# Applied Econometrics

## Lecture 2: Instrumental Variables, 2SLS and GMM

Måns Söderbom*

3 September 2009

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

## 1. Introduction

- Last time we talked about the unobservability problem in econometrics, and how this impacts on our ability to interpret regression results causally.
- We discussed how, under certain assumptions, a proxy variable approach can be used to mitigate or even eliminate the bias posed by (for example) omitted variables. As the name suggests, the proxy variable 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:

Wooldridge (2002), Chapters 5; 6.2; 8 and 14

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

Wooldridge, J.M. (2001) Applications of Generalized Method Moments Estimation, Journal of Economic Perspectives 15:4, pp.87-100.

In addition, there is a rather long chapter in Angrist & Pischke entitled "Instrumental variables in action", which we will discuss later in the course

## 2. Instrumental Variables: Motivation and Recap

- Population model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_K x_K + u, \tag{2.1}$$

where $E(u) = 0$, and $cov(x_j, u) = 0$, for $j = 1, 2, ..., K-1$ (from now on, we assume the "variable" $x_1$ is the constant), but where $x_K$ might be correlated with $u$, thus potentially endogenous, in which case OLS is inconsistent.

- If an **instrument** is available, the method of **instrumental variables (IV)** can be used to address the endogeneity problem, and provide consistent estimates of the structural parameters $\beta_j$.

- Note: We thus focus initially on the special case where there is **one** endogenous explanatory variable and **one** instrument.

- For the IV estimator to be consistent, the instrument $z_1$ has to satisfy two conditions:

1. The instrument must be exogenous, or **valid**:

$$cov(z_1, u) = 0.$$

This is often referred to as an **exclusion restriction**.

2. The instrument must be **informative**, or **relevant**. That is, the instrument $z_1$ must be correlated with the endogenous regressor $x_K$, conditional on all exogenous variables in the model (i.e. $x_2, ..., x_{K-1}$). That is, if we write the linear projection of $x_K$ onto all the exogenous variables,

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K, \tag{2.2}$$

where by definition of a linear projection error, $r_K$, is mean zero and uncorrelated with all the variables on the right-hand side, we require $\theta_1 \neq 0$.

- A corollary of these two conditions is that the instruments are not allowed to be explantory variables in the original equation.

- Hence, if $z_1$ is a valid and informative instrument, and $\beta_K \neq 0$, $z_1$ impacts on $y$ *but only indirectly, through the variable* $x_K$.

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

## 3. Identification & Definition

The **assumptions** above (validity and relevance) enable us to **identify** the parameters of the model. Loosely speaking, identification means that we can write the parameters in the structural model

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_K x_K + u,$$

in terms of **moments in observable variables**. Sticking to the example introduced, recall that we are happy to assume exogeneity for $x_2, ..., x_{K-1}$, so that

$$\begin{aligned} E(1 \cdot u) &= 0 \\ E(x_2 u) &= 0 \\ E(x_3 u) &= 0 \\ &\quad (...) \\ E(x_{K-1} u) &= 0, \end{aligned} \tag{3.1}$$

however we did **not** want to assume

$$E(x_K u) = 0,$$

because we suspect $x_K$ is endogenous: $E(x_K u) \neq 0$.

Now, if all we have are the moment conditions in (3.1), the parameters of the model are **not** identified. The reason is simple: with only $K-1$ moment conditions, we cannot solve for $K$ parameters. This model is therefore **underidentified**.

If the instrument $z_1$ is available (available = we have the data, and we believe the variable satisfies relevance and validity), we are in business, because the instrument validity assumption provides the additional moment condition

$$E(z_1 u) = 0.$$

Hence, using matrix notation as follows

$$\begin{aligned}\mathbf{x} &= \begin{bmatrix} 1 & x_2 & x_3 & ... & x_K \end{bmatrix} \\ \mathbf{z} &= \begin{bmatrix} 1 & x_2 & ... & x_{K-1} & z_1 \end{bmatrix},\end{aligned}$$

where each matrix element is a size $N$ column vector, we write the structural model as

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

and the moment conditions (or orthogonality conditions) as

$$E\left(\mathbf{z}'u\right) = \mathbf{0}.$$

Combining these two equations, we get

$$\begin{aligned} E\left(\mathbf{z}'u\right) &= \mathbf{0} \\ E\left(\mathbf{z}'\left(y - \mathbf{x}\boldsymbol{\beta}\right)\right) &= \mathbf{0} \\ E\left(\mathbf{z}'\mathbf{x}\right)\boldsymbol{\beta} &= E\left(\mathbf{z}'y\right), \end{aligned}$$

which is a system of $K$ linear equations (recall: $\boldsymbol{z}'$ is $K \times N$, $\boldsymbol{x}$ is $N \times K$, $\boldsymbol{\beta}$ is $K \times 1$, and $y$ is $N \times 1$). Provided the matrix $E\left(\mathbf{z}'\mathbf{x}\right)$ has full rank, i.e.

$$\text{rank } E\left(\boldsymbol{z}'\boldsymbol{x}\right) = K,$$

we can invert $E\left(\mathbf{z}'\mathbf{x}\right)$ and solve for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = \left[E\left(\mathbf{z}'\mathbf{x}\right)\right]^{-1} E\left(\mathbf{z}'y\right).$$

This solves for $K$ unknown parameters $\boldsymbol{\beta}$ from $K$ linear equations, hence this model is **exactly identified.**

While $\boldsymbol{\beta}$ is expressed here as a function of population moments, we can use sample moments ("data"; recall the analogy principle) to consistently estimate $\boldsymbol{\beta}$, provided we have a random sample of observations on $y, \mathbf{x}, z_1$. This defines the **instrumental variable estimator**:

$$\hat{\boldsymbol{\beta}}^{IV} = \left( N^{-1} \sum_{i=1}^{N} \mathbf{z}_i' \mathbf{x}_i \right)^{-1} \left( N^{-1} \sum_{i=1}^{N} \mathbf{z}_i' y_i \right),$$

or, in full matrix notation,

$$\hat{\boldsymbol{\beta}}^{IV} = \left( \boldsymbol{Z}'\boldsymbol{X} \right)^{-1} \left( \boldsymbol{Z}'Y \right), \tag{3.2}$$

where $\boldsymbol{Z}, \boldsymbol{X}, Y$ are data matrices.

- Whilst it is clear how the validity condition enabled us to identify the model, the role of the second condition - instrument relevance - may appear less clear. Recall that the instrument must be correlated with the endogenous explanatory variable, conditional on the other exogenous variables in the model.

- We need this condition, because otherwise the rank of $E(\mathbf{z}'\mathbf{x})$ will be less than $K$, and so the model would be underidentified. We skip the proof (problem 5.12 in Wooldridge provides some hints), because the intuition is very clear: if $\theta_1 = 0$ in

  $$x_K = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K,$$

  then that amounts to not having an instrument, in which case the model is underidentified as we have already seen.

You may want to be convinced that the IV estimator defined in (3.2) is consistent, under the assump-

tions we have made. Notice that

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta}+u)) \\ \hat{\boldsymbol{\beta}}^{IV} &= \boldsymbol{\beta}+(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'u).\end{aligned}$$

Using Slutsky's theorem, we get

$$\begin{aligned}p\lim\hat{\boldsymbol{\beta}}^{IV} &= \boldsymbol{\beta}+[E(\mathbf{Z}'\mathbf{X})]^{-1}E(\mathbf{Z}'u) \\ p\lim\hat{\boldsymbol{\beta}}^{IV} &= \boldsymbol{\beta},\end{aligned}$$

hence consistent: as the sample size $N$ goes to infinity, the IV estimator converges in probability to the true population value $\boldsymbol{\beta}$.

- *Student checkpoint*: Convince yourself - and ideally someone else too - that you are able to prove that for the model,

  $$y = \beta_1 + \beta_2 x_2 + u,$$

  where $x_2$ is endogenous and an instrument $z_1$ is available (satisfying the validity and relevance conditions above):

  $$x_2 = \theta_1 z_1 + r$$

  we can obtain the IV estimate of $\beta_2$ by means of a two-stage procedure:

  1. Regress the endogenous variable $x_2$ on the instrument $z_1$ using OLS. Calculate the predicted values of $x_2$.
  2. Use the predicted values (instead of the actual values) of $x_2$ from the first regression as the explanatory variable in the structural equation, and estimate using OLS. The resulting estimate of the coefficient on predicted $x_2$ is the IV estimate of $\beta_2$. Interpret this in terms of 'purging' the endogenous variable of the correlation with the residual.

- Notice that if I use $x_2$ as its own instrument in the first stage (i.e. $z_1 = x_2$), I obtain OLS estimates in the second stage. So in a sense, OLS can actually be viewed as an IV estimator in which all variables are assumed exogenous.

As already discussed, the validity and relevance conditions are equally important in identifying $\beta_2$. There is one important difference between them, however:

- The relevance condition can be tested, for example by computing the $t$-statistic associated with $\hat{\theta}_1$ in the reduced form (first stage) regression.

- The validity condition, however, cannot be tested, because the condition involves the unobservable residual $u$. Therefore, this condition has to be taken on faith, which is why relating the validity condition 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.

[EXAMPLE: Earnings, education and distance to school - Section 1 in the appendix]

## 4. Multiple Instruments: Two-Stage Least Squares

- We considered above the simple IV estimator with one endogenous explanatory variable, and one instrument. As already noted, this is a case of **exact identification**. Similarly, if you have two endogenous explanatory variables and two instruments, the model is again exactly identified.

- If you have less instruments than endogenous regressors, the model is **underidentified**.

- 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.

- Suppose we have $M$ instrumental variables for $x_K$: $z_1, z_2, ..., z_M$. Suppose each of these instruments satisfies the validity condition

$$cov\,(z_h, u) = 0,$$

  for all $h$. If each of these has some partial correlation with $x_K$ (relevance condition), we could then in principle compute $M$ different IV estimators.

- Of course, that's neither practical nor efficient.

- Theorem 5.3 in Wooldridge asserts that the **Two-Stage Least Squares (2SLS)** estimator is the most efficient IV estimator. The 2SLS estimator is obtained by using **all** the instruments simultaneously in the first stage regression:

$$x_K = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \theta_2 z_2 + ... + \theta_M z_M + r_K.$$

By definition, the OLS estimator of the first stage regression will construct the **linear combination** of the instruments most highly correlated with $x_K$. By assumption all the instruments are exogenous, hence this procedure retains more exogenous variation in $x_K$ 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 $x_K$:

$$\hat{x}_K = \hat{\delta}_1 + \hat{\delta}_2 x_2 + ... + \hat{\delta}_{K-1} x_{K-1} + \hat{\theta}_1 z_1 + \hat{\theta}_2 z_2 + ... + \hat{\theta}_M z_M,$$

  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 $\theta_j$ in the first stage has to be nonzero for the model to be identified.

You might 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.

### 4.1. 2SLS: The General Case

So far we have focussed on the case where there is only one endogenous explanatory variable. In my view, this is a useful approach for studying the 2SLS estimator, because the main mechanisms carry over to the more general case with several endogenous explanatory variables. Therefore I will discuss the general case rather briefly - you can refer to Section 5.2 in Wooldridge for details.

- The validity and relevance conditions in the general case, where several elements of $x$ may be correlated with $u$, are as follows:

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

$$\text{rank}(\boldsymbol{z}'\boldsymbol{x}) = K, \qquad \text{(Relevance)}$$

  where $\boldsymbol{z}$ is $1 \times L$, and $\text{rank}(\boldsymbol{z}'\boldsymbol{z}) = L$, ruling out collinearity amongst the instruments. In this notation, any exogenous element of $\boldsymbol{x}$, including a constant, are included in $\boldsymbol{z}$.

- The validity condition is straightforward to understand, but the relevance condition perhaps is not. Clearly for the relevance condition, stated here as a **rank condition**, to hold, we need at least as many instruments as there are explanatory variables: $L \geq K$. This is known as the **order condition**. However, whilst necessary, $L \geq K$ is not sufficient for $\text{rank}(\boldsymbol{z}'\boldsymbol{x}) = K$: the elements of $\boldsymbol{z}$ must also be appropriately correlated with the elements of $\boldsymbol{x}$.

- Testing the rank condition formally is tedious and somewhat involved, and so we will not go into details here (neither does Wooldridge). It is useful, of course, to look carefully at the first stage results. We will have as many first-stage regressions as there are endogenous explanatory variables, and you need at least one significant coefficient on the instruments in each reduced form regression for the model to be well identified. This is a necessary, not sufficient condition, however. To see this, consider the model

$$y = \beta_1 + \beta_2 x_2 + \beta_2 x_3 + \beta_3 x_4 + u,$$

  where $x_3$ and $x_4$ are endogenous. We need at least two instruments, say $z_1$ and $z_2$, and these enter

in the reduced form equations for $x_3$ and $x_4$:

$$\begin{aligned} x_3 &= \pi_1 + \pi_2 x_2 + \pi_3 z_1 + \pi_4 z_2 + \varepsilon_1, \\ x_4 &= \gamma_1 + \gamma_2 x_2 + \gamma_3 z_1 + \gamma_4 z_2 + \varepsilon_2. \end{aligned}$$

- If $\pi_3 = 0, \pi_4 \neq 0, \gamma_3 = 0, \gamma_4 \neq 0$, the structural equation is **not** identified, because the instrument $z_1$ is irrelevant in both equations - hence, effectively we only have one instrument.
- If $\pi_3 = 0, \pi_4 \neq 0, \gamma_3 \neq 0, \gamma_4 = 0$, the structural equation **is** identified, because the instrument $z_1$ is relevant in the equation determining $x_4$, while the $z_2$ is relevant for $x_3$.

From a practical point of view, you will almost certainly notice if identification fails. If your model is literally not identified, because you have too few instruments or because the instruments are collinear, then Stata will report this and stop. If your instruments are very weakly correlated with the endogenous explanatory variables, the coefficients on the instruments in the first stage may be insignificant, and the 2SLS standard errors very large - correctly telling you that you are not learning anything from the current model.

**General expression for the 2SLS estimator.**

- The algebra of the 2SLS estimator is more involved than that of the IV estimator. Using matrix algebra helps us understand the general mechanisms. Recall that, for the IV estimator, we have

$$\hat{\boldsymbol{\beta}}^{IV} = \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \boldsymbol{Z}'Y. \tag{4.1}$$

It is straightforward to show that this expression can be expressed as

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

i.e. OLS using predicted instead of actual values of the explanatory variables (for the exogenous

variables in $X$, predicted and actual values coincide, of course).

- The same expression holds for 2SLS:

$$\hat{\boldsymbol{\beta}}^{2SLS} = \left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}\hat{\boldsymbol{X}}'Y, \tag{4.2}$$

however because the model is overidentified this does not give an expression for 2SLS equivalent to (4.1). To see what we get if we 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 (4.2):

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

A common way of writing this is as

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

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

## 5. Generalized Method of Moments

References: Chapters 8 and 14 in Wooldridge (2002). Wooldridge, J.M. (2001) Applications of Generalized Method Moments Estimation, Journal of Economic Perspectives 15:4, pp.87-100.

- Consider the usual linear model:

$$y_i = \boldsymbol{X}_i\boldsymbol{\theta} + u_i, \tag{5.1}$$

  where $\boldsymbol{X}_i$ is a $1 \times P$ matrix of explanatory variables, $\boldsymbol{\theta}$ is a $P \times 1$ vector of parameters, and $y_i$ and $u_i$ are scalars.

- We have seen how the OLS estimator and the 2SLS estimator can be derived from a set of moment conditions of the following form:

$$E\left(\mathbf{Z}_i' u_i\right) = 0,$$

  where $\mathbf{z}$ is a vector of instruments (OLS is obtained by setting $\mathbf{Z}_i = \mathbf{X}_i$). This way of deriving these estimators fits very well into the framework of **Generalized Method of Moments (GMM)**.

- The formalization of GMM is usually attributed to Hansen (1982). Hansen showed that every previously suggested instrumental variables estimator, in linear or nonlinear models, with cross-section, time series or panel data, could be cast as a GMM estimator. GMM is therefore sometimes viewed as a **unifying framework** for inference in econometrics.

- In this section I derive the general expression for the GMM estimator of the linear model, and show how GMM relates to conventional estimators like OLS and 2SLS. A discussion of how GMM can be used to test hypotheses of interest is provided below. I also show some empirical examples. I do **not** provide a complete rigorous theoretical treatment of GMM however, because that would take me too long. If you want to understand the theoretical foundations of the GMM estimator, please consult Chapter 14 in Wooldridge (2002).

### 5.1. The basis for the GMM estimator: Moment functions

- GMM is based on **moment functions** that depend on observable random variables and unknown parameters, and that have **zero expectation** in the population when evaluated at the true parameters. Adopting the general notation of Wooldridge (Chapter 14), this is formalized as

$$E\left[\boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)\right] = 0,$$

  where $\boldsymbol{g}$ denotes the moment function, $\boldsymbol{w}_i$ is a vector of observable random variables (e.g. our instruments) and $\boldsymbol{\theta}$ is a $P \times 1$ vector of unknown parameters. The moment function $\boldsymbol{g}$ can be linear or nonlinear.

- In **linear models** a natural way of writing the moment function is as $\boldsymbol{Z}_i'\left(y_i - \boldsymbol{X}_i\boldsymbol{\theta}\right)$, since

$$E\left[\boldsymbol{Z}_i'\left(y_i - \boldsymbol{X}_i\boldsymbol{\theta}\right)\right] = 0 \tag{5.2}$$

  often can be related to theory (or at least theoretical reasoning). This is also exactly what the moment conditions underlying 2SLS look like.

- Provided we have a random sample, we can appeal to the **analogy principle** and replace population moments by sample moments. This enables us to estimate $\boldsymbol{\theta}$ based on the data.

- If the model is **exactly identified**, so that $L = P$, where $L$ is the number of instruments, there are $L$ moment conditions in (5.2) and they all **hold exactly** (because we solve for $P$ parameters based on $L = P$ equations). As we have already seen, this gives either the IV estimator or the OLS estimator, depending on how the moments are written.

- If the model is **overidentified**, $L > P$, then in general there is **no unique solution** to (5.2), because not all $L$ sample moments corresponding to (5.2) will hold exactly - this is simply because there are "too many" equations.

- Hansen (1982) proposed a solution to this problem, based on bringing the sample moments as close to zero as possible. This is achieved by **minimizing the quadratic form**

$$\min_{\theta} \underset{(1 \text{ x } L)}{\left[\sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)\right]'} \cdot \underset{(L \text{ x } L)}{C} \cdot \underset{(L \text{ x } 1)}{\left[\sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)\right]}$$

with respect to the parameters $\boldsymbol{\theta}$, where $C$ is a positive definite $L \times L$ weighting matrix (more on the latter below). It can be shown that this yields a **consistent estimator** of $\boldsymbol{\theta}$, under certain regularity conditions (see Theorem 14.1 in Wooldridge, 2002).

- Provided that the moment functions are continuously differentiable, the GMM estimator satisfies the **first-order condition**

$$\underset{(P \text{ x } L)}{\left[\sum_{i=1}^{N} \nabla_{\theta} \boldsymbol{g}\left(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}\right)\right]'} \cdot \underset{(L \text{ x } L)}{C} \cdot \underset{(L \text{ x } 1)}{\left[\sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}\right)\right]} = \underset{(P \text{ x } 1)}{\mathbf{0}}, \tag{5.3}$$

where

$$\nabla_{\theta} \boldsymbol{g}\left(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}\right) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \boldsymbol{g}\left(\boldsymbol{w}_i; \hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_P\right)' \\ \frac{\partial}{\partial \theta_2} \boldsymbol{g}\left(\boldsymbol{w}_i; \hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_P\right)' \\ (...) \\ \frac{\partial}{\partial \theta_P} \boldsymbol{g}\left(\boldsymbol{w}_i; \hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_P\right)' \end{bmatrix}$$

is an $L \times P$ vector of derivatives of the moment function $g$ with respect to the first, second, etc. element of the parameter vector $\boldsymbol{\theta}$, (remember: $\boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)$ is $L \times 1$, hence $L \times P$).

- Let's write this for the linear model that we are familiar with. We have

$$\boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) = \boldsymbol{Z}_i' \left(y_i - \boldsymbol{X}_i \boldsymbol{\theta}\right),$$

hence

$$\begin{aligned}
\frac{\partial}{\partial\theta_1}\boldsymbol{g}\left(\boldsymbol{w}_i;\theta_1,\theta_2,...,\theta_P\right) &= -\boldsymbol{Z}_i'x_{1i} \qquad (L\times 1)\\
\frac{\partial}{\partial\theta_2}\boldsymbol{g}\left(\boldsymbol{w}_i;\theta_1,\theta_2,...,\theta_P\right) &= -\boldsymbol{Z}_i'x_{2i} \qquad (L\times 1)\\
&\qquad (...)\\
\frac{\partial}{\partial\theta_P}\boldsymbol{g}\left(\boldsymbol{w}_i;\theta_1,\theta_2,...,\theta_P\right) &= -\boldsymbol{Z}_i'x_{Pi} \qquad (L\times 1),
\end{aligned}$$

and so

$$\nabla_\theta \boldsymbol{g}\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right) = \left[\begin{array}{cccc} -\boldsymbol{Z}_i'x_{1i} & -\boldsymbol{Z}_i'x_{2i} & (...) & -\boldsymbol{Z}_i'x_{Pi}\end{array}\right] = -\boldsymbol{Z}_i'\boldsymbol{X}_i.$$

- The first-order condition becomes

$$\left[\sum_{i=1}^{N}\boldsymbol{Z}_i'\boldsymbol{X}_i\right]'\cdot C\cdot\left[\sum_{i=1}^{N}\boldsymbol{Z}_i'\left(y_i-\boldsymbol{X}_i\hat{\boldsymbol{\theta}}\right)\right]=\mathbf{0}.$$

It follows that

$$\left[\sum_{i=1}^{N}\boldsymbol{Z}_i'\boldsymbol{X}_i\right]'\cdot C\cdot\left[\sum_{i=1}^{N}\boldsymbol{Z}_i'y_i\right]=\left[\sum_{i=1}^{N}\boldsymbol{Z}_i'\boldsymbol{X}_i\right]'\cdot C\cdot\left[\sum_{i=1}^{N}\boldsymbol{Z}_i'\boldsymbol{X}_i\hat{\boldsymbol{\theta}}\right],$$

hence the solution for $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}}=\left(\left[\sum_{i=1}^{N}\boldsymbol{Z}_i'\boldsymbol{X}_i\right]'\cdot C\cdot\sum_{i=1}^{N}\boldsymbol{Z}_i'\boldsymbol{X}_i\right)^{-1}\left[\sum_{i=1}^{N}\boldsymbol{Z}_i'\boldsymbol{X}_i\right]'\cdot C\cdot\left[\sum_{i=1}^{N}\boldsymbol{Z}_i'y_i\right],$$

which can be written in data matrices as:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{GMM} &= \left(\left(\boldsymbol{Z}'\boldsymbol{X}\right)'\cdot C\cdot\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)'\cdot C\cdot\boldsymbol{Z}'y.\\
\hat{\boldsymbol{\theta}}^{GMM} &= \left(\left(\boldsymbol{X}'\boldsymbol{Z}\right)\cdot C\cdot\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\right)\cdot C\cdot\boldsymbol{Z}'y.
\end{aligned}$$

- This **defines** the GMM estimator for the linear model.

- For nonlinear models it is typically not possible to solve for $\hat{\boldsymbol{\theta}}^{GMM}$ analytically, in which case numerical optimization methods can be used to find the parameter vector that satisfies the first-order conditions (5.3) (cf. MLE)

- This illustrates the point that the GMM framework is general indeed.

### 5.2. The Linear Model: GMM and Conventional Estimators

- We just saw that, for the linear model,

$$\hat{\boldsymbol{\theta}}^{GMM} = \left(\left(\boldsymbol{X}'\boldsymbol{Z}\right) \cdot C \cdot \left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{Z}\right) \cdot C \cdot \boldsymbol{Z}'y.$$

- Notice that if $L = P$ (just identified) then $\boldsymbol{X}'\boldsymbol{Z}$ is a **square** matrix ($C$ is always square), and so we have

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

  The GMM estimator therefore reduces to

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

  which **is** the IV estimator (cf. eq (1.3) in lecture 2). So you see the IV estimator fits in the GMM framework.

- You also see that in this case the matrix $C$ **plays no role**.

- Of course, if $\boldsymbol{Z} = \boldsymbol{X}$, then the GMM estimator for the linear model coincides with the OLS estimator:

$$\hat{\boldsymbol{\theta}}^{GMM} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'y.$$

  Again, $C$ plays no role.

- In contrast, in the **overidentified** case ($L > P$), the choice of the weight matrix $C$ is **important**. I have already announced (without proof) that the GMM estimator is consistent - and this happens to be true **for any** $C$, provided that $C$ is positive definite. In other words, if all you are worrying about is consistency of your estimator, then you can use **any matrix you want**, as long as it is positive definite.

- However, the choice of $C$ is nevertheless important, because in finite samples different $C$-matrices will lead to different point estimates!

- Consider first

$$C = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}.$$

We then have

$$\hat{\boldsymbol{\theta}}^{GMM} = \Big(\left(\boldsymbol{X}'\boldsymbol{Z}\right)\cdot\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\cdot\left(\boldsymbol{Z}'\boldsymbol{X}\right)\Big)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\right)\cdot\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\cdot\mathbf{Z}'y,$$

which is the **2SLS** estimator. So the 2SLS estimator can also be cast as a GMM estimator.

- If different choices of $C$ lead to different results in practice, you might think this somewhat unsatisfactory - two researchers with the same model and same data will report different results if they choose different $C$-matrices, and both researchers can claim they are using a consistent estimator.

- Fortunately, there is a way of choosing among all the possibilities. It can be shown that, among all possible candidates for the weight matrix $C$, the "best" one is the **inverse of the covariance of the moments**:

$$C = \left[Var\left(\mathbf{Z}_i'u_i\right)\right]^{-1}.$$

This is the "best" choice because this weight matrix produces the GMM estimator with the smallest variance, asymptotically (see Section 8.3.3 in Wooldridge, 2002). Referring back to the minimization problem (now written specifically for the linear regression model):

$$\min_{\theta}\underset{(1\text{ x L})}{\left[\sum_{i=1}^{N}\mathbf{Z}_i'u_i\right]'}\cdot\underset{(\text{L x L})}{C}\cdot\underset{(\text{L x 1})}{\left[\sum_{i=1}^{N}\mathbf{Z}_i'u_i\right]}$$

this actually has some intuitive appeal: low-variance moments will be given higher weight in the criterion function than high-variance moments.

- The problem is that the inverse of the covariance matrix for the moments (for the population, of course) is **unknown**, and so an estimator based on the covariance matrix for the moments in the population is **infeasible**.

- The feasible solution is to **estimate** this covariance matrix using a judicious, pre-specified, choice for $C$. Remember any positive definite $C$ achieves consistency, and so if I can estimate $\boldsymbol{\theta}$ consistently I can also estimate the covariance matrix of the moments consistently.

- This line of reasoning suggests the following 2-step procedure for obtaining a consistent and asymptotically efficient GMM estimator (see Wooldridge, 2002, p. 193).

**STEP 1**: Obtain a "preliminary" GMM estimate of $\boldsymbol{\theta}$, denoted $\breve{\boldsymbol{\theta}}$, using a suitable weight matrix $C_1$. Typically, the 2SLS estimator is used to produce $\breve{\boldsymbol{\theta}}$, which, as we have seen, amounts to using $C = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$. Then predict the residuals:

$$\breve{u}_i = y_i - \boldsymbol{X}_i'\breve{\boldsymbol{\theta}}.$$

With these in hand, we can now consistently estimate the covariance of the moments:

$$\hat{Var}\left(\mathbf{Z}_i'u_i\right) \equiv \hat{\Lambda} = \left(N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\breve{u}_i\breve{u}_i'\mathbf{Z}_i\right).$$

**STEP 2**: Use the inverse of the covariance of the moments as the weight matrix, and re-estimate the model to obtain an asymptotically efficient (optimal), and consistent, GMM estimator:

$$\hat{\boldsymbol{\theta}}^{GMM} = \left(\left(\boldsymbol{X}'\boldsymbol{Z}\right)\cdot\left(\hat{\Lambda}\right)^{-1}\cdot\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\right)\cdot\left(\hat{\Lambda}\right)^{-1}\cdot\boldsymbol{Z}'y. \tag{5.4}$$

- *Student checkpoint:* Suppose the residuals are homoskedastic, and suppose you have a linear single-equation model. What are you gaining by using GMM compared to 2SLS?

- [EXAMPLE: Ivreg2, GMM, optimal weight matrix - section 2 in the appendix.]

### 5.3. Why use GMM?

- What are the relative merits of GMM compared to the conventional estimators? That is, why would you ever use GMM? Let me now try to shed some light on this.[1] I focus on the linear regression model only.

- First, as should be clear now, if the moments we are using are of the form

$$E\left(\mathbf{Z}_i' u_i\right) = 0,$$

and the model is exactly identified, the GMM estimator coincides with the IV estimator (or, if $\boldsymbol{Z} = \boldsymbol{X}$, with OLS). Hence, there can be no advantages to using GMM for the linear model in the exactly identified case.

- For overidentified models, we have seen that the GMM estimator is asymptotically efficient if the weight matrix is optimal:

$$\hat{\boldsymbol{\theta}}^{GMM} = \left( \left(\boldsymbol{X}'\boldsymbol{Z}\right) \cdot \left(\hat{\Lambda}\right)^{-1} \cdot \left(\boldsymbol{Z}'\boldsymbol{X}\right) \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{Z}\right) \cdot \left(\hat{\Lambda}\right)^{-1} \cdot \boldsymbol{Z}' y,$$

with

$$\Lambda = E\left(\mathbf{Z}_i' u_i u_i' \mathbf{Z}_i\right).$$

However, if the error term is homoskedastic (and, if there is a time dimension in the data, non-autocorrelated), this reduces to

$$\Lambda = \sigma_u^2 E\left(\mathbf{Z}_i' \mathbf{Z}_i\right),$$

and so once we replace $E\left(\mathbf{Z}_i' \mathbf{Z}_i\right)$ by $N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Z}_i$ we get the 2SLS estimator. So in that case (homoskedasticity) also there is no reason to "use GMM" rather than 2SLS (since they are the same).

[1] See Wooldridge, J.M. (2001) Applications of Generalized Method Moments Estimation, Journal of Economic Perspectives 15:4, pp.87-100.

- It follows from the above that GMM will be an improvement over 2SLS if $N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Z}_i$ is **not** the the optimal weight matrix.

- If there is heteroskedasticity, for example, the covariance of the moments is **not** $\sigma_u^2 E\left(\mathbf{Z}_i'\mathbf{Z}_i\right)$, and so 2SLS is not asymptotically efficient. By estimating the covariance of the moments in the first step and then computing the weight matrix as $\hat{\Lambda} = \left(N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\breve{u}_i\breve{u}_i'\mathbf{Z}_i\right)$ we can expect the GMM estimator to be more efficient than 2SLS.

- Similarly, in time series applications in which there is serial correlation in the error terms, the covariance of the moments is not $\sigma_u^2 E\left(\mathbf{Z}_t'\mathbf{Z}_t\right)$. Calculating the covariance matrix of the moments from the data based on an initial estimator, and using that to form the weight matrix, may thus be more efficient than doing 2SLS.

- Whenever one estimates a **system** of equations, the benefits of GMM estimation become more apparent. We will see that when studying the Arellano-Bond (1991) and Blundell-Bond (1998) models for dynamic panel data models.

**Potential drawbacks:**

- Definition of the weight matrix for the first step is **arbitrary**, and different choices will lead to different point estimates in the second step

  - One possible remedy is to not stop after two iterations, but continue to update the weight matrix until some sort of convergence has been achieved. This estimator can be obtained by using the cue (continuously updated estimators) options within ivreg2

- Inference problems because the optimal weight matrix is **estimated**. This can lead to sometimes severe downward bias in the estimated standard errors for the GMM estimator - see for example the Monte Carlo evidence reported by Arellano and Bond (1991).

  - Frank Windmeijer has proposed a method that appears to work well however. This is not

available (I think) in ivreg2, but it is available in xtabond2. We will return to this in the lectures on dynamic panel data models.

## 6. Variance and Efficiency

### 6.1. Variance of the 2SLS estimator

Recall the 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

$$\hat{Avar}\left(\hat{\boldsymbol{\beta}}^{2SLS}\right) = \hat{\sigma}^2\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}, \tag{6.1}$$

where

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

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

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

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 $\rho_{xz}$ 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 (6.1) 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.

2. Use a Generalized Method of Moments (GMM) estimator.

[EXAMPLE: The Card (1995) data - Section 3 in the appendix.]

### 6.2. Variance of the GMM estimator

- Suppose that our weight matrix $C$ is "optimal" (asymptotically efficient), thus constructed as the inverse of the covariance of the moments:

$$C = \left[Var\left(\mathbf{Z}_i' u_i\right)\right]^{-1}.$$

  In practice, we estimate this as follows:

$$\hat{Var}\left(\mathbf{Z}_i' u_i\right) \equiv \hat{\Lambda} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \breve{u}_i \breve{u}_i' \mathbf{Z}_i\right).$$

  where the $\breve{u}_i$ are the residuals from the first step (preliminary estimator).

- For the linear model we have

$$\hat{\boldsymbol{\theta}}^{GMM} = \left(\left(\boldsymbol{X}'\boldsymbol{Z}\right) \cdot \left(\hat{\Lambda}\right)^{-1} \cdot \left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{Z}\right) \cdot \left(\hat{\Lambda}\right)^{-1} \cdot \boldsymbol{Z}' y.$$

- The formula for the variance of $\hat{\boldsymbol{\theta}}^{GMM}$ is relatively straightforward, namely:

$$V\left(\hat{\boldsymbol{\theta}}^{GMM}\right) = \left[\left[\sum_{i=1}^{N} \nabla_{\theta} \boldsymbol{g}\left(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}\right)\right]' \left(\hat{\Lambda}\right)^{-1} \left[\sum_{i=1}^{N} \nabla_{\theta} \boldsymbol{g}\left(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}\right)\right]\right]^{-1},$$

  in general (check Wooldridge, pp. 423-424 for its origins) which reduces to

$$V\left(\hat{\boldsymbol{\theta}}^{GMM}\right) = \left[\left(\boldsymbol{X}'\boldsymbol{Z}\right)\left(\hat{\Lambda}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right]^{-1},$$

  for the linear model.

- An important issue in finite samples: This formula is derived under the assumption that the weight matrix is non-stochastic. In practice, the weight matrix is actually 'noisy', since the residuals in the first stage are affected by sampling error. The upshot is that step 2 standard errors tend to be too

good. Methods now exist that enable you to correct for sampling error in the first step (e.g. the Windmeijer procedure, commonly used these days when estimating dynamic panel data models).

## 7. Testing for exogeneity and validity of overidentifying restrictions

Reference: Wooldridge (2002), Chapter 6.

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

### 7.1. Testing for exogeneity: 2SLS

- 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 to me, namely

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

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.

### 7.1.1. 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 very hard to interpret given that $H$ is non-negative in theory (follows a Chi-squared distribution under the null).

### 7.1.2. A regression-based Hausman test

Hausman has also derived a regression-based form of the test just outlined, that 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. $E(y_2u_1) = 0$, then it **must** be that $E(v_2u_1) = 0$, 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. $E(y_2u_1) \neq 0$, then it **must** be that $E(v_2u_1) \neq 0$, 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

$$H_0 : E(v_2u_1) = 0,$$

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

$$y_1 = \boldsymbol{z}_1 \boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 \hat{v}_2 + error, \tag{7.1}$$

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, this gives estimates of the parameters $\boldsymbol{\delta}_1$,$\alpha_1$ that are *numerically identical to* 2SLS - as we shall see later in this course, this is quite useful. There is one minor issue though: the OLS standard errors associated with (7.1) 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.

### 7.2. 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. This relates directly to the GMM framework. Let's turn to this now.

### 7.3. Specification Tests with GMM

- GMM estimation offers a very general method for specification testing, based on the minimized value of the criterion function with and without the restrictions imposed There is a strong similarity between this **criterion based test** and the log likelihood ratio test after maximum likelihood estimation (more on MLE later in the course). The approach is as follows.

- Suppose that we have obtained the optimal GMM estimator $\hat{\boldsymbol{\theta}}^{GMM}$ defined in (5.4). This may be your **unrestrictive estimator**, or your **restrictive estimator,** depending on what you are doing.

- Suppose you want to investigate whether you're 'allowed' to rely on a more restrictive model, nested in your current model (e.g. one where you omit a sub-set of the explanatory variables).

- In this case your current model is the unrestrictive estimator.

- Now impose $Q$ restrictions on the original model, and re-estimate the parameters using GMM with those restrictions imposed. Use the same set of instruments and the same weight matrix as for the unrestricted model. Refer to this as the **restrictive estimator**, denoted $\hat{\boldsymbol{\theta}}_R^{GMM}$.

- Now suppose we want to test whether the restrictions imposed underlying $\hat{\boldsymbol{\theta}}_R^{GMM}$ are valid ones. This we can easily do by comparing the values of the **criterion function** for $\hat{\boldsymbol{\theta}}^{GMM}$ and $\hat{\boldsymbol{\theta}}_R^{GMM}$. It can be shown that, under the null hypothesis that the restrictions are valid, the **GMM distance statistic** is distributed as chi-square with degrees of freedom equal to the number of restrictions imposed on the general model (i.e. $Q$):

$$J = \left[ \left( \sum_{i=1}^{N} \boldsymbol{Z}_i' \hat{u}_{i(R)} \right)' \hat{C}_2 \left( \sum_{i=1}^{N} \boldsymbol{Z}_i' \hat{u}_{i(R)} \right) - \left( \sum_{i=1}^{N} \boldsymbol{Z}_i' \hat{u}_i \right)' \hat{C}_2 \left( \sum_{i=1}^{N} \boldsymbol{Z}_i' \hat{u}_i \right) \right] / N \sim \chi^2_Q, \qquad (7.2)$$

where

$$\hat{u}_{i(R)} = y_i - \boldsymbol{X}_i \hat{\boldsymbol{\theta}}_R^{GMM},$$

is the vector of residuals based on the restrictive model, and

$$\hat{u}_i = y_i - \boldsymbol{X}_i \hat{\boldsymbol{\theta}}^{GMM},$$

is the vector of residuals based on the unrestrictive model.

- Because constrained optimization (i.e. minimization subject to the $Q$ restrictions imposed) cannot result in a smaller objective function than unconstrained minimization, equation (7.2) is always nonnegative, and usually strictly positive.

- This approach can be used to test for exogeneity and the validity of overidentifying restrictions, as well as for exogeneity (among other things).

### 7.3.1. Testing for exogeneity

- Suppose we suspect that some of the explanatory variables are endogenous - for convenience the last $p$ columns in $\boldsymbol{X}_i$.

- In this case, the **unrestrictive** model does **not** use the $p$ moment conditions

$$E\left[\check{\boldsymbol{X}}_i' \left(y_i - \boldsymbol{X}_i \boldsymbol{\theta}_o\right)\right] = 0, \tag{7.3}$$

where $\check{\boldsymbol{X}}_i'$ are the dubious instruments, whereas the **restrictive** model **does** use them (it's restrictive because it imposes the restriction that these potentially endogenous variables are in fact exogenous, i.e. orthogonal to the residual).

- Hence, we can test for exogeneity by comparing criterion values with and without the "dubious" moments (7.3) using the distance statistic (7.2).

- EXAMPLE: OVERID restrictions and exogeneity - see appendix.

### 7.3.2. Testing validity of overidentifying restrictions

- Suppose we have more instruments than explanatory variables. How can we use the approach above to test for the validity of the overidentifying restrictions? In particular, what does the unrestrictive and restrictive models look like?

- Clearly the fact that we have more instruments than explanatory variables implies restrictions: we are imposing the restrictions that $Q = L - P > 0$ variables do not belong in the structural equation. So this model is the restrictive model.

- Now consider relaxing that restriction, by adding $Q$ instruments to the set of explanatory variables in the model (remember, we can't add all the instruments, as you need as many exclusion restrictions as there are endogenous explanatory variables to identify the model). That model, clearly, would be exactly identified. Now, we know that in the exactly identified case all sample moments hold exactly:

$$N^{-1}\left(\sum_{i=1}^{N} \boldsymbol{Z}_i'\hat{u}_{i(R)}\right) = \mathbf{0},$$

  and so the criterion value of the unrestrictive model is exactly zero.

- This implies that if the null hypothesis is that the overidentifying restrictions are valid, the test statistic reduces to

$$J = \left(\sum_{i=1}^{N} \boldsymbol{Z}_i'\hat{u}_{i(R)}\right)' \hat{C}_2 \left(\sum_{i=1}^{N} \boldsymbol{Z}_i'\hat{u}_{i(R)}\right)/N \sim \chi^2_Q.$$

  Notice that this is simply the minimized value of the criterion function for the overidentified model.

## 8. 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.

### 8.1. 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.

### 8.2. 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:

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

**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 4 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 finate 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 5 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 →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

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

  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.

[3] To derive this formula you need to know a few matrix tricks. Check out http://web.mit.edu/gfischer/www/Downloads/Metrics383/383-7-Finite.pdf for a very detailed (handwritten) exposition.

- 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: Applied Econometrics**
**Department of Economics, University of Gothenburg**
**Appendix Lecture 2**

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. Earnings and education in Kenya

We now revisit the basic earnings model for Kenyan workers (see section 1 in the appendix to the lecture 1 notes):

$$lw_i = \beta_0 + \beta_1 \cdot ed_i + residual_i .$$

We suspect that education is endogenous because it is correlated with unobserved ability, the latter being part of the residual. To address this problem, we use data on the distance to primary school for the individual when s/he was seven years old. The idea is that distance to school is uncorrelated with unobserved ability, but correlated with education (since distance affects the cost of going to school), hence fulfilling the validity and relevance conditions  More precisely, the instrument is a dummy variable = 1 if the individual lived more than 6 km from the nearest primary school at the age of seven. This dummy is denoted *gt6km*.

Summary statistics are as follows:

```
. tabstat lw ed gt6km, s(mean N min max p50);

   stats |        lw        ed     gt6km
---------+------------------------------
    mean |  4.205216  9.933684  .0905263
       N |       950       950       950
     min |  2.227541         0         0
     max |  8.134265        17         1
     p50 |  4.019301        11         0
----------------------------------------
```

While we are suspicious of the OLS estimator in this context, this is usually a good place to start. One reason is that it gives us potentially interesting descriptive statistics (e.g. the correlation between lw and ed is given by the square root of the R-squared). Another reason is that it is often useful to have a benchmark estimator with which we can compare the IV estimates. So here are the OLS estimates, obtained by means of Stata:

```
. reg lw ed ;

      Source |       SS       df       MS              Number of obs =     950
-------------+------------------------------           F(  1,   948) =  181.38
       Model |  84.4673729     1  84.4673729           Prob > F      =  0.0000
    Residual |   441.47884   948  .465694979           R-squared     =  0.1606
-------------+------------------------------           Adj R-squared =  0.1597
       Total |  525.946213   949  .554210973           Root MSE      =  .68242

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .1042316   .0077394    13.47   0.000     .0890433    .1194198
       _cons |   3.169813   .0800051    39.62   0.000     3.012805     3.32682
------------------------------------------------------------------------------
```

Thus, if you believe education is exogenous, you would learn from this regression that the return to an additional year of schooling is about 10%, with a pretty tight confidence interval (between 9% and 12%).

Now consider the IV estimator. We see from the summary statistics that only 9% of the individuals (i.e. some 90 individuals in the sample) lived more than 6km from a primary school at the age of seven. We might worry that the instrument is not going to be particularly informative (relevant). ivreg2 reports a rich set of statistics that help us assess whether the key conditions required for identification are fulfilled. At this point, don't worry if you don't understand all the output produced by ivreg2.

```
. ivreg2 lw (ed=gt6km), first;

First-stage regressions
-----------------------

First-stage regression of ed:

OLS estimation
--------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      950
                                                      F(  1,   948) =     6.69
                                                      Prob > F      =   0.0098
Total (centered) SS     =  7774.822105                Centered R2   =   0.0070
Total (uncentered) SS   =       101519                Uncentered R2 =   0.9240
Residual SS             =  7720.309647                Root MSE      =    2.854

------------------------------------------------------------------------------
          ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       gt6km |  -.8348407   .3226777    -2.59   0.010    -1.468086   -.2015956
       _cons |   10.00926   .0970859   103.10   0.000     9.818731    10.19979
------------------------------------------------------------------------------
Included instruments: gt6km
------------------------------------------------------------------------------
Partial R-squared of excluded instruments:   0.0070
Test of excluded instruments:
  F(  1,   948) =     6.69
  Prob > F      =   0.0098



Summary results for first-stage regressions
-------------------------------------------

Variable    | Shea Partial R2 |   Partial R2    |  F(  1,   948)    P-value
ed          |     0.0070      |     0.0070      |        6.69       0.0098

Underidentification tests
Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)
Ha: matrix has rank=K1 (identified)
Anderson canon. corr. N*CCEV LM statistic   Chi-sq(1)=6.66     P-val=0.0099
Cragg-Donald N*CDEV Wald statistic          Chi-sq(1)=6.71     P-val=0.0096

Weak identification test
Ho: equation is weakly identified
Cragg-Donald Wald F-statistic                        6.69
See main output for Cragg-Donald weak id test critical values

Weak-instrument-robust inference
Tests of joint significance of endogenous regressors B1 in main equation
Ho: B1=0 and overidentifying restrictions are valid
Anderson-Rubin Wald test     F(1,948)= 2.00      P-val=0.1579
Anderson-Rubin Wald test     Chi-sq(1)=2.00      P-val=0.1572
Stock-Wright LM S statistic  Chi-sq(1)=2.00      P-val=0.1576

Number of observations              N  =        950
Number of regressors                K  =          2
Number of instruments               L  =          2
Number of excluded instruments      L1 =          1
```

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

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =       950
                                                      F(  1,   948) =      2.31
                                                      Prob > F      =    0.1285
Total (centered) SS     =  525.9462132                Centered R2   =    0.1391
Total (uncentered) SS   =  17325.59593                Uncentered R2 =    0.9739
Residual SS             =  452.8115283                Root MSE      =     .6904

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .1424103   .0935081     1.52   0.128    -.0408622    .3256827
       _cons |   2.790557   .9291496     3.00   0.003     .9694576    4.611657
------------------------------------------------------------------------------
Underidentification test (Anderson canon. corr. LM statistic):           6.661
                                                   Chi-sq(1) P-val =    0.0099
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                6.694
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           0.000
                                                 (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         ed
Excluded instruments: gt6km
------------------------------------------------------------------------------
```

Notice that the option *first*, added after a comma in the ivreg instruction, gives me the first-stage regression, in addition to the IV regression. Of course, the first-stage regression is very useful and important in this context, as it sheds light on whether we have an informative (relevant) instrument or not. Basically, if the instrument is insignificant in the first-stage, we effectively have not identified the model, and so we will not learn anything from the IV results.

Based on the first-stage results, it appears that we can at least be hopeful that we will be able to identify the parameter of interest in the second stage (i.e. the coefficient on education), since the coefficient on the distance dummy is negative (as expected) and significant at the 1% level. The point estimate implies that children living more than 6 km from the nearest primary school at the age of seven accumulate nearly one year (0.83) less education than those that do not. That seems to make sense.

Turning to the IV estimator, we see that the point estimate of the education coefficient is somewhat higher than the OLS estimate (0.14 vs. 0.10). This would seem to contradict the underlying prior, which is that education is positively correlated with unobserved ability and hence with the residual. In fact, this is a very common result in the empirical literature, see Card (2001) for a review.

We also see that the estimated standard error is very high indeed. Recall that the confidence interval for the OLS estimator is 9-11%. For the IV estimator, the 95% confidence interval is -4 to 33%, and so we cannot reject the hypothesis that the coefficient is zero. Thus, we appear to learn very little from this exercise. Why that may be so is an important issue that we will return to later.

Finally, note that in the (very) special case where there is one endogenous explanatory variable and one instrument, which is a dummy variable, one can show (indeed you will be asked to do so in the first problem set) that the IV estimator $\hat{\beta}_1$ can be written as

$$\hat{\beta}_1 = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0},$$

where $\bar{y}_0$ and $\bar{x}_0$ are the sample averages of $y_i$ and $x_i$ over the part of the sample with $z_i = 0$, and $\bar{y}_1$ and $\bar{x}_1$ are the sample averages of $y_i$ and $x_i$ over the part of the sample with $z_i = 1$. In our case we have:

```
. tabstat lw ed, s(mean) by(gt6km);

Summary statistics: mean
  by categories of: gt6km

   gt6km |        lw        ed
---------+--------------------
       0 |  4.215979  10.00926
       1 |  4.097089  9.174419
---------+--------------------
   Total |  4.205216  9.933684
------------------------------
```

Using the formula above thus yields (4.097089-4.215979)/(9.174419-10.00926) = 0.14241. This estimator is sometimes referred to as the Wald estimator. You might find this helps you with the intuition of the IV estimator.

## 2. Estimation by GMM

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

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

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                      Number of obs =     2220
                                                      F( 15,  2204) =    38.40
                                                      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

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1017497   .0130693     7.79   0.000     .0761343    .1273652
       exper |   .1004833   .0096578    10.40   0.000     .0815544    .1194123
     expersq |   -.002493   .0004108    -6.07   0.000    -.0032983   -.0016878
       black |  -.1549702    .026376    -5.88   0.000    -.2066661   -.1032742
       south |  -.1226742   .0343093    -3.58   0.000    -.1899193   -.0554292
        smsa |   .1244044   .0233825     5.32   0.000     .0785754    .1702333
      reg661 |   -.080592   .0456197    -1.77   0.077    -.1700049    .0088209
      reg662 |   .0056286   .0335724     0.17   0.867     -.060172    .0714293
      reg663 |   .0411136   .0324834     1.27   0.206    -.0225527    .1047799
      reg664 |  -.0486601   .0413421    -1.18   0.239    -.1296891    .0323688
      reg665 |    .013062   .0454376     0.29   0.774    -.0759939     .102118
      reg666 |   .0314252   .0497923     0.63   0.528    -.0661658    .1290162
      reg667 |   .0172291   .0482101     0.36   0.721    -.0772609    .1117191
      reg668 |  -.1598693   .0552705    -2.89   0.004    -.2681974   -.0515412
      smsa66 |   .0276992    .021362     1.30   0.195    -.0141696     .069568
       _cons |    4.23282    .227809    18.58   0.000     3.786322    4.679317
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):           169.520
                                                   Chi-sq(4) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Kleibergen-Paap rk Wald F statistic):         56.318
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.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         6.236
                                                   Chi-sq(3) P-val =    0.1007
-endog- option:
Endogeneity test of endogenous regressors:                               3.720
                                                   Chi-sq(1) P-val =    0.0538
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
------------------------------------------------------------------------------
```

. ivreg2 lwage  (educ= nearc2 nearc4 motheduc fatheduc) exper expersq black south smsa reg661 reg662 reg663 re
> g664 reg665
> reg666 reg667 reg668 smsa66, gmm2s robust endog(educ);

**2-Step GMM estimation**

---

```
Estimates efficient for arbitrary heteroskedasticity
Statistics robust to heteroskedasticity

                                                      Number of obs =     2220
                                                      F( 15,  2204) =    38.74
                                                      Prob > F      =   0.0000
Total (centered) SS     =  428.9994844                Centered R2   =   0.2614
Total (uncentered) SS   =  88133.52155                Uncentered R2 =   0.9964
Residual SS             =   316.878457                Root MSE      =    .3778

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1003219   .0130403     7.69   0.000     .0747635    .1258804
       exper |   .0997653    .009651    10.34   0.000     .0808498    .1186808
     expersq |  -.0024939   .0004108    -6.07   0.000    -.0032991   -.0016887
       black |  -.1586361   .0262778    -6.04   0.000    -.2101396   -.1071325
       south |   -.119613   .0342555    -3.49   0.000    -.1867525   -.0524735
        smsa |   .1268335   .0233467     5.43   0.000     .0810749    .1725921
      reg661 |  -.0860177   .0454949    -1.89   0.059    -.1751862    .0031507
      reg662 |   .0042036   .0335305     0.13   0.900     -.061515    .0699222
      reg663 |   .0393087   .0324169     1.21   0.225    -.0242272    .1028446
      reg664 |  -.0506996   .0412997    -1.23   0.220    -.1316455    .0302463
      reg665 |   .0080002   .0453793     0.18   0.860    -.0809415     .096942
      reg666 |   .0198662   .0494659     0.40   0.688    -.0770852    .1168175
      reg667 |   .0111105   .0481193     0.23   0.817    -.0832017    .1054226
      reg668 |  -.1654134   .0551271    -3.00   0.003    -.2734605   -.0573664
      smsa66 |   .0256423   .0213386     1.20   0.229    -.0161805    .0674651
       _cons |   4.260154    .227212    18.75   0.000     3.814826    4.705481
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):           169.520
                                                     Chi-sq(4) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Kleibergen-Paap rk Wald F statistic):         56.318
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.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
```

```
Hansen J statistic (overidentification test of all instruments):         6.236
                                                      Chi-sq(3) P-val =    0.1007
-endog- option:
Endogeneity test of endogenous regressors:                               3.720
                                                      Chi-sq(1) P-val =    0.0538
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
------------------------------------------------------------------------------
```

## 3. Illustration using the CARD.RAW data

Note: The CARD.RAW data are used in various problems in Wooldridge, e.g. problem 5.4 and 6.1.

```
. use "C:\teaching_gbg07\applied_econ07\lectures\wooldat\CARD.dta", clear
```

Table 3.1 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
------------------------------------------------------------------------------
```

Table 3.2: 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 3.3 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 3.4: 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 3.5: 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 4: 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 5: 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
```