i) exogenous or ii) endogenous.

- If y_2 is exogenous, i.e. $E(y_2 u_1) = 0$, then it **must** be that $E(v_2 u_1) = 0$, given that $\boldsymbol{z}$ is uncorrelated with v_2 and u_1 (otherwise y_2 would be correlated with u_1)
- If y_2 is endogenous, i.e. $E(y_2 u_1) \neq 0$, then it **must** be that $E(v_2 u_1) \neq 0$, given that $\boldsymbol{z}$ is uncorrelated with v_2 and u_1 (there is no other way y_2 can be correlated with u_1).

It is thus clear that our exogeneity test can be formulated as

$$H_0 : E(v_2 u_1) = 0,$$

i.e. the null hypothesis is that the two residuals are uncorrelated. Now write the linear projection of the residual u_1 on the reduced form error u_2:

$$u_1 = \rho_1 v_2 + \xi_1.$$

This implies $E(v_2, u_1) = \rho_1 \sigma_v^2$, hence we can rewrite the null hypothesis of exogeneity as

$$H_0 : \rho_1 = 0.$$

Thus, y_2 is exogenous if and only if $\rho_1 = 0$. To see how this is useful from an applied point of view, now replace u_1 by $\rho_1 v_2 + \xi_1$ in the structural equation:

$$y_1 = \boldsymbol{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + \xi_1.$$

Of course, v_2 is not directly observed, but it can be **estimated** from the reduced form equation:

$$\hat{v}_2 = y_2 - \boldsymbol{z}\hat{\boldsymbol{\pi}},$$

and we can then run the structural regression

$$y_1 = \boldsymbol{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 \hat{v}_2 + error, \tag{4.1}$$

using OLS (note!) and actual, not predicted, y_2.

- The exogeneity test can now be done as a simple t-test of the null that $\rho_1 = 0$.
- A heteroskedasticity-robust t-test can be used if you suspect there is heteroskedasticity under the null.
- Incidentally, using OLS to estimate (4.1) gives estimates of the parameters $\boldsymbol{\delta}_1$,α_1 that are *numeri-*

cally identical to 2SLS. However, the OLS standard errors associated with (**??**) are valid under the null that $\rho_1 = 0$, but not under the alternative that $\rho_1 \neq 0$. In the latter case, the conventional standard errors are downward biased. One implication of this is that, if you do not reject the null hypothesis based on standard errors that are possibly too low, you certainly wouldn't do so based on the correct standard errors.

4.2. Testing for validity of overidentifying restrictions: 2SLS

- In an exactly identified model we **cannot** test the hypothesis that the instrument is valid, i.e. that the exclusion restriction is a valid one. In that case, the assumption that the instrument is valid will essentially have to be taken on faith - i.e. you have to believe the theoretical arguments underlying the exclusion restriction.[2]

- If our model is overidentified, we can test for the **validity of the overidentifying restrictions**. Please note that this is **not** a test of the hypothesis that "the instruments are valid". Rather, it is as follows:

  - Under the assumption - which we can't test - that G_1 instruments are valid with **certainty**, where G_1 is the number of endogenous explanatory variables, we can test the null hypothesis that the $Q_1 = L_2 - G_1$ overidentifying instruments (where L_2 is the total number of instruments) are orthogonal to the residual in the structural equation.

- So what's the point of considering this test, then, given that it does not shed light on the issue that we are interested in (which is instrument validity, in general)? You can view the OVERID test as a first hurdle that needs to be overcome in the context of IV estimation, in the following sense:

- If the OVERID test indicates you should **reject** the null hypothesis, then this is pretty clear evidence your model is mis-specified. You then have no choice but to respecify the model. When doing so, think carefully about the implications of the test outcome. Whenever the OVERID test implies rejection of the null, this usually means at least one of the instruments would have a

[2] To see the intuition of why we cannot test for the validity of this assumption, consider the exactly identified model

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 y_2 + u_1, \\ y_2 &= \pi_0 + \pi_1 z_1 + v_2. \end{aligned}$$

Express the structural equation as a function of the predicted value of Y_2:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 (\hat{\pi}_0 + \hat{\pi}_1 z_1) + u_1 \\ &= (\beta_0 + \beta_1 \hat{\pi}_0) + \beta_1 (\hat{\pi}_1 Z_1) + u_1. \end{aligned}$$

We cannot test the hypothesis $cov(z_1, u_1) = 0$, simply because u_1 is not observed and, without further information, we cannot obtain an estimate of u_1 unless we assume $cov(z_1, u_1) = 0$. That is, the estimate of u_1 will be uncorrelated with z_1 by construction.

significant effect in the structural equation. Think about the economics of that. For example, if you are instrumenting education with distance to primary school at the age of seven, and mother's education, you might think mother's education is a dubious instrument as it may be correlated with unobserved ability. So the next step could be to re-estimate the model without mother's education in the instrument set.

- If the OVERID test suggests you should **accept** the null hypothesis, then what to make of this depends largely on the faith you have in your instruments in general. If you are almost certain that G_1 instruments are valid, then you might be inclined to conclude that the model passing the OVERID test means that **all** your instruments are valid (perhaps some of your instruments are less credible than others, in which case this might be useful knowledge).

Intuition of the OVERID test. Suppose the model is

$$\begin{aligned} y_2 &= \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2 \\ y_1 &= \beta_0 + \beta_1 y_2 + u_1, \end{aligned}$$

which is overidentified. We know we can obtain IV estimates of the structural equation here by using only z_1 as an instrument. Because in that case z_2 is not used in the estimation, we can check whether z_2 and the estimated residual $\hat{u}_1$ are correlated. If they are, then z_2 would not be a valid instrument, under the assumption that z_1 is a valid instrument (we need this assumption, otherwise the model is not identified of course).

Clearly we can then reverse the roles of z_1 and z_2 and examine whether z_1 is uncorrelated with $\hat{u}_1$ if z_2 is used as an instrument.

Which test should we use? It turns out that this choice does not matter. Remembering that, in this case, the validity of at least one IV must be taken on faith.

Mechanics of the basic OVERID test for 2SLS. Such a test can be carried out as follows:

1. Estimate the structural equation with 2SLS / IV and obtain the estimated residuals $\hat{u}_1$.

2. Regress $\hat{u}_1$ on **all** exogenous variables (in the example above, z_1 and z_2). Obtain the R-squared.

3. Under the null hypothesis that the instruments are uncorrelated with u_1, the statistic $N \times R^2$ follows a chi-squared distribution with Q_1 degrees of freedom. If $N \times R^2$ exceeds the relevant critical value then we conclude that some of the instruments are not uncorrelated with u_1, in which case they are not valid instruments.

There is an equivalent way of carrying out the OVERID test, which is based on the **criterion function** that is (implicitly) being minimized to yield the 2SLS results. See Section 4.2.2 in AP for a brief discussion.

[Discuss Card (1995); See section 1 in the appendix]

5. Discussion: Using IV in practice

Reference: Murray, Michael P.(2006) "Avoiding Invalid Instruments and Coping with Weak Instruments," Journal of Economic Perspectives, 2006, vol. 20, issue 4, pages 111-132

- The survey paper by Murray (2006) is an excellent survey paper on the instrumental variable estimator, stressing intuition and implications rather than technicalities.

- He begins by discussing some studies using instruments to identify causal effects. He then asks: should instrumental variable be thought of as a panacea (a cure for all diseases)? He argues not. Two reasons:

  - Instruments may be invalid. This would result in inconsistent estimates and possibly greater bias than for OLS. Indeed, since you can never be certain that your instruments are valid, there's a "dark cloud of invalidity" hanging overhead all instruments when they are arrive on the scene.

– Instruments may be so weakly correlated with the endogenous explanatory variables (referred to as 'troublesome' variables in the paper) that in practice it's not possible to overcome the bias of the OLS estimator. Weak instruments lead to bias, and misleading inference (common result: standard errors far too low), in instrumental variable estimation.

5.1. Supporting an instrument's validity

In order to chase away the dark cloud of instrument invalidity, you need to use economic arguments combined with statistical analysis.

1. You need to advance theoretical arguments as to why your instruments are valid ones. A very common view in the profession is that how much credence should be granted to IV studies depends to a large extent on the quality of the arguments in support of the instruments' validity. You will see a good example of this in the Miguel et al. paper (Lab 1).

2. Test for the validity of overidentifying restrictions. Of course, to have maximum faith in such a test you need to know with certainty that an exactly identifying subset of the instruments are valid. In practice, typically you don't know. But if you're using different instruments with different rationales, so that one might be valid while the other is not, then your audience will have more faith in the instruments if the OVERID test is passed. If your instruments are basically variants on the same theme - e.g. all measures of institutional quality - then it seems more unlikely that some can be valid whilst others are not. In any case, what you're definitely not allowed to do is say, because the OVERID restrictions look valid, that "the instruments are valid". You can never be sure.

3. Be diligent about omitted variables. Omitted variables bias is a relevant concern in the context of IV estimation - but in a somewhat different form, compared to OLS. In particular, IV estimation is biased if an omitted relevant variable is correlated either with the included non-endogenous explanatory variables (X) or the instrumental variables (Z). So there are good reasons for adding

control variables, even if you're estimating with instrumental variables. With panel data we may want to control for fixed effects, for example.

4. Use alternative instruments (rotate the instruments). This in the spirit of the OVERID test. If you have many instruments, then try adding them one by one and check if your results are robust. If parameter estimates vary a lot depending on which instruments are being used, this would be a sign that not all your instruments are valid.

5.2. Coping with weak instruments

Estimation and inference with weak instruments - instruments only weakly correlated with the endogenous variables - is an area of active research. Some of the theoretical arguments are rather technical, but the main points are pretty straightforward. Let's start by looking at some straightforward results.

Weak instruments imply high variance: We have seen that if the instruments and the endogenous regressor(s) are only weakly correlated, the variance of the IV estimator can be rather high - recall that, in the single-regressor single-instrument model:

$$\hat{Var}\left(\hat{\beta}_1^{IV}\right) = \hat{\sigma}^2 \frac{1}{N\rho_{xz}^2\sigma_x^2}.$$

Weak instruments exacerbate the bias caused by invalid instruments: Another implication of weak instruments is that the IV estimate may be quite badly inconsistent even as the sample size tends to infinity. To see this, recall that

$$\begin{aligned}
p\lim \hat{\beta}_1^{IV} &= \beta_1 + p\lim \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})\,u_i}{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})\,(x_i - \bar{x})}, \\
p\lim \hat{\beta}_1^{IV} &= \beta_1 + \frac{cov\,(z_i, u_i)}{cov\,(z_i, x_i)}, \\
p\lim \hat{\beta}_1^{IV} &= \beta_1 + \frac{corr\,(z_i, u_i)}{corr\,(z_i, x_i)}\frac{\sigma_u}{\sigma_x}.
\end{aligned}$$

Clearly, the inconsistency in the IV estimator can be large if $corr\,(z_i, u_i) \neq 0$ and $corr\,(z_i, x_i)$ is relatively small.

- *Student checkpoint*: Show that the OLS estimator will have smaller asymptotic bias than the 2SLS estimator whenever

$$corr\,(x_i, u_i) < \frac{corr\,(x_i, u_i)}{corr\,(z_i, x_i)}.$$

Clearly, if z_i and x_i are not correlated at all and $corr\,(z_i, u_i) \neq 0$, the asymptotic bias of the IV estimator tends to infinity. Thus it is important to establish whether z_i and x_i are correlated or not.

Weak instruments lead to small sample bias, even if $corr\,(z_i, u_i) = 0$ in the population:

- A much more subtle point than those raised above is that, even if $corr\,(z_i, u_i) = 0$ in the population (so that the instrument is valid) it is now well understood that instrumental variable methods can give very misleading results - biased parameter estimates, downward biased standard errors - in small samples.
- Problems can become particularly serious if we have
  - **Weak** instruments; and/or
  - **Many** instruments (large number of overidentifying restrictions)
- You might think having a large sample solves these problems, but that is not necessarily the case. Angrist and Krueger (1991) used more than 300,000 observations to estimate the returns to education, but because they used a very large number of instruments, some of the inference reported in that paper is not reliable, as shown by Bound, Jaeger and Baker (1996). So the issue is not sample size, but how informative your data are.

[EXAMPLE on small sample & strong instruments vs. large sample & weak instruments - section 2 in the appendix.]

- When instruments are only weakly correlated with the endogenous explanatory variable(s), two serious problems emerge:

1. Biased parameter estimates: Even though 2SLS estimates are consistent (i.e. they almost certainly approach the true value as N goes to infinity), the estimates are **always biased** in finite samples. When the instruments are weak, this bias can be large - even in large samples.

2. Biased standard errors: When the instruments are weak, 2SLS standard errors tend to become too small - i.e. you'd reject the null too often.

The combination of these problems is disturbing: the mid-point of your confidence interval is in the wrong place, and the width of the confidence interval is too narrow.

[EXAMPLE. Results from a simulation based on a model with many instruments, *all of which are uninformative (irrelevant)* - section 3 in the appendix].

- There is now quite a large literature on the implications of weak/many instruments for inference. This literature is fairly technical. Practitioners need to be aware of the pitfalls however. ivreg2 produces several tests that shed light on whether weak instruments are likely to be a problem in practice. Murray (2006) provides a useful discussion. The rest of this section draws heavily on his exposition.

Biased parameter estimates. Here's an argument that should make it immediately obvious to you that 2SLS can be biased in finite samples: suppose you have one endogenous regressor, and suppose the number of instruments is equal to the number of observations. In this case the first stage regression will result in $R^2 = 1$, and the predicted value of the endogenous variable in the first stage will coincide with the actual value. Your 2SLS estimator coincides exactly with the OLS estimator (the one you were suspicious of in the first place).

We can be a bit more precise. Consider the following simple model:

$$\begin{aligned} Y_{1i} &= \beta_0 + \beta_1 Y_{2i} + \varepsilon_i, \\ Y_{2i} &= \alpha_0 + Z_i \alpha_1 + \mu_i, \end{aligned}$$

where $Var(\varepsilon_i) = Var(\mu_i) = 1$ for convenience.

The explanatory variable Y_{2i} is endogenous if $corr(\varepsilon, \mu) \neq 0$. Define $\rho = corr(\varepsilon, \mu)$.

Hahn and Hausman (2005) show that, for this specification, the finite-sample bias of 2SLS for the overidentified model ($l > 1$), where l is the number of instruments in the Z_i vector, can be written

$$E\left[\hat{\beta}_{1,2SLS} - \beta_1\right] \approx \frac{l\rho\left(1 - R^2\right)}{nR^2},$$

where R^2 is the R-squared from the first stage, and n is the number of observations.[3]

- Key insight: The bias rises with three factors -
  - The number of instruments used
  - The correlation between the residuals (strength of endogeneity)
  - Weakness of the instruments (weak instruments $\rightarrow$low R^2 in the first stage).
- Clearly these problems will be more severe in small samples.
- Recall that adding instruments might be thought a good idea on the grounds that standard errors decrease. Now you see there is a cost associated with that, in terms of bias. Note in particular that this cost will be high if the instruments are **weak** - why?
- Example: Suppose $l = 15$, $\rho = 0.5, R^2 = 0.20$, $n = 200$, $\beta_1 = 1$. In this case, we would have

$$E\left[\hat{\beta}_{1,2SLS} - \beta_1\right] \approx \frac{15 \times 0.5 \times 0.8}{200 \times 0.2} = 0.15,$$

[3] To derive this formula you need to know a few matrix tricks. Let me know if you are interested.

i.e. a bias of 15%.

- *Student checkpoint:* Can you derive the bias in the OLS estimator for this model? How do the 2SLS and OLS estimators compare, in terms of bias? Can OLS ever be less biased? This is a fundamental question - the whole point of using 2SLS is to reduce the bias produced by OLS.

- *Student task* (optional - but should be fun): Can you write a Stata program that computes the bias above by means of simulations? Are the simulations results consistent with the analytical formula?

- I will now partly reveal the answer to the question set above: yes, if the instruments are too weak and/or too many, then the 2SLS estimator may be more biased than the OLS estimator.

- Stock and Yogo (2005) provide a formal test for when an IV is "too weak" to be trustworthy. The null hypothesis in this test is that bias of 2SLS is some fraction of the bias of OLS (e.g. less than 10%).

- In the simplest case where there's just one endogenous explanatory variable, the key test statistic is the F-statistic in the first stage (with non-standard critical values, however).

Biased standard-error estimates.

- The estimated variance of 2SLS is generally biased downward in finite samples - and the bias can become large when the instruments are weak. This means that you will tend to reject the null hypothesis too often if you rely on the 2SLS standard errors.

- Stock and Yogo (2005) proposed a test of the null hypothesis that the true significance of hypothesis tests about the endogenous regressor's coefficient is smaller than 10% (and 15,20,25%) when the usually stated significance level is 5%. Such tests are reported by ivreg2. Clearly, if your test statistic is lower than, say, 25% maximal IV size, then your standard errors are very unreliable (strongly downward biased).

PhD Programme: Econometrics II
Department of Economics, University of Gothenburg
Appendix Lecture 4
Måns Söderbom

Instrumental Variable Estimation in Stata

I will use the Stata command **ivreg2**, which has been developed by Stata users (not Stata Corp.). If this command is not already on your computer, you should be able to install it by typing

ssc install ivreg2

in the Stata command window.

In version 10 of Stata, the command **ivregress** is available, which is similar to ivreg2 (though not quite as comprehensive). Older versions of Stata have the command ivreg, which is a little bit too limited for our purposes.

1. Illustration using the CARD.RAW data

Using wage data for 1976, Card (1995) uses a dummy variable indicating whether a man grew up in the vicinity of a four-year college as an IV for years of schooling.[1] The data can be downloaded from the web (the file name is CARD.RAW). These data are used by Wooldridge (2003; Econometric Analysis of Cross Section and Panel Data).

```
. use CARD.dta, clear
```

Table 1.1 OLS

```
      Source |       SS       df       MS              Number of obs =    2220
-------------+------------------------------           F( 15,  2204) =   54.96
       Model |  116.783056    15  7.78553706           Prob > F      =  0.0000
    Residual |  312.216429  2204  .141658997           R-squared     =  0.2722
-------------+------------------------------           Adj R-squared =  0.2673
       Total |  428.999484  2219  .193330097           Root MSE      =  .37638

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0770086   .0040714    18.91   0.000     .0690243    .0849928
       exper |   .0898502   .0079036    11.37   0.000     .0743509    .1053495
     expersq |  -.0024481   .0003967    -6.17   0.000    -.0032261   -.0016702
       black |  -.1761354   .0239043    -7.37   0.000    -.2230128   -.1292581
       south |   -.125071   .0312269    -4.01   0.000    -.1863083   -.0638338
        smsa |   .1376717   .0235462     5.85   0.000     .0914967    .1838468
      reg661 |  -.0865621   .0457195    -1.89   0.058    -.1762199    .0030956
      reg662 |  -.0020709   .0318752    -0.06   0.948    -.0645795    .0604378
      reg663 |   .0314867    .031107     1.01   0.312    -.0295154    .0924888
      reg664 |  -.0503983    .040855    -1.23   0.217    -.1305165      .02972
      reg665 |   .0036234   .0422329     0.09   0.932     -.079197    .0864438
      reg666 |   .0182858   .0488216     0.37   0.708    -.0774553    .1140269
      reg667 |   .0048968   .0459144     0.11   0.915    -.0851432    .0949367
      reg668 |  -.1557652   .0520945    -2.99   0.003    -.2579245   -.0536058
      smsa66 |   .0279434   .0227061     1.23   0.219    -.0165842     .072471
       _cons |   4.656564   .0833419    55.87   0.000     4.493128    4.820001
------------------------------------------------------------------------------
```

[1] Card. D. (1995). "Using geographic variation in college proximity to estimate the return to schooling," in *Aspects of Labour Market Bevhavior: Essays in Honour of John Vanderkamp,* ed. L.N. Christophides, E. K. Grant, and R. Swidinsky. Toronto: University of Toronto Press, 201-222.

Table 1.2: Reduced form education for education

```
      Source |       SS       df       MS              Number of obs =    2220
-------------+------------------------------           F( 18,  2201) =  115.63
       Model |  7221.94718    18  401.219288           Prob > F      =  0.0000
    Residual |  7636.97669  2201  3.46977587           R-squared     =  0.4860
-------------+------------------------------           Adj R-squared =  0.4818
       Total |  14858.9239  2219  6.69622527           Root MSE      =  1.8627

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      nearc2 |   .0180446    .087154     0.21   0.836     -.152868    .1889573
      nearc4 |   .2604735   .0983896     2.65   0.008     .0675272    .4534197
    motheduc |   .1324826   .0170677     7.76   0.000     .0990122    .1659531
    fatheduc |   .1111796   .0145968     7.62   0.000     .0825547    .1398045
       exper |  -.3805367   .0382972    -9.94   0.000    -.4556392   -.3054343
     expersq |   .0025954   .0019641     1.32   0.187    -.0012563     .006447
       black |  -.3459218   .1219798    -2.84   0.005    -.5851293   -.1067143
       south |  -.0518041   .1548235    -0.33   0.738    -.3554196    .2518113
        smsa |   .4218089   .1167867     3.61   0.000     .1927854    .6508325
      reg661 |  -.3795599   .2283522    -1.66   0.097    -.8273683    .0682485
      reg662 |  -.3169284   .1583069    -2.00   0.045    -.6273748    -.006482
      reg663 |  -.3542991   .1570864    -2.26   0.024    -.6623522    -.046246
      reg664 |  -.0814964   .2059201    -0.40   0.692    -.4853145    .3223218
      reg665 |  -.2797824   .2111526    -1.33   0.185    -.6938616    .1342969
      reg666 |  -.4014203   .2431572    -1.65   0.099    -.8782619    .0754213
      reg667 |  -.2318261   .2296505    -1.01   0.313    -.6821804    .2185282
      reg668 |   .0818341   .2624031     0.31   0.755    -.4327495    .5964177
      smsa66 |  -.2201582   .1174246    -1.87   0.061    -.4504328    .0101165
       _cons |   14.02289   .2995127    46.82   0.000     13.43554    14.61025
------------------------------------------------------------------------------

. test nearc2 nearc4 motheduc fatheduc ;

 ( 1)  nearc2 = 0
 ( 2)  nearc4 = 0
 ( 3)  motheduc = 0
 ( 4)  fatheduc = 0

       F(  4,  2201) =   65.48
            Prob > F =    0.0000

. predict res, res;
```

Table 1.3 Regression based Hausman test for endogeneity

```
      Source |       SS       df       MS              Number of obs =    2220
-------------+------------------------------           F( 16,  2203) =   51.88
       Model |  117.405539    16   7.3378462           Prob > F      =  0.0000
    Residual |  311.593945  2203  .141440738           R-squared     =  0.2737
-------------+------------------------------           Adj R-squared =  0.2684
       Total |  428.999484  2219  .193330097           Root MSE      =  .37609

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1017497   .0124755     8.16   0.000     .0772848    .1262147
       exper |   .1004833   .0093841    10.71   0.000     .0820808    .1188859
     expersq |   -.002493    .000397    -6.28   0.000    -.0032715   -.0017146
       black |  -.1549702   .0259292    -5.98   0.000    -.2058184    -.104122
       south |  -.1226742   .0312237    -3.93   0.000    -.1839053   -.0614432
        smsa |   .1244044   .0243632     5.11   0.000     .0766271    .1721816
      reg661 |   -.080592   .0457728    -1.76   0.078    -.1703544    .0091703
      reg662 |   .0056286   .0320614     0.18   0.861    -.0572452    .0685025
      reg663 |   .0411136   .0314199     1.31   0.191    -.0205022    .1027294
      reg664 |  -.0486601   .0408319    -1.19   0.233    -.1287332    .0314129
      reg665 |    .013062   .0424395     0.31   0.758    -.0701636    .0962876
      reg666 |   .0314252   .0491844     0.64   0.523    -.0650274    .1278778
      reg667 |   .0172291   .0462541     0.37   0.710     -.073477    .1079353
      reg668 |  -.1598693   .0520911    -3.07   0.002     -.262022   -.0577166
      smsa66 |   .0276992   .0226889     1.22   0.222    -.0167947    .0721931
         res |  -.0276853   .0131969    -2.10   0.036    -.0535649   -.0018056
       _cons |   4.232819   .2184829    19.37   0.000     3.804366    4.661273
------------------------------------------------------------------------------
```

Table 1.4: 2SLS estimates

```
. ivreg2 lwage  (educ= nearc2 nearc4 motheduc fatheduc) exper expersq black
south smsa reg661 reg662 reg663 reg664 reg665 reg666 reg667 reg668 smsa66,
endog(educ);

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =     2220
                                                      F( 15,  2204) =    34.95
                                                      Prob > F      =   0.0000
Total (centered) SS     =  428.9994844                Centered R2   =   0.2600
Total (uncentered) SS   =  88133.52155                Uncentered R2 =   0.9964
Residual SS             =  317.4474881                Root MSE      =    .3781

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1017497   .0125438     8.11   0.000     .0771643    .1263351
       exper |   .1004833   .0094355    10.65   0.000     .0819901    .1189765
     expersq |   -.002493   .0003991    -6.25   0.000    -.0032754   -.0017107
       black |  -.1549702   .0260712    -5.94   0.000    -.2060688   -.1038715
       south |  -.1226742   .0313948    -3.91   0.000    -.1842068   -.0611416
        smsa |   .1244044   .0244966     5.08   0.000     .0763919    .1724169
      reg661 |   -.080592   .0460235    -1.75   0.080    -.1707964    .0096124
      reg662 |   .0056286    .032237     0.17   0.861    -.0575548    .0688121
      reg663 |   .0411136    .031592     1.30   0.193    -.0208056    .1030328
      reg664 |  -.0486601   .0410555    -1.19   0.236    -.1291275    .0318072
      reg665 |    .013062   .0426719     0.31   0.760    -.0705735    .0966975
      reg666 |   .0314252   .0494538     0.64   0.525    -.0655024    .1283528
      reg667 |   .0172291   .0465074     0.37   0.711    -.0739237     .108382
      reg668 |  -.1598693   .0523764    -3.05   0.002    -.2625251   -.0572135
      smsa66 |   .0276992   .0228132     1.21   0.225    -.0170138    .0724122
       _cons |    4.23282   .2196795    19.27   0.000     3.802256    4.663383
------------------------------------------------------------------------------
Underidentification test (Anderson canon. corr. LM statistic):         236.081
                                                   Chi-sq(4) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                65.478
Stock-Yogo weak ID test critical values:  5% maximal IV relative bias    16.85
                                         10% maximal IV relative bias    10.27
                                         20% maximal IV relative bias     6.71
                                         30% maximal IV relative bias     5.34
                                         10% maximal IV size             24.58
                                         15% maximal IV size             13.96
                                         20% maximal IV size             10.26
                                         25% maximal IV size              8.31
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           6.556
                                                   Chi-sq(3) P-val =    0.0875
-endog- option:
Endogeneity test of endogenous regressors:                               4.426
                                                   Chi-sq(1) P-val =    0.0354
Regressors tested:    educ
------------------------------------------------------------------------------
Instrumented:         educ
Included instruments: exper expersq black south smsa reg661 reg662 reg663
reg664
```

```
                    reg665 reg666 reg667 reg668 smsa66
Excluded instruments: nearc2 nearc4 motheduc fatheduc
------------------------------------------------------------------------------

. ivendog;

Tests of endogeneity of: educ
H0: Regressor is exogenous
    Wu-Hausman F test:                 4.40102  F(1,2203)   P-value = 0.03603
    Durbin-Wu-Hausman chi-sq test:     4.42614  Chi-sq(1)   P-value = 0.03539
```

Table 1.5: 2SLS estimates excluding parents' education (dubious IVs)

```
. ivreg2 lwage  (educ= nearc2 nearc4 ) exper expersq black south smsa reg661
reg662 reg663 reg664 reg665
> reg666 reg667 reg668 smsa66, endog(educ);

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =     2220
                                                      F( 15,  2204) =    27.70
                                                      Prob > F      =   0.0000
Total (centered) SS     =  428.9994844                Centered R2   =   0.1739
Total (uncentered) SS   =  88133.52155                Uncentered R2 =   0.9960
Residual SS             =  354.3903925                Root MSE      =    .3995

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1472587   .0702897     2.10   0.036     .0094935    .2850239
       exper |    .120042   .0312972     3.84   0.000     .0587006    .1813834
     expersq |  -.0025757     .00044    -5.85   0.000     -.003438   -.0017134
       black |  -.1160388   .0651608    -1.78   0.075    -.2437517    .0116741
       south |  -.1182655   .0338386    -3.49   0.000     -.184588   -.0519431
        smsa |   .1000003   .0451679     2.21   0.027     .0114729    .1885278
      reg661 |  -.0696106   .0514015    -1.35   0.176    -.1703557    .0311345
      reg662 |   .0197911   .0402696     0.49   0.623    -.0591358    .0987181
      reg663 |   .0588214   .0428444     1.37   0.170     -.025152    .1427949
      reg664 |   -.045463   .0436489    -1.04   0.298    -.1310134    .0400873
      reg665 |   .0304235   .0522139     0.58   0.560    -.0719139    .1327609
      reg666 |   .0555939   .0638295     0.87   0.384    -.0695097    .1806974
      reg667 |   .0399133   .0599879     0.67   0.506    -.0776608    .1574875
      reg668 |  -.1674185   .0565124    -2.96   0.003    -.2781807   -.0566562
      smsa66 |   .0272501   .0241137     1.13   0.258    -.0200119    .0745121
       _cons |   3.453381   1.204836     2.87   0.004     1.091946    5.814816
------------------------------------------------------------------------------
Underidentification test (Anderson canon. corr. LM statistic):           8.394
                                                   Chi-sq(2) P-val =    0.0150
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                4.180
Stock-Yogo weak ID test critical values: 10% maximal IV size             19.93
                                         15% maximal IV size             11.59
                                         20% maximal IV size              8.75
                                         25% maximal IV size              7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
```

```
Sargan statistic (overidentification test of all instruments):           3.495
                                                     Chi-sq(1) P-val =    0.0615
-endog- option:
Endogeneity test of endogenous regressors:                               1.138
                                                     Chi-sq(1) P-val =    0.2861
Regressors tested:    educ
------------------------------------------------------------------------------
Instrumented:         educ
Included instruments: exper expersq black south smsa reg661 reg662 reg663
reg664
                      reg665 reg666 reg667 reg668 smsa66
Excluded instruments: nearc2 nearc4
------------------------------------------------------------------------------
```

SECTION 2: SAMPLE SIZE & IV ESTIMATION

1. Small sample & strong IVs vs. large sample & weak IVs

Model:

x = alpha*z + v2

y = beta*x + u1

No endogeneity. How well does the IV estimator do? Results from 200 simulations based on artificial data based on alpha=beta=1.

Case 1: Small sample (N=50), strong instrument (t-stat 1st stage = 6.0)

Variable	Obs	Mean	Std. Dev.	Min	Max
E(alpha_ols)	200	1.004607	.161436	.5434976	1.466404
E(beta_ols)	200	.9712768	.1705391	.5916569	1.518729
E(beta_iv)	200	.9688918	.2606616	.2876143	1.824383

Case 2: Large sample (N=2000), weak instrument (t-stat 1st stage = 2.0)

. sum store1 store2 store3

N=2000

Variable	Obs	Mean	Std. Dev.	Min	Max
E(alpha_ols)	200	1.019581	.5327355	-.4075418	2.487735
E(beta_ols)	200	.977441	.1674216	.5277573	1.43578
E(beta_iv)	200	-.5020311	12.56567	-136.0609	23.53888

SECTION 3: Too many instruments

2. Too many instruments

True model:

```
ge e2=std_v2*invnorm(uniform())
ge e1=std_e1*invnorm(uniform())

ge u1=e1+e2

ge x = 1*z + e2
ge y = 0*x + u1
```

where z, which is a valid and informative instrument, is drawn from std normal distribution.

True coefficient, denoted beta, on x is **zero**, but OLS is biased since x is correlated with u1. The plim of the OLS estimator is 0.5 here.

Now consider using as instruments for x 50 random variables w1,w2,…,w50 that are totally uncorrelated with x in theory. We do not use z (assume not available).

Question: how does the 2SLS estimator perform?

Variable	Obs	Mean	Std. Dev.	Min	Max
E(beta_ols)	200	.4927751	.0171272	.4467664	.5439443
E(beta_2sls)	200	.413747	.1071855	.1153944	.7246853
E(std error 2sls)	200	.1144694	.0110274	.0935858	.1683483
E(beta_LIML)	200	.035033	2.161731	-10.14505	12.20792
E(std error LIML)	200	1.731708	6.296244	.1315755	47.98719

=> 2SLS IS CLEARLY BIASED TOWARDS OLS! The Limited Information Maximum Likelihood (LIML) estimator appears much more robust in this context.

3. Same model as in (2) but with using only 5 instruments

Variable	Obs	Mean	Std. Dev.	Min	Max
E(beta_ols)	200	.4927751	.0171272	.4467664	.5439443
E(beta_2sls)	200	.3097683	.4376534	-1.724144	1.178116
E(std error 2sls)	200	.471942	.3071753	.2225561	3.96013
E(beta_LIML)	200	.2219852	11.76237	-120.5036	88.33043
E(std error LIML)	200	70.10669	564.0644	.2356987	6711.553

The Stata program generating these results can be found below.

```
/*

Illustration: Too many instruments

*/


clear
local N=2000
local seedn=457387+`N'
set seed `seedn'

set matsize 1600

set obs `N'
set more off

ge z=invnorm(uniform())
scalar std_v2 = 1
scalar std_e1 = 1


forvalues i = 1(1)50 {
generate w`i' = uniform()
}


local k=1

mat store=J(200,5,0)

qui{
while `k'<=200{

ge e2=std_v2*invnorm(uniform())
ge e1=std_e1*invnorm(uniform())

ge u1=e1+e2

ge x = 1*z + e2

ge y = 0*x + u1

if `k'==1 {
noi reg y x
mat store[`k',1]=_b[x]        /* ols coefficient */
noi ivreg2 y (x=w1-w50 )
mat store[`k',2]=_b[x]        /* iv coefficient */
mat V=e(V)
mat store[`k',3]=sqrt(V[1,1])       /* iv std error*/

noi ivreg2 y (x=w1-w50 ), liml
mat store[`k',4]=_b[x]        /* liml coefficient */
mat V=e(V)
mat store[`k',5]=sqrt(V[1,1])       /* liml std error*/

}

if `k'>1 {
reg y x
mat store[`k',1]=_b[x]        /* ols coefficient */
```

```
ivreg2 y (x=w1-w50)
mat store[`k',2]=_b[x]        /* iv coefficient */
mat V=e(V)
mat store[`k',3]=sqrt(V[1,1])       /* iv std error*/
ivreg2 y (x=w1-w50), liml
mat store[`k',4]=_b[x]        /* liml coefficient */
mat V=e(V)
mat store[`k',5]=sqrt(V[1,1])       /* liml std error*/

}

disp `k'
drop e1 e2 x y u1

local k=`k'+1
}
}
svmat store
/* note:
mean(store1) = E(b_ols)
mean(store2) = E(b_2sls)
mean(store3) = se(b_2sls)
mean(store4) = E(b_liml)
mean(store5) = se(b_liml)
*/

sum store1 store2 store3 store4 store5
```