

The Two Sample Problem Revisited *

PHOEBUS J. DHRYMES

Columbia University

May 2004

Abstract

This paper re-examines the two sample instrumental variable estimator suggested by Angrist and Krueger (1992), AK. It shows that this is not a different estimator but is merely an Aitken-like procedure in which various moments have been obtained from different sources. The procedure is made fully implementable by providing consistent estimators for the requisite covariance matrices. Finally, various two-sample based estimators are compared and ranked and it is established that the two-sample based estimator examined herein is inefficient relative to the feasible Aitken-like estimator when all requisite information is available at the same time from the same source.

*©Phoebus J. Dhrymes, 2004

Preliminary material; not to be quoted or disseminated without permission of the author.

1 Introduction

A number of recent empirical papers make use of a procedure given in Angrist and Krueger (1992 (AK)); this is an econometric procedure to rectify certain anomalies, when all requisite data are not available from the same source. The procedure described in AK is somewhat more complex than it needs to be; moreover the estimation of certain covariance matrices essential to the method's implementation is not spelled out.

2 Problem Formulation

We shall formulate the problem in a manner different from that given in AK. First we shall obtain the efficient estimator assuming that all required information is available; second, we shall examine what modifications, if any, are required in view of the fact that all requisite data are not available at the same time, from the same source; third, we shall show that the estimator obtained by AK is simply one that involves the utilization, therein, of sample moments obtained from two different samples; fourth, we shall examine several possible alternative estimators and rank them, and fifth, we shall show that the most efficient two-sample based estimator is **inefficient** relative to the one obtainable when all requisite information is available at the same time from the same source.

2.1 The Model

Let

$$y_i = x_i \beta + \epsilon_i$$

$$x_{i.} = z_{i.}A + u_{i.}, \quad i = 1, 2, 3, \dots, n \quad (1)$$

where y is the scalar variable of interest, $x_{i.}$ is a k -element row vector of explanatory variables, some of which may be correlated with the structural error, ϵ , $z_{i.}$ is the m -element (row) vector of **instrumental** variables, and A and β are parameters.

The usual assumptions for such models are invoked, viz.,

- i. The sequence ϵ_i is one of independent heteroskedastic random variables with mean zero and variance $0 < \sigma_{ii} < \infty$; the vectors $u_{i.}$ are a sequence of i.i.d. random vectors with mean zero and a certain covariance matrix Φ , which need not be positive definite because some of the x 's may be contained in z .
- ii. The matrices $X = (x_{i.})$, $Z = (z_{i.})$ are of full column rank and

$$\frac{X'X}{n} \longrightarrow M_{xx} > 0, \quad \frac{Z'Z}{n} \rightarrow M_{zz} > 0.$$

- iii. ϵ_i and $u_{i.}$ are **correlated** and their covariance (vector) is given by $\rho_{i.}^* = (\rho_{i1}, \rho_{i2}, \dots, \rho_{ik})$.

Using the method described in Dhrymes (1969), we take the system in Eq. (1) and transform it (the first equation), using the “reduced form” of the second equation therein so that the transformed first equation obeys all the requirements for the standard GLM (general linear model);¹ if we then apply OLS to the transformed equation we shall obtain, by the Gauss-Markov theorem, the efficient estimator. Thus,

$$Z'y = Z'X\beta + Z'\epsilon \quad (2)$$

¹This approach has also been used, *mutatis mutandis* by Amemiya (), Jorgenson and Laffont (), and Hansen () to produce the nonlinear 2SLS, 3SLS and GMM estimators.

gives an equation in which the explanatory variables $Z'X$ are (asymptotically) uncorrelated with the error term $Z'\epsilon$. The problem is that the error term of the transformed system has a non-scalar covariance matrix

$$\text{Cov}(Z'\epsilon) = Z'\Sigma Z, \quad \Sigma = \text{diag}(\sigma_{11}, \sigma_{22}, \sigma_{33}, \dots, \sigma_{nn}). \quad (3)$$

Since by assumption the matrix above is **positive definite** there exists a nonsingular matrix R such that

$$RR' = Z'\Sigma Z. \quad (4)$$

Consider then

$$R^{-1}Z'y = R^{-1}Z'X\beta + R^{-1}Z'\epsilon. \quad (5)$$

In this transformed system we have, using Eq. (4)

$$\text{Cov}(R^{-1}Z'\epsilon) = I_m, \quad \frac{X'Z(Z'\Sigma Z)^{-1}Z'\epsilon}{n} \xrightarrow{P} 0. \quad (6)$$

Consequently,

$$\begin{aligned} \hat{\beta} &= [X'Z(Z'\Sigma Z)^{-1}Z'X]^{-1}X'Z(Z'\Sigma Z)^{-1}Z'y \\ &= \beta + [X'Z(Z'\Sigma Z)^{-1}Z'X]^{-1}X'Z(Z'\Sigma Z)^{-1}Z'\epsilon, \end{aligned} \quad (7)$$

is the **efficient** estimator, which may be easily shown to obey

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Psi^{-1}), \quad \Psi = \text{plim}_{n \rightarrow \infty} \frac{X'Z(Z'\Sigma Z)^{-1}Z'X}{n}. \quad (8)$$

Whether this estimator is **feasible** depends on our ability to estimate the matrix $Z'\Sigma Z/n$. If we can, the estimator above becomes feasible, and its limiting distribution will be as indicated.

3 Two Sample Complication

Suppose now that we do not have information on y , Z , X from a single sample; instead, we have information on y and Z from sample one, of size n_1 , say y_1 , Z_1 , and information on X , Z from a second sample of size n_2 , say X_2 , Z_2 .

We now add assumption iv: The two samples are mutually independent.

A simple solution to the two sample problem would be to substitute the appropriate sample moments as they are available, leading to

$$\hat{\beta} = [X_2'Z_2(Z_1'\Sigma Z_1)^{-1}Z_2'X_2]^{-1}X_2'Z_2(Z_1'\Sigma Z_1)^{-1}Z_1'y_1. \quad (9)$$

It may be shown that **if** $\lim_{n_1, n_2 \rightarrow \infty} (n_1/n_2) = 1$, the estimator above is **consistent**. If, however, the limit equals something different from one it is not, and it converges to a multiple of β . For this reason, and also considerations regarding its limiting distribution, we adjust for sample size at this stage, leading to the estimator,

$$\tilde{\beta} = [X_2'Z_2(Z_1'\Sigma Z_1)^{-1}Z_2'X_2]^{-1} \frac{n_2}{n_1} X_2'Z_2(Z_1'\Sigma Z_1)^{-1}Z_1'y_1 \quad (10)$$

$$= \left[\frac{X_2'Z_2}{n_2} \left(\frac{Z_1'\Sigma Z_1}{n_1} \right)^{-1} \frac{Z_2'X_2}{n_2} \right]^{-1} \left[\frac{X_2'Z_2}{n_2} \left(\frac{Z_1'\Sigma Z_1}{n_1} \right)^{-1} \frac{Z_1'y_1}{n_1} \right], \text{ and}$$

$$\tilde{\beta} - \beta = \left[\frac{X_2'Z_2}{n_2} \left(\frac{Z_1'\Sigma Z_1}{n_1} \right)^{-1} \frac{Z_2'X_2}{n_2} \right]^{-1} C(n_1, n_2), \quad \text{where} \quad (11)$$

$$C(n_1, n_2) = \frac{X_2'Z_2}{n_2} \left(\frac{Z_1'\Sigma Z_1}{n_1} \right)^{-1} \frac{Z_1'X_1}{n_1} \beta - \frac{X_2'Z_2}{n_2} \left(\frac{Z_1'\Sigma Z_1}{n_1} \right)^{-1} \frac{Z_2'X_2}{n_2} \beta + \left(\frac{X_2'Z_2}{n_2} \left(\frac{Z_1'\Sigma Z_1}{n_1} \right)^{-1} \frac{Z_1'\epsilon_1}{n_1} \right). \quad (12)$$

It follows therefore that

$$\sqrt{n_1}(\tilde{\beta} - \beta) \sim (M_{xz}M_{zz}^{*-1}M_{zx})^{-1}M_{xz}M_{zz}^{*-1}\sqrt{n_1}\zeta_n, \quad (13)$$

where

$$\zeta_n = \left(\frac{Z_1'X_1\beta}{n_1} - M_{zx}\beta + \frac{Z_1'\epsilon_1}{n_1} \right) - \left(\frac{Z_2'X_2\beta}{n_2} - M_{zx}\beta \right)$$

$$M_{zz}^{-1} = \text{plim}_{n_1 \rightarrow \infty} \frac{1}{n_1} Z_1' \Sigma Z_1, \quad M_{zx} = \text{plim}_{n_1 \rightarrow \infty} Z_1' X_1 = \text{plim}_{n_2 \rightarrow \infty} \frac{1}{n_2} Z_2' X_2. \quad (14)$$

Because the two samples are independent, the first and second expressions in rounded brackets are mutually independent; however, the two terms in the first rounded brackets are dependent because of the incidence of simultaneity or other reasons for the correlation of ϵ_i and u_i , or more precisely v_i , to be defined below.

Remark 1. The complication with the two sample case arises from the first two terms of Eq. (12). While the difference

$$\Delta = \frac{X_2'Z_2}{n_2} \left(\frac{Z_1'\Sigma Z_1}{n_1} \right)^{-1} \frac{Z_1'X_1}{n_1} \beta - \frac{X_2'Z_2}{n_2} \left(\frac{Z_1'\Sigma Z_1}{n_1} \right)^{-1} \frac{Z_2'X_2}{n_2} \beta$$

obeys $\Delta \xrightarrow{P} 0$, $\sqrt{n_1}\Delta$ **does not converge to zero in probability**. Instead, it converges in distribution to a well defined normal vector, more specifically

$$\sqrt{n_1}\Delta \sim M_{xz}M_{zz}^{*-1} \left(\frac{1}{\sqrt{n_1}} Z_1' X_1 \beta - \sqrt{\frac{n_1}{n_2}} \frac{1}{\sqrt{n_2}} Z_2' X_2 \right),$$

which converges in distribution to a well defined random vector, to be derived with more detail below.

This is the proper juncture to examine more closely the equations of the model. In the first equation of Eq. (1), $y_i = x_i\beta + \epsilon_i$, suppose that k_1

of the explanatory variables are exogenous, i.e. they are independent of ϵ and as such they have been included in the matrix of instruments Z ; the remaining k_2 variables ($k_1 + k_2 = k$) are then either actually or potentially correlated with the structural error ϵ . Partition $X = (X_{k_1}, X_{k_2})$ and notice that $Z = (X_{k_1}, Z_{m-k_1})$, and note that in the second equation of Eq. (1)

$$X = ZA + U \text{ implies the restrictions } A = \begin{bmatrix} I_{k_1} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \text{ and } U = (0, U_{k_2}).$$

For simplicity of notation denote $U_{k_2} = V = (v_i)$, and assume

- i. the v_i are independent identically distributed with mean zero and covariance matrix Γ ;
- ii. (ϵ_i, v_i) are independent not identically distributed with mean zero and covariance matrix

$$\text{Cov}(\epsilon_i, v_i) = \begin{bmatrix} \sigma_{ii} & \rho_i \\ \rho_i' & \Gamma \end{bmatrix},$$

where ρ_i is the appropriate sub-vector of ρ_i^* , as defined in connection with assumption iii.

Note that ρ_i has dimension $1 \times k_2$, and Γ is $k_2 \times k_2$.

It will suffice now to find the limiting distribution of

$$\sqrt{n_1}\zeta_n = \left(\frac{Z_1' X_1 \beta}{\sqrt{n_1}} - \sqrt{n_1} M_{zx} \beta + \frac{Z_1' \epsilon_1}{\sqrt{n_1}} \right) - \frac{\sqrt{n_1}}{\sqrt{n_2}} \left(\frac{Z_2' X_2 \beta}{\sqrt{n_2}} - \sqrt{n_2} M_{zx} \beta \right), \quad (15)$$

which is accomplished by finding (separately in view of independence) those of the entities in round brackets. The first bracket may be written as

$$\left(\frac{Z_1' X_1 \beta}{\sqrt{n_1}} - \sqrt{n_1} M_{zx} \beta + \frac{Z_1' \epsilon_1}{\sqrt{n_1}} \right) = \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} ([z_{i1}' x_i \beta - M_{zx} \beta] + (z_{i1}' \epsilon_i)). \quad (16)$$

The first component of the right member (in square brackets) has mean zero and covariance matrix

$$\Phi_{22i} = \frac{1}{n_1} (\beta'_{k_2} \Gamma \beta_{k_2}) z'_{i1} \cdot z_{i1}, \quad (17)$$

which, upon summing and taking limits, yields the covariance matrix of that term as

$$\Phi_{22} = \lim_{n_1 \rightarrow \infty} (\beta'_{k_2} \Gamma \beta_{k_2}) \frac{Z'_1 Z_1}{n_1} = (\beta'_{k_2} \Gamma \beta_{k_2}) M_{zz}. \quad (18)$$

The covariance matrix of the term $Z'_1 \epsilon / \sqrt{n_1}$ is easily established to be

$$\Phi_{11} = \lim_{n_1 \rightarrow \infty} \frac{Z'_1 E[\epsilon \epsilon'] Z_1}{n_1} = \lim_{n_1 \rightarrow \infty} \frac{Z'_1 \Sigma Z_1}{n_1} = M_{zz}; \quad (19)$$

finally, the cross-covariance is given by

$$\Phi_{12} = 2 \lim_{n_1 \rightarrow \infty} \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} z'_{i1} \cdot E[\epsilon_i v_j] \beta_{k_2} z_{j1} = \lim_{n_1 \rightarrow \infty} \frac{Z'_1 \Sigma_{12} Z_1}{n_1}, \quad (20)$$

$$\Sigma_{12} = \text{diag}(\rho_1 \cdot \beta_{k_2}, \rho_2 \cdot \beta_{k_2}, \dots, \rho_{n_1} \cdot \beta_{k_2}).$$

By the same argument, we may show that the second term in Eq.(15) converges in distribution to a random vector with distribution

$$N(0, \alpha \Phi_{22}), \quad \alpha = \lim_{n_1, n_2 \rightarrow \infty} \frac{n_1}{n_2}.$$

Consequently we have established that

$$\begin{aligned} \sqrt{n_1}(\tilde{\beta} - \beta) &\xrightarrow{d} N(0, \Psi_1), \quad \Psi_1 = \Psi^{-1} M_{xz} M_{zz}^{*-1} \Omega_1 M_{zz}^{*-1} M_{zx} \Psi^{-1}, \\ \Omega_1 &= \Phi_{11} + 2\Phi_{12} + (1 + \alpha)\Phi_{22}. \end{aligned} \quad (21)$$

4 Implementation for Inference

To make the estimator(s) above fully implementable for purposes of inference, and more definitively obtain the most efficient feasible estimator, we need to produce consistent estimators of the variance covariance matrices discussed above. Thus, for example if we cannot find a consistent estimator of $Z_1' \Sigma Z_1 / n_1$, the estimator whose distribution we obtained in the previous section is **not implementable**. We begin by considering sample 1 and noting that we may obtain the residuals

$$e = y_1 - \hat{X}_1 \tilde{\beta}, \quad \text{wehre } \hat{X}_1 = Z_1 \tilde{A}, \quad \tilde{A} = (Z_2' Z_2)^{-1} Z_2' X_2, \quad (22)$$

and $\tilde{\beta}$ is an **initial consistent** estimator. Developing the right member above we find

$$e = \epsilon_1 + (X_1 - \hat{X}_1)\beta - \hat{X}_1(\tilde{\beta} - \beta), \quad X_1 - \hat{X}_1 = U_1 - Z_1(\tilde{A} - A). \quad (23)$$

The i th component of this vector is given by

$$e_i = \epsilon_{i1} + v_{i1} \cdot \beta_{k_2} - z_{i1} \cdot (\tilde{A} - A)\beta - z_{i1} \cdot \tilde{A}(\tilde{\beta} - \beta). \quad (24)$$

In view of the consistency of \tilde{A} , $\tilde{\beta}$, for sufficiently large n_1 , n_2 ,

$$e_i \sim \epsilon_{i1} + v_{i1} \cdot \beta_{k_2}, \quad (25)$$

so that

$$e_{i1}^2 \sim \epsilon_{i1}^2 + \beta_{k_2}' v_{i1}' \cdot v_{i1} \cdot \beta_{k_2} + 2\epsilon_{i1} v_{i1} \cdot \beta_{k_2}, \quad (26)$$

and

$$\begin{aligned} \frac{1}{n_1} \sum_{i=1}^{n_1} z_{i1}' \cdot z_{i1} \cdot e_{i1}^2 &\xrightarrow{P} \lim_{n_1 \rightarrow \infty} \frac{1}{n_1} \sum_{i=1}^{n_1} z_{i1}' \cdot z_{i1} \cdot (\sigma_{ii} + \beta_{k_2}' \Gamma \beta_{k_2} + 2\rho_i \cdot \beta_{k_2}) \\ &= \Phi_{11} + \Phi_{22} + 2\Phi_{12}, \quad \text{where } \Phi_{22} = \beta_{k_2}' \Gamma \beta_{k_2} M_{zz}. \end{aligned} \quad (27)$$

Unfortunately, we **cannot** separately estimate the three components of the matrix above, even though we could estimate consistently Φ_{22} , from sample 2, for example

$$\tilde{\Phi}_{22} = \frac{1}{n_2^2} \tilde{\beta}'_{k_2} \tilde{V}'_2 \tilde{V}_2 \tilde{\beta}_{k_2} Z'_2 Z_2 \quad (28)$$

5 Issues of Efficiency

We recall from Eq. (13) that the estimator examined earlier (asymptotically) has two components; the first component $(M_{xz} M_{zz}^{*-1} M_{zx})^{-1} M_{xz} M_{zz}^{*-1}$ involves a “weighting matrix, M_{zz}^{*-1} and other entities provided by the data of the problem; it also involves $\sqrt{n_1} \zeta_n$, which **does not contain** the “weighting matrix. Thus, irrespective of what “weighting matrix we use its behavior, asymptotically, does not change. To find an efficient estimator in this context, we require

Lemma 1. Let H_1 , H_2 , D be conformable matrices of full column rank; H_2 , D are (square) positive definite matrices and H_1 is otherwise completely arbitrary. Put

$$A_1 = (H'_1 H_2^{-1} H_1)^{-1} H'_1 H_2^{-1} D H_2^{-1} H_1 (H'_1 H_2^{-1} H_1)^{-1}, \quad A_2 = (H'_1 D^{-1} H_1)^{-1},$$

then

$$A_1 - A_2 \geq 0.$$

Proof: Without loss of generality, define C by

$$(H'_1 H_2^{-1} H_1)^{-1} H'_1 H_2^{-1} = (H'_1 D^{-1} H_1)^{-1} H'_1 D^{-1} + C, \quad (29)$$

and note that $CH_1 = 0$. Post-multiply the equation above by D to obtain

$$(H'_1 H_2^{-1} H_1)^{-1} H'_1 H_2^{-1} D = (H'_1 D^{-1} H_1)^{-1} H'_1 + CD, \quad (30)$$

and multiply the corresponding sides of Eq.(29) by (the transpose of) of Eq. (30) obtain

$$A_1 A_2 = CDC' \quad (31)$$

q.e.d.

Remark 2. The matrix CDC' is at least positive semi-definite, and it is the zero matrix only if $C = 0$, which will be the case if $H_2 = D$.

The relevance of Lemma 1 to our problem is that the estimator given in Eq. (13) is **inefficient unless we take** $M_{zz}^* = \Omega_1$. We have therefore proved

Proposition 1. The estimator

$$\hat{\beta} = [X_2' Z_2 \hat{\Omega}_1^{-1} Z_2' X_2]^{-1} \frac{n_2}{n_1} X_2' Z_2 \hat{\Omega}_1^{-1} Z_1' y_1 \quad (32)$$

and its limiting distribution is given by

$$\sqrt{n_1}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Psi_1^{-1}), \quad \Psi_1 = (M_{xz} \Omega_1^{-1} M_{zx}). \quad (33)$$

An immediate consequence of Proposition 1 is

Proposition 2. The efficient estimator in the two-sample case is **inefficient** relative to the estimator obtained earlier when all requisite data are available at the same time from the same source, provided

$$\sigma_{ii}^{1/2} \leq \frac{1 + \alpha}{2} (\beta'_{k_2} \Gamma \beta_{k_2})^{1/2}.$$

Proof: The estimator in the latter case (also termed $\hat{\beta}$) has a limiting distribution, given in Eq. (8), which is $N(0, \Psi^{-1})$, where $\Psi = M_{xz} \Phi_{11}^{-1} M_{zx}$.

But

$$\Psi_1^{-1} - \Psi^{-1} \geq 0, \quad \text{if and only if } \Psi - \Psi_1 \geq 0.$$

Now,

$$\Psi - \Psi_1 = M_{xz}[\Phi_{11}^{-1} - \Omega_1]M_{zx} \geq 0, \text{ if and only if } \Omega_1 - \Phi_{11} = 2\Phi_{12} + (1 + \alpha)\Phi_{22} \geq 0.$$

The last inequality, however, is not guaranteed unless

$$|\rho_{i \cdot} \beta_{k_2}| \leq \frac{1 + \alpha}{2} \beta'_{k_2} \Gamma \beta_{k_2}.$$

Since

$$(\rho_{i \cdot} \beta_{k_2})^2 \leq \sigma_{ii} \beta'_{k_2} \Gamma \beta_{k_2}, \quad (34)$$

it follows that

$$|\rho_{i \cdot} \beta_{k_2}| \leq \sigma_{ii}^{1/2} (\beta'_{k_2} \Gamma \beta_{k_2})^{1/2} \leq \frac{1 + \alpha}{2} \beta'_{k_2} \Gamma \beta_{k_2}, \text{ because } \sigma_{ii}^{1/2} \leq \frac{1 + \alpha}{2} (\beta'_{k_2} \Gamma \beta_{k_2})^{1/2}. \quad (35)$$

q.e.d.

Remark 3. A concise description of the procedure is as follows:

- i. Obtain an initial consistent estimator, say

$$\tilde{\beta} = [X'_2 Z_2 (Z'_1 Z_1)^{-1} Z'_1 X_2]^{-1} \frac{n_2}{n_1} X'_2 Z_2 (Z'_1 Z_1)^{-1} Z'_1 y_1. \quad (36)$$

- ii. Obtain the residuals

$$e = y_1 \hat{X}_1 \tilde{\beta}, \quad \hat{X}_1 = Z_1 \tilde{A}, \quad \tilde{A} = (Z'_2 Z_2)^{-1} Z'_2 X_2. \quad (37)$$

- iii. Use the residuals above to estimate part of Ω_1 as

$$\hat{\Omega}_1 - \alpha \Phi_{22} = \frac{1}{n_1} \sum_{i=1}^{n_1} z'_{i1} z_{i1} e_i^2. \quad (38)$$

- iv. Use the residuals, \tilde{V}_2 from the estimation of A to obtain

$$\hat{\Phi}_{22} = \frac{1}{n_2} \tilde{\beta}'_{k_2} \tilde{V}'_2 \tilde{V}_2 \tilde{\beta}_{k_2} Z'_2 Z_2. \quad (39)$$

Add $\alpha \hat{\Phi}_{22}$ to the result in item iii. to obtain the complete estimate of $\hat{\Omega}_1$.

v. Use the preceding to obtain the efficient estimator

$$\hat{\beta} = [X_2' Z_2 \hat{\Omega}_1^{-1} Z_2' X_2]^{-1} \frac{n_2}{n_1} X_2' Z_2 \hat{\Omega}_1^{-1} Z_1' y_1. \quad (40)$$