Columbia University

Department of Economics
Discussion Paper Series

Estimation of Models with Grouped and Ungrouped Data by Means of "2SLS"

Phoebus J. Dhrymes
Adriana Lleras-Muney

Discussion Paper #:0102-66

Department of Economics
Columbia University
New York, NY 10027

July 2002

Estimation of Models with Grouped and Ungrouped Data by Means of "2SLS"

Phoebus J. Dhrymes
Adriana Lleras-Muney
Columbia University

September 2001; revised November 2001

Abstract

This paper deals with a special case of estimation with grouped data, where the dependent variable is only available for groups, whereas the endogenous regressor(s) is available at the individual level. In this situation, the solution adopted by researchers is to aggregate the individual data and then use standard 2SLS estimation. However when some data is available at the individual level, it might be possible to gain efficiency by estimating the first stage using the available individual data, and then estimating the second stage at the aggregate level. This estimation procedure yields a consistent and asymptotically normal estimator that we refer to as Mixed-2SLS. Depending on the parametric configuration of the model, the Mixed-2SLS estimator can be more or less efficient than standard 2SLS. The standard 2SLS estimator of this literature is asymptotically equivalent to the OLS estimator based on group data alone.

*We wish to acknowledge useful comments received from the participants of a seminar at Princeton University

A number of simulations are carried out that illustrate or confirm theoretical findings.

1 Introduction and Review

Individual data is not always available for empirical estimation, but often grouped data can be obtained. As is well known, grouped data estimation of well-specified linear models yields unbiased and consistent estimates of the parameters, see e.g. Prais and Aitchison (1954). However, it is often the case that the specified model contains one or more explanatory variables (regressors) which are correlated with the structural error term. This situation arises either because the model is truly a system of simultaneous equations or because there is an omitted variable that is correlated with a regressor. A standard solution to this problem uses instrumental variables to obtain consistent estimates of the parameters. Instrumental variable estimation can easily be done using grouped data by the standard "2SLS" procedure common in this literature or, equivalently, estimators of the parameters of the individual model can be obtained by GLS estimation of data that has been grouped using the relevant instruments, see Angrist (1991a), and Moffit (1995). This paper was motivated by, and deals with, a particular case of instrumental variables for grouped data, where the dependent variable is only available for groups, whereas the endogenous regressor(s) is available at the individual level. This will be the case for example when matching data from multiple sources, or if the data being used is restricted so that only aggregates are available to the researcher. In general, the situation described in this paper applies to any estimation done at the aggregate level, where the first stage can potentially be estimated using disaggregate data. Recent papers where the data available is of this type include the works of Angrist (1991b), Pritchett and Summers (1996), Winter Ebmer and Steven (1999), Dee and Evans (1999) and Lleras-Muney (2001). In this situation, the solution often adopted by researchers has been to aggregate the individual data and then use standard "2SLS" estimation. However when some data is available at the individual level, it may be possible to gain efficiency by estimating the first stage using the available individual data, and then estimating the second stage at the aggregate level. This estimation procedure yields a consistent and asymptotically normal estimator that we refer to as Mixed-2SLS and, depending on the parametric configuration of the model, the Mixed-2SLS estimator can be more or less efficient than standard 2SLS. The estimator discussed here can be compared to the Two-Sample Instrumental Variable estimator developed by Angrist and Krueger (1994). Other aspects of the previous literature on aggregation of linear models explore the efficiency issues that arise when using grouped data. For example, Feser and Ronchetti (1997) and Im (1998) derive efficient estimators for grouped data. The consequences of heteroskedasticity were explored by Blackburn (1997) and Dickens (1990). Moulton (1990) discussed the problem of intra-group correlations. Shore-Sheppard (1996) and Hoxby and Paserman (1998) looked at the implication of within-group correlation when using instrumental variables. No other paper however, has examined the special case when grouped and ungrouped data is available. This paper is organized as follows. Section one contains an introduction to the problem and a review of the literature; section two provides the formulation for the general problem and derives the Mixed-2SLS estimator; section three shows that the estimator is consistent and asymptotically normal; section four compares this estimator with the standard "2SLS" estimator that is most commonly used in empirical work; section five discusses a variety of issues that may arise in the empirical implementation of the estimator. Each section includes results from a number of simulations reported in Appendices 1 and 2, that are designed to illustrate or confirm, for finite samples, findings obtained by asymptotic theory. Section six concludes.

2 Formulation of the Problem

Consider the model

$$y = X\beta + u, \qquad X = (X_1, x_{\cdot k}), \tag{1}$$

where y is the $n \times 1$ vector of observations on the dependent variable, X is the $n \times k$ matrix of observations on the k explanatory variables, β is a conformable vector of unknown parameters and u is the (structural) error vector whose components are asserted to be i.i.d. with mean 0 and variance $0 < \sigma_{11} < \infty$. It is asserted that one of the variables, say x_k, is correlated with the error, while the variables in X_1 are independent of the structural error vector u. This situation arises either because the model is truly a system of simultaneous equations where y and x_k are jointly determined, or because there is an omitted variable that is correlated with x_k. It is further asserted that we may represent the observations on the correlated explanatory variable as

$$x_{\cdot k} = Z\gamma + v, \tag{2}$$

where Z is the $n \times m$ matrix of observations on the "instruments", which are asserted to be independent of v and u, as is also X_1. By the process, often inappropriately termed 2SLS, of regressing $x_{\cdot k}$ on Z, and then y on $\hat{X} = (X_1, \hat{x}_{\cdot k})$, we may obtain consistent estimators of β,

$$\hat{\beta}_{i2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = \beta + (\hat{X}'\hat{X})^{-1}\hat{X}'(u + \beta_k Pzv) \tag{3}$$

where

$$\hat{x}_{\cdot k} = Z(Z'Z)^{-1}Z'x_{\cdot k}, \quad \hat{v} = x_{\cdot k} - \hat{x}_{\cdot k} = [I - Z(Z'Z)^{-1}Z']v = P_z v. \tag{4}$$

Because individual level data is not available for y, it is desired to estimate the parameter β not as indicated above, but by means of grouped data, after the variables of Eqs. (3) and (4) have been obtained. The grouping is as follows: The n observations are divided into G groups such that the i th group contains n_i observations and

$$\sum_{i=1}^{G} n_i = n. \tag{5}$$

Without loss of generality, we may rearrange the observations so that the first n_1 observations belong to group 1, the next n_2 observations belong to group 2 and so on. Grouping is effected by means of the (grouping) $G \times n$ matrix $H = (h_{i\cdot})$, where $h_{i\cdot}$ contains all zero elements, except for an n_i-element row vector, i.e.

$$h_{i\cdot} = (0, \frac{1}{\sqrt{n_i}} e_{i\cdot}, 0), \quad e_{i\cdot} = (1, 1, 1, \ldots, 1) \tag{6}$$

where its nonzero components appear in the positions $n_1 + n_2 + \ldots + n_{i-1} + 1, \ldots, n_1 + n_2 + \ldots + n_{i-1} + n_i$. The grouped data are thus given by

$$Hy = (\sqrt{n_1}\bar{y}_1, \sqrt{n_2}\bar{y}_2, \ldots, \sqrt{n_G}\bar{y}_G)', \;\; Hu = (\sqrt{n_1}\bar{u}_1, \sqrt{n_2}\bar{u}_2, \ldots, \sqrt{n_G}\bar{u}_G)',$$

$$H\hat{X} = (\sqrt{n_1}\bar{x}_1', \sqrt{n_2}\bar{x}_2', \ldots, \sqrt{n_G}\bar{x}_G')', \;\; HZ = (\sqrt{n_1}\bar{z}_1', \sqrt{n_2}\bar{z}_2', \ldots, \sqrt{n_G}\bar{z}_G')',$$

$$Hv = (\sqrt{n_1}\bar{v}_1, \sqrt{n_2}\bar{v}_2, \ldots, \sqrt{n_G}\bar{v}_G)',$$

where $\bar{u}_i, \; \bar{v}_i$, denote the (scalar) means of the corresponding variables in the ith group, and $\bar{x}_i, \; \bar{z}_i$, are k-and m-element row vectors, respectively, containing the ith group means of the x-and z-variables, respectively.

Using these definitions, the equation of interest can be written as:

$$Hy = HX\beta + Hu = H\hat{X}\beta + (Hu + \beta_k HP_z v).$$

The Mixed-2SLS estimator whose properties we shall now establish is given by

$$\hat{\beta} = (\hat{X}'H'H\hat{X})^{-1}\hat{X}'H'Hy = \beta + (\hat{X}'H'H\hat{X})^{-1}\hat{X}'H'H[u + \beta_k P_z v],$$

$$\hat{X} = (X_1, Z\hat{\gamma}), \quad \hat{\gamma} = (Z'Z)^{-1}Z'x_{\cdot k}, \quad P_z = I - Z(Z'Z)^{-1}Z'. \tag{7}$$

Intuitively, this estimator is derived by obtaining predicted values of X using individual data; upon grouping these predicted values we then estimate the equation of interest by OLS, where HX has been replaced by $H\hat{X}$.

There are three major issues to be discussed relative to this problem. First, what are the properties of the resulting estimator if we follow the procedure outlined above. Second, what are the properties of the resulting estimator **if** x_k is regressed on Z **using grouped data**, i.e. if one follows the standard "2SLS" estimation procedure. Third, if only grouped data are used is the "2SLS" significantly different from the OLS estimator using grouped data.

3 Properties of Estimators

3.1 Consistency

We shall establish the properties of the Mixed estimator in Eq. (7) under the following assumptions:

i. The matrices X and Z obey

$$\frac{1}{n}X'X \rightarrow M_{xx} > 0, \quad \frac{1}{n}Z'Z \rightarrow M_{zz} > 0;$$

ii. $G > \max(k, m)$;

iii. the random vectors $w_{s\cdot} = (u_s, v_s), \quad s = 1, 2, \ldots,$ are an i.i.d. sequence with

$$Ew_{s\cdot} = 0, \quad \mathrm{Cov}(w_{s\cdot}) = \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}, \quad \sigma_{12} \neq 0$$

iv. $\lim_{n\to\infty} \frac{n_i}{n} = \alpha_i \in (0,1)$ such that $\alpha_1 + \alpha_2 + \ldots + \alpha_G = 1$.

v. $\lim_{n\to\infty} \bar{x}_i = \xi_i$, $\lim_{n\to\infty} \bar{z}_i = \zeta_i$, $i = 1, 2, \ldots, G$, where the limits are to be understood as ordinary convergence if the variables are not random, or as limits in probability if they are random.

It is an immediate consequence of i and v that

$$\frac{1}{n}\hat{X}'H'H\hat{X} = \sum_{i=1}^{G}\left(\frac{n_i}{n}\right)\bar{x}_i'\bar{x}_i \rightarrow \sum_{i=1}^{G}\alpha_i\xi_i'\xi_i = M_{\xi\xi} > 0. \qquad (8)$$

Notice that, because of Eq. (2), the last diagonal element of M_{xx} is given by

$$m_{(xx),kk} = \gamma' M_{zz}\gamma + \sigma_{22}$$

Remark 1. In most applied problems the matrix H is a primitive, i.e. it is suggested by the nature of the problem investigated. However, because of Eq. (8) H cannot be such that $M_{\xi\xi}$ is singular. This will eventuate when, for example, the **means** $\bar{x}_i$ (or their limits) are all the same, in which case $\xi_i\xi_i'$ will be the same for all i. This will imply that the matrix $M_{\xi\xi}$ is of rank **one**. Another implicit requirement, which will be important at another stage of the argument, is that the division into classes cannot be made on the basis of a(n) (explanatory) variable which is correlated with the structural error. This is so because when H is based on classifications of an explanatory variable independent of the structural error we can assert that

$$\frac{1}{\sqrt{n}} H u \overset{\text{P or a.c.}}{\to} 0, \tag{9}$$

while if it is based on classifications of an explanatory variable, or any other basis, which is correlated with the structural error **we cannot** make the assertion above. At this stage the matrix H should not be thought of as an instrumental matrix, although in some circumstances it may.

The consistency of the estimator (in the sense of (at least) convergence in probability) will be established if we prove that

$$\frac{1}{n}\hat{X}'H'H[u+\beta_k P_z v] \overset{\text{P}}{\to} 0. \tag{10}$$

To show this we first note that

$$\frac{1}{n}\hat{X}'H'H[u+\beta_k v] = \sum_{j=1}^{G}\left(\frac{n_j}{n}\right)\bar{x}_j'[\bar{u}_j+\beta_k\bar{v}_j] \overset{\text{P}}{\to} 0$$

due to the fact that $w_{s\cdot}$ is a sequence of i.i.d. random vectors with mean zero; in fact,

$$\bar{u}_j+\beta_k\bar{v}_j \overset{\text{a.c.}}{\to} 0$$

by Kolmogorov's strong law of large numbers, see Dhrymes (1989), p. 188, and Remark 1.

Next consider

$$\frac{1}{n}\hat{X}'H'HZ\left(\frac{1}{n}Z'Z\right)^{-1}\frac{1}{n}Z'v.$$

The last term converges to zero, at least in probability; thus the proof of consistency will be complete if we can show that

$$\frac{1}{n}\hat{X}'H'HZ \to \sum_{j=1}^{G}\left(\frac{n_j}{n}\right)\xi_j'\zeta_j = M_{\xi\zeta},$$

and the last matrix is well defined, i.e. it has finite elements. But this is evident by the Cauchy inequality and the fact that, due to the assumptions made,

$$\frac{1}{n}Z'H'HZ \rightarrow \sum_{j=1}^{G} \left(\frac{n_j}{n}\right) \zeta_j'\zeta_j = M_{\zeta\zeta} > 0,$$

and it, as well as $M_{\xi\xi}$ have finite elements.[1]

4 Alternative Estimators

Before we proceed to the limiting distribution and questions of relative efficiency, let us set forth the "2SLS" and the OLS estimators using **only** grouped data. The OLS estimator is evidently given by

$$\hat{\beta}_{OLS} = (X'H'HX)^{-1}X'H'Hy = \beta + (X'H'HX)^{-1}X'H'Hu, \qquad (11)$$

while the "2SLS" estimator is given by

$$\tilde{\beta}_{\text{“}2SLS\text{”}} = [(\widetilde{HX})'(\widetilde{HX})]^{-1}(\widetilde{HX})'Hy = \beta + [(\widetilde{HX})'(\widetilde{HX})]^{-1}(\widetilde{HX})'[Hu + \beta_k P_{Hz}Hv]. \qquad (12)$$

That these two estimators are **consistent** may be established as follows. In the case of the OLS estimator we have

$$\frac{1}{n}X'H'HX \xrightarrow{\text{P}} M_{\xi\xi}, \quad \frac{1}{n}X'H'Hu \xrightarrow{\text{a.c.}} 0,$$

which shows consistency. The argument for the consistency of the "2SLS" estimator is essentially the same as that for the Mixed-2SLS. This is so because the former is given by

$$\tilde{\beta} = \beta + [(\widetilde{HX})'(\widetilde{HX})]^{-1}(\widetilde{HX})'[Hu + \beta_k P_{Hz}Hv], \qquad (13)$$

and $\widetilde{HX} = (HX_1, \widetilde{Hx}_{\cdot k})$, where

$$\widetilde{Hx}_{\cdot k} = (Z'H'HZ)^{-1}Z'H'Hx_{\cdot k}. \qquad (14)$$

To see this more clearly, observe that both procedures go through the intermediate step of estimating the vector γ, one using ungrouped data,

[1] Notice that this result requires the subsidiary assumption that, in the limit, the group means ζ_i are not all equal.

the other using grouped data. In either case the resulting estimator (of γ) is consistent.

Remark 3. To simplify the discussion of limiting distribution and relative efficiency issues regarding these estimators we argue as follows: suppose the matrix of "instruments", Z , contains X_1 as a submatrix, i.e. suppose

$$Z = (X_1, P), \quad P \text{ is n } \times (m-k+1) \text{ and independent of } w\,. \tag{15}$$

Consequently, we may write

$$\widetilde{HX} = (HX_1, (HZ)\tilde{\gamma}) = HZ(I^*_{k-1}, \tilde{\gamma}), \quad (\widetilde{HX})' P_{Hz} = 0. \tag{16}$$

When this is the case, the "2SLS" **is equivalent to the OLS estimator**, which implies that the widespread empirical practice of including X_1 as part of the instrumental matrix Z renders the "2SLS" estimator superfluous.

4.1 Limiting Distribution

Since there is a great deal of similarity in the arguments establishing the limiting distribution of all three estimators we shall deal with them simultaneously. Thus,

$$\sqrt{n}(\hat{\beta} - \beta)_{OLS} = \left(\frac{X'H'HX}{n}\right)^{-1} \frac{1}{\sqrt{n}} X'H'Hu; \tag{17}$$

$$\sqrt{n}(\hat{\beta} - \beta)_{\text{“}2SLS\text{”}} = \left(\frac{(\widetilde{HX})'(\widetilde{HX})}{n}\right)^{-1} \frac{1}{\sqrt{n}} (\widetilde{HX})'Hu;$$

$$\sqrt{n}(\hat{\beta} - \beta)_{\text{Mixed}-2SLS} = \left(\frac{\hat{X}'H'H\hat{X}}{n}\right)^{-1} \frac{1}{\sqrt{n}} \hat{X}'H'H(u + \beta_k Pzv); \tag{18}$$

To facilitate discussion introduce the notation

$$\bar{X} = (X_1, Z\gamma) \tag{19}$$

and note that

$$\frac{1}{\sqrt{n}} H(\bar{X} - X) \stackrel{\text{a.c.}}{\to} 0 \tag{20}$$

owing to the fact that

$$\frac{1}{\sqrt{n}} H(\bar{X} - X) = (0, -p), \;\; p = (n_1/n)^{1/2}\bar{v}_1, (n_2/n)^{1/2}\bar{v}_2, \cdots, (n_G/n)^{1/2}\bar{v}_G]', \tag{21}$$

and the group means converge to zero by Kolmogorov's strong law of large numbers. Similarly,

$$\frac{1}{\sqrt{n}} H(\hat{X} - X) = (0, p^*) \stackrel{\text{P}}{\to} 0, \tag{22}$$

because $p^* = \left(\frac{1}{\sqrt{n}}\right) HZ(\hat{\gamma} - \gamma) - p$.

Consequently, by the consistency of the estimator of γ in both Mixed-2SLS and "2SLS" we need only deal with the relations above where X or $\hat{X}$ is replaced by $\bar{X}$. Moreover, since

$$\frac{1}{\sqrt{n}} X'H'Hu = \frac{1}{\sqrt{n}} \bar{X}'H'Hu + \frac{1}{\sqrt{n}}(0, v)'H'Hu, \quad \frac{1}{\sqrt{n}}(0, v)'H'Hu \stackrel{\text{P}}{\to} 0, \tag{23}$$

we need only deal with

$$\sqrt{n}(\hat{\beta} - \beta)_{OLS} \sim \left(\frac{(H\bar{X})'(H\bar{X})}{n} \right)^{-1} \frac{1}{\sqrt{n}} (H\bar{X})'Hu; \tag{24}$$

$$\sqrt{n}(\hat{\beta} - \beta)_{\text{``}2SLS\text{''}} \sim \left(\frac{(\bar{X}H)'H\bar{X}}{n} \right)^{-1} \frac{1}{\sqrt{n}} (H\bar{X})'Hu;$$

$$\sqrt{n}(\hat{\beta} - \beta)_{\text{Mixed}-2SLS} \sim \left(\frac{\bar{X}'H'H\bar{X}}{n} \right)^{-1} \frac{1}{\sqrt{n}} \bar{X}'H'H(u + \beta_k P_z v). \tag{25}$$

Remark 4. As noted earlier in Remark 3, the simplification in the "2SLS" estimator is occasioned by the fact that the matrix of "instruments", Z, contains, as a submatrix, X_1. An earlier version of the paper dealt with the case in which Z is not so restricted, and showed that when we restrict it as in the discussion above we obtain the result just given. Since most practitioners routinely include X_1, we chose to retain only the simplified discussion.

Although evidently the limiting distribution is slightly different for the two cases, the results of the comparison with the Mixed-2SLS estimator are, in substance, precisely the same whether one includes or does not include X_1 as a submatrix of Z.

For the first two estimators, the limiting distribution is determined by the behavior of

$$\frac{1}{\sqrt{n}}(H\bar{X})'Hu \sim (\alpha_1^{1/2}\xi_1', \alpha_2^{1/2}\xi_2', \cdots, \alpha_G^{1/2}\xi_G)'Hu \xrightarrow{\mathrm{d}} N(0, \sigma_{11}M_{\xi\xi}), \quad (26)$$

using the central limit theorem for i.i.d. random variables, see Dhrymes (1988) p. 264. Therefore, the limiting distribution of the OLS and "2SLS" estimators is given by

$$\sqrt{n}(\hat{\beta}-\beta)_{``2SLS} \sim \sqrt{n}(\hat{\beta}-\beta)_{OLS} \sim N(0,\Phi), \quad \Phi = \sigma_{11}M_{\xi\xi}^{-1}. \quad (27)$$

This development makes clear that, in the context of the problem as we have formulated it and **using grouped data**, there is **no** reason to employ the "2SLS" estimator.[2]

To deal with the Mixed-2SLS estimator, given the preceding discussion, we need only deal with

$$\frac{1}{\sqrt{n}}\bar{X}'H'H(u+\beta_k P_z v) = \left(\frac{1}{\sqrt{n}}\bar{X}'H'\right)H(u+\beta_k v)$$
$$-\beta_k\left(\frac{\bar{X}'H'HZ}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\frac{1}{\sqrt{n}}Z'v. \quad (28)$$

The equation above makes clear that we are dealing with a sequence of independent non-identically distributed random vectors obeying the Lindeberg condition, see Dhrymes (1989) p. 265; thus, we conclude that

$$\sqrt{n}(\hat{\beta}-\beta)_{\mathrm{Mixed}-2SLS} \xrightarrow{\mathrm{d}} N(0,\Psi), \quad (29)$$

[2]This result may suggest to some that H is an instrumental matrix, even though its origin lies with the manner in which the data becomes available and does reflect an action by the investigator to define an instrumental matrix.

where

$$\Psi = \eta M_{\xi\xi}^{-1} - (\eta - \sigma_{11}) M_{\xi\xi}^{-1} M_{\xi\zeta} M_{zz}^{-1} M_{\zeta\xi} M_{\xi\xi}^{-1},$$

$$\eta = (1, \beta_k) \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} (1, \beta_k)'. \tag{30}$$

To render the Mixed-2SLS estimator fully operational for inference purposes we need to produce an estimator for Ψ. The major problem in this connection is to produce estimators for σ_{11} and σ_{12} since the estimators for the other quantities are readily available.

Define

$$\hat{u}^* = H(y - \hat{X}\hat{\beta}) = H(u + \beta_k \hat{v}) - H\hat{X}(\hat{\beta} - \beta) = P_{H\hat{X}} H(u + \beta_k \hat{v}),$$

$$\hat{w} = \hat{u}^* - \beta_k P_{H\hat{X}} H\hat{v} = P_{H\hat{X}} Hu. \tag{31}$$

For suitable constants $K_i, \ i = 1, 2$

$$\tilde{\sigma}_{11} = \frac{1}{K_1} \hat{w}'\hat{w}, \quad \tilde{\sigma}_{12} = \frac{1}{K_2} \hat{w}' H\hat{v},$$

are **unbiased** estimators of σ_{11}, σ_{12}, respectively. [3] To see this note that

$$E\hat{w}'\hat{w} = E\mathrm{tr}[H' P_{H\hat{X}} Huu'] = \sigma_{11}(G - k).$$

Thus, for $K_1 = G - k$, $\tilde{\sigma}_{11}$ is an unbiased estimator. Similarly,

$$E\hat{w}' H\hat{v} = E\mathrm{tr}[H' P_{H\hat{X}} HP_z vu'] = \sigma_{12} \mathrm{tr}[H' P_{H\hat{X}} HP_z].$$

Thus, with $K_2 = \mathrm{tr}[H' P_{H\hat{X}} HP_z]$, σ_{12} is an unbiased estimator. Since all entities involved in the calculations above are directly available (save for β_k which is strongly consistently estimable) the problem of estimating the covariance matrix of the limiting distribution is solved.

The preceding fully elucidates the properties of the estimator using grouped and ungrouped data (Mixed-2SLS). For "2SLS consistent estimators for the parameters σ_{11}, σ_{12} do not exist, and by contrast with the

[3]It should be noted that this is strictly true only if $\hat{X}$ is replaced by $\bar{X}$, otherwise it is an approximation owing to the fact that γ is estimated. But for large samples this is a very good approximation.

previous case the same is true for σ_{22}. However, the latter is completely irrelevant for the OLS estimator as well the "2SLS" estimator when X_1 is included as a submatrix of Z.

Unbiased estimators for the first parameter (σ_{11}) may be found by the same method given above.

Remark 5. Typically, consistent estimators of the second moments of the structural errors do not exist; this is due to the fact that with group data we are dealing with (group) means. From

$$Hy = HX\beta + Hu$$

we see that, under our assumptions,

$$\frac{1}{n}(Hu)'Hu = \sum_{j=1}^{G} \left(\frac{n_i}{n}\right) \bar{u}_j^2$$

and, moreover,

$$\bar{u}_j \stackrel{\text{a.c.}}{\to} 0, \quad \text{for all} \quad j.$$

By a property of limits almost certainly (a.c.) or in probability (P)

$$\bar{u}_j^2 \stackrel{\text{a.c.}}{\to} 0.$$

Hence, the sum of squares above converges to zero, which helps explain why estimators of the structural parameter β based on group data are consistent, even with a fixed number of groups.

5 Relative Efficiency

Since the three estimators examined in the previous section are both consistent and asymptotically normal and, moreover, the OLS and "2SLS" are equivalent, the question of relative efficiency entails only a comparison of the covariance matrices of the limiting distribution of the OLS and Mixed-2SLS. Thus, consider

$$\Psi - \Phi = (\eta - \sigma_{11})[M_{\xi\xi}^{-1} - M_{\xi\xi}^{-1} M_{\xi\zeta} M_{zz}^{-1} M_{\zeta\xi} M_{\xi\xi}^{-1}], \tag{32}$$

where $\eta - \sigma_{11} = 2\beta_k\sigma_{12} + \beta_k^2\sigma_{22}$.

We show that in Eq. (32) the matrix in square brackets is positive semi-definite, using a number of results from Dhrymes (2000), chapter 3.

The matrix in question is positive semi-definite if and only if $M_{\xi\xi} - M_{\xi\zeta}M_{zz}^{-1}M_{\zeta\xi} \geq 0$. The latter, however, is the limit (after division by n), of

$$\begin{aligned} A_n &= \bar{X}'H'H\bar{X} - \bar{X}'H'HZ(Z'Z)^{-1}Z'H'H\bar{X} \\ &= \bar{X}'H'[H(I - Z(Z'Z)^{-1}Z')H']H\bar{X}. \end{aligned} \quad (33)$$

The matrix of Eq. (33) is positive semi-definite if the matrix in square brackets is. But $P_z = I - Z(Z'Z)^{-1}Z'$ is a symmetric idempotent matrix of dimension n and rank m (the column dimension of Z). Let J be the matrix of characteristic roots of P_z (which consists of m unities and $n-m$ zeros) and Q the (orthogonal) matrix of characteristic vectors. Then, we have the representation

$$A_n = \bar{X}'H'HQJQ'H'H\bar{X}, \quad (34)$$

which is evidently positive semi-definite for all $n \geq m$. Hence, the limit is also positive semi-definite, thus concluding the proof that the matrix in square brackets of Eq. (32) is positive semi-definite. Consequently, it is immediately evident that $\Phi - \Psi \geq 0$, i.e. the OLS estimator is inefficient relative to Mixed-2SLS, if and only if

$$\eta - \sigma_{11} = 2\beta_k\sigma_{12} + \beta_k^2\sigma_{22} < 0, \quad (35)$$

and it is relatively efficient if and only if

$$\eta - \sigma_{11} = 2\beta_k\sigma_{12} + \beta_k^2\sigma_{22} > 0. \quad (36)$$

Remark 6. In the discussion above, we have shown that the estimator referred to in this literature as “2SLS is, given the standard assumptions on individuals, **asymptotically equivalent** to the OLS estimator using all group data. In different contexts this may not be the case. However, the use of the term “2SLS for such procedures is inappropriate.

The term two stage least squares (2SLS) has a very well defined meaning in the theory of complete systems of simultaneous equations. Precisely, it refers to the procedure of regressing the endogenous variables on

all predetermined (i.e. lagged dependent and exogenous) variables and using the "predicted" values of the dependent variables, so obtained, in the second stage. The procedure widely employed in this literature is more appropriately referred to as "instrumental variables with an ill defined model", see Dhrymes (1994) pp. 106-108; moreover, it is shown therein that when relevant predetermined variables are omitted from the first stage, the resulting structural estimator is (consistent but) inefficient relative to the case where **all** predetermined variables are included; it is **only** in the latter case that it is appropriate to use the term 2SLS. Indeed it may be shown that, in a limited information sense, 2SLS is **an optimal** instrumental variables estimator when the admissible class of instruments is of the form $\{Z : XA\}$, where X is the matrix of **all** predetermined variables and A is a suitable matrix of full column rank. See Dhrymes (1970), Theorem 2, pp. 302-303.

Since in the context of this literature the complete model is not specified, we cannot possibly (logically) assert that the instruments we have used contain **all** the (predetermined) variables that are independent of, or uncorrelated with, the structural error. Thus, we seem to attempt to co-opt the favorable connotations attaching to the term 2SLS, without providing the necessary foundation. It is also evident that we cannot use the argument that such procedures, being 2SLS, are automatically efficient relative to other (IV) limited information estimators.

In fact, the results obtained in this section show that there are cases in which what are referred to as "2SLS are inefficient relative to Mixed-2SLS procedures. This should not be surprising given that the fact that the two procedures use slightly different information.

The intuition behind the result in this section is, roughly speaking, as follows: Using individual data in the first stage utilizes more information and as such contributes to greater efficiency. However, because of subsequent grouping, the (grouped) residuals from that stage are **not necessarily orthogonal** to the grouped variables ($H\hat{X}$) in the second stage, so that the error term in the second stage is, in the derivation of the limiting distribution, different from the original structural error. The variance of the structural error is σ_{11} and, in a limiting sense, we may think of η as the variance of the error term in the second stage.

The result then states that if $\eta - \sigma_{11} < 0$, we have efficiency for the Mixed-2SLS estimator, while if $\eta - \sigma_{11} > 0$ we do not, "because" we have added to the variability of the equation error.

5.1 Monte Carlo Results

Tables 1.1 and 1.2, in Appendix 1, report on the empirical (sampling) distribution of the Mixed-2SLS and "2SLS" estimators. Precisely, from each replication we obtain one estimate of the parameter vector β. We may think of that as one observation from the finite distribution of the two estimators, respectively. By taking their means and standard deviation we give some information about the first two moments of the finite sample distributions. The last column gives the characteristic roots of the difference of the two empirically obtained covariance matrices. The fact that all roots are non-negative confirms the result that one of the estimators is efficient relative to the other, depending on the parametric configuration $2\beta_k\sigma_{12} + \beta_k^2\sigma_{22}$, as obtained by asymptotic theory in the discussion(s) of the previous sections.

Table 1.3 is of the genre as Tables 1.1 and 1.2, except that the sample size is 500, a rather small sample by the standards of the literature. The table shows that the theoretical results obtained by reliance on asymptotic theory continue to hold even for a sample of relatively modest size.

6 Issues arising in Empirical applications

In this section we raise and answer a number of questions of relevance in the empirical implementation of the estimator(s) discussed in this paper. First we address the issue of how the grouping matrix H should be chosen. In the first section we point out the restrictions that H must satisfy for the results on consistency to hold. Then we show that if one has a choice on how to group the data, finer groupings always increase efficiency for either estimator. Although it is evident from the discussion in a previous section that the relative efficiency of the estimators discussed herein does not depend on how the data are grouped (i.e which estimator is more efficient does not depend on H), the efficiency of each estimator does depend on that choice.

Then we go on to address two circumstances that are often encountered in empirical applications. It is common for researchers to use "instruments" that are already defined at the aggregate level. For example several papers use laws defined at the state level as their instruments. We answer the question of whether it is still worthwhile using individual level data in the first stage even in this circumstance.

Finally, we point out that the estimation procedure we labeled mixed-2SLS can be used when matching data from different sources. In other words, one can estimate the first and second stage from different data sources, as is suggested by Angrist and Krueger.

6.1 Choice of the grouping matrix H

In the discussion above we have asserted that, on the assumption

$$n, \quad n_i \longrightarrow \infty \quad \text{such that} \quad \frac{n_i}{n} \longrightarrow \alpha_i > 0,$$

$$\frac{1}{\sqrt{n}} Hu \xrightarrow{\mathrm{P}} 0 \tag{37}$$

where u is a vector of i.i.d. random variables with mean zero and variance $\sigma_{11} > 0$. Can the grouping matrix, otherwise, be chosen arbitrarily? The answer is generally yes, provided the grouping **is not chosen on the basis of a variable that is correlated with** u**.** Although this is generally acknowledged in the oral tradition of this literature, no rigorous derivation of this result is available. We provide a suitable argument to that effect. Thus, let $u_{\cdot i} = (u_{i1}, u_{i2})'$, $i = 1, 2, \ldots, n$ be a sequence of independent identically distributed random vectors with

$$Eu_{\cdot i} = \mu, \qquad \mathrm{Cov}(u_{\cdot i}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} > 0. \tag{38}$$

If we group on the basis of u_2, it means that observations in group i have the property $u_{s2} \in (k_{1i}, k_{2i}]$, for some constants k_{1i}, k_{2i} and $s = n^*_{i-1} + 1, n^*_{i-1} + 2, \ldots, n^*_i$, where $n^*_i = \sum_{j=1}^{i} n_j$; moreover, this holds for all i. To answer the question posed we need to determine the conditional mean of u_1 **given that** $u_2 \in (k_1, k_2]$. Although the argument may be made for an arbitrary distribution, the exposition can be considerably simplified if we assume normality, in which case we have a readily available expression for the conditional distribution. Thus, consider

$$E(u_1|k_1 < u_2 < k_2) = \int_{k_1}^{k_2} f_2(u_2)\left(\int_{-\infty}^{\infty} u_1 f(u_1|u_2)du_1\right)du_2, \tag{39}$$

where $f(u_1|u_2)$ is the conditional density of u_1 given u_2. Carrying out the integration within the large round brackets we obtain

$$\int_{-\infty}^{\infty} u_1 f(u_1|u_2)du_1 = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(u_2 - \mu_2).$$

Carrying out the remaining integration we have to evaluate

$$\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}\frac{1}{\sqrt{2\pi\sigma_{22}}}\int_{k_1}^{k_2}(u_2 - \mu_2)e^{-\frac{1}{2\sigma_{22}}(u_2-\mu_2)^2}du_2 = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}\left[f(k_1) - f(k_2)\right],$$

where f is the density of a normal variable with mean μ_2 and variance σ_{22}. It is, thus, quite evident that unless $\sigma_{12} = 0$, the rightmost member of the equation above cannot possibly be μ_1 for all groups. Therefore the group means of u_1 will not converge to μ_1 if σ_{12} is different than zero.

In Table 2.1, Appendix 2, we verify empirically the results given in this section. The two tables refer to two sets of simulations as follows: In the first table we obtain 1,000 samples (replications) of 10,000 observations each, on the bivariate vector (x_1, x_2), such that their mean is **zero**, and the covariance between them (the parameter σ_{12}) is about .89. The observations in each replication are first ranked on the basis of the **magnitude of** x_2, and divided into 10 groups each containing 1,000 observations. Then we compute the group means for the two variables. The results speak for themselves; even with such great number of observations, the group means of x_1 are "significantly" different from zero for **all** groups.

The second table is constructed in the same manner as the first, except that the covariance between the two variables obeys $\sigma_{12} \approx .1$. In this table, while many of the group means for x_1 are still significantly different from zero, **some are not**. This implies that the inconsistency entailed by grouping based on an "endogenous" variable tends to be less significant the lower the correlation between this variable and the structural error term.

A subsidiary question to be answered next is whether finer or coarser groups are preferable, i.e. given the total number of observations is it better to have a small number of groups, each containing a larger number of observations, or to have a larger number of groups, each containing a smaller number of observations?

6.2 Is finer or coarser grouping more efficient

In this section we answer the question: if the problem and the data permit multiple groupings, i.e. if we can define groups equally well so as to contain more or fewer of the "individual" observations, does it make a difference, in terms of asymptotic efficiency, which is being chosen?

Without loss of generality let us pose the problem as one in which we consolidate two adjacent groups to form a new, larger group. Thus, suppose the initial grouping matrix is H as defined by the discussion surrounding Eq. (5), while after consolidation it is given by

$$H_2 = DH, \quad D = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & d_g \end{bmatrix},$$

$$d_{2s+1\cdot} = \left(\sqrt{\frac{n_{2s+1}}{n_{2s+1}+n_{2(s+1)}}}, \sqrt{\frac{n_{2(s+1)}}{n_{2s+1}+n_{2(s+1)}}} \right), \quad g = \frac{G}{2}, \quad (40)$$

for $s = 0, 1, \ldots, g-1$, on the assumption that G is even. The limiting distribution of the estimator in the two cases is given, respectively, by

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{\mathrm{d}} N(0, \Psi_1), \quad (41)$$

$$\sqrt{n}(\tilde{\gamma} - \gamma) \xrightarrow{\mathrm{d}} N(0, \Psi_2)$$

$$\Psi_1 = \sigma_{22} \operatorname*{plim}_{n\to\infty} \left(\frac{W_1' H' H W_1}{n} \right)^{-1}, \quad \Psi_2 = \sigma_{22} \operatorname*{plim}_{n\to\infty} \left(\frac{W_1' H_2' H_2 W_1}{n} \right)^{-1}.$$

To show that the estimator using finer groups is efficient we need to show that $\Psi_2 - \Psi_1 \geq 0$. Using the results in Dhrymes (2000), chapter 3, it is sufficient to show that

$$W_1' H' H W_1 - W_1' H_2' H_2 W_1 \geq 0, \text{ or alternatively that } H'H - H_2' H_2 \geq 0.$$

The last matrix difference is block diagonal and its s th diagonal block, $s = 0, 1, \ldots, G-1$, is given by

$$A_s = \begin{bmatrix} c\zeta e_{2s+1}' e_{2s+1} & -c e_{2s+1}' e_{2(s+1)} \\ -c e_{2(s+1)}' e_{2s+1} & \frac{c}{\zeta} e_{2(s+1)}' e_{2s+1} \end{bmatrix}, \tag{42}$$

where

$$c = \frac{1}{n_{2s+1} + n_{2(s+1)}}, \quad \zeta = \frac{n_{2(s+1)}}{n_{2s+1}}.$$

Thus, for every s, we can write the matrix A_s above as

$$A_s = c\zeta \left(e_{2s+1\cdot}, \quad -\frac{1}{\zeta} e_{2(s+1)\cdot} \right)' \left(e_{2s+1\cdot}, \quad -\frac{1}{\zeta} e_{2(s+1)\cdot} \right) \geq 0, \tag{43}$$

which is evidently positive semi-definite.

Recalling the discussion of Remark 6, we conclude that in the Mixed-2SLS estimator, if we must use grouped data in the first stage, we need to use the finest possible grouping allowed by the data in order to gain maximum efficiency.

It should be noted that Prais and Aitchison (1954), on entirely intuitive grounds, argue that efficiency increases when observations are grouped as to maximize the between group variance. This is a special case of the result proved here which shows that *ipso facto* finer grouping is more efficient than coarser grouping, without provisos. It is further worth noting that Feige and Watts (1972), working with bank data from the Federal Reserve System find, in a purely empirical sense, that coarser aggregation results in a significant loss of efficiency.

Remark 7. The preceding discussion has established that if fine grouping is used, the first stage of 2SLS is efficient relative to the case where coarser grouping is used. What connection does this have to the estimation of the structural parameter of interest, viz. β? This is answered most easily in the case where only grouped data are used in estimation.

The limiting distribution of the estimator of β with coarse groups is normal with mean zero and covariance matrix

$$\sigma_{11} \operatorname*{plim}_{T\to\infty} \left(\frac{1}{n}\bar{X}' H_2' H_2 \bar{X}\right)^{-1}.$$

The corresponding entity with finer groups is given by

$$\sigma_{11} \operatorname*{plim}_{T\to\infty} \left(\frac{1}{n}\bar{X}' H' H \bar{X}\right)^{-1}.$$

The matrix difference

$$J = \sigma_{11}\bar{X}'[H'H - H_2'H_2]\bar{X}$$

is positive semi-definite if the matrix in square brackets is. But we have shown this to be so in an earlier discussion. Thus, finer groups yield relatively efficient estimators of the structural parameters of interest when only grouped data is used. The same is true for the mixed estimator, but the demonstration of this is too complex to discuss here.

Table 2.2 (Appendix 2) illustrates and confirms, using Monte Carlo simulations, the results obtained by asymptotic theory. We see in particular that with samples of 10,000 observations, increasing the number of groups from 100 to 200 results in substantial increase in precision (lower MSE) for both estimators, while just increasing the sample size (from 10,000 to 20,000 observations), while keeping the number of groups fixed (at 200) does not materially increase the precision of the estimators, i.e. it does not appreciably reduce their MSE.

6.3 Instruments available only at the aggregate level

In this part we analyze the following problem: in the first stage we need to estimate the relationship

$$x_{\cdot k} = Z\gamma + v.$$

Let $Z = (X_1, P^*)$, where P^* is a matrix containing only **exogenous** variables. The problem is that P^* is not available. What we do have is P, which refers to all the exogenous variables **at the aggregate**

(group) level. Do we gain efficiency by "blowing up" such variables to the individual level, and if so how should this be done? Since

$$Hx_{\cdot k} = (HX_1, P)\gamma + Hv$$

is the correct representation of the model in aggregate form, we must define the variables in P at the individual level as $H_1 P = P^*$, so that $HP^* = P$. This implies that we should take

$$H_1 = H' \quad \text{because then} \quad HH_1 = HH' = I_G. \tag{44}$$

In such cases, we take the individual data based model to be

$$x_{\cdot k} = W_1 H_1^* \gamma + v, \quad H_1^* = (I_n, H_1), \quad W_1 = \begin{bmatrix} X_1 & 0 \\ 0 & P \end{bmatrix} \tag{45}$$

By the arguments given earlier, the OLS estimator, $\hat{\gamma}$, for the model in Eq. (45) has the limiting distribution

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{\mathrm{d}} N(0, \Psi_1), \quad \Psi_1 = \sigma_{22} \operatorname*{plim}_{n\to\infty} \left(\frac{W_1' H_1^{*\prime} H_1^* W_1}{n} \right)^{-1}. \tag{46}$$

If we use the aggregate version of the model, viz.

$$Hx_{\cdot k} = (HX_1, P)\gamma + Hv = W_1 H^* \gamma + Hv, \quad H^* = (H, I_G), \tag{47}$$

the OLS estimator from this model, $\tilde{\gamma}$, has the limiting distribution

$$\sqrt{n}(\tilde{\gamma} - \gamma) \xrightarrow{\mathrm{d}} N(0, \Psi_2), \quad \Psi_2 = \sigma_{22} \operatorname*{plim}_{n\to\infty} \left(\frac{W_1' H^{*\prime} H^* W}{n} \right)^{-1}.$$

To determine whether the OLS estimator from the individual model is efficient relative to the one from the aggregate model it is sufficient to establish that

$$J = W_1' H_1^{*\prime} H_1^* W_1 - W_1' H^{*\prime} H^* W_1 \geq 0,$$

see Dhrymes (2000), pp. 89. But

$$J = W_1' \begin{bmatrix} I_n - H'H & 0 \\ 0 & 0 \end{bmatrix} W_1,$$

which is positive semi-definite if

$$I_n - H'H \geq 0. \tag{48}$$

It is easily shown by direct multiplication that

$$I_n - H'H = \text{diag}\left(I_{n_1} - \frac{e'_{1\cdot}e_{1\cdot}}{n_1}, \quad I_{n_2} - \frac{e'_{2\cdot}e_{2\cdot}}{n_2}, \ldots, I_{n_G} - \frac{e'_{G\cdot}e_{G\cdot}}{n_G}\right), \tag{49}$$

where $e_{i\cdot}$ is an n_i-element row vector of unities. Consequently, for **every** i,

$$I_{n_i} - \frac{e'_{i\cdot}e_{i\cdot}}{n_i} \geq 0, \tag{50}$$

which shows that the estimator based on **individual data** is efficient, even though the information in the matrix P is only available at the aggregate level. This is due to the presence of actual individual information as contained in the matrix X_1. As pointed out in Remark 6, efficiency (i.e. a smaller limiting covariance matrix) in the first stage implies efficiency in the second stage for both the "2SLS and Mixed-21SLS estimators.

Evidently in the absence of individual level information, beyond $x_{\cdot k}$, "individual"-based estimators will be identical to aggregate-based estimators! We give a formal demonstration in the remark below.

Remark 8. The model in question is

$$y = x_{\cdot k}\beta_k + u, \quad x_{\cdot k} = H_1 P\gamma + v, \tag{51}$$

where $x_{\cdot k}$ is available in individual data form, while y is only available in group form, i.e. we only have the observations Hy. The precise question is this: do we gain anything by regressing $x_{\cdot k}$ on $H_1 P$, and then using $H\hat{x}_{\cdot k}$ in the second stage? If we follow the procedures just mentioned the OLS estimator of γ is given by

$$\hat{\gamma} = (P'H'_1H_1P)^{-1}P'H'_1x_{\cdot k}. \tag{52}$$

Noting that $H_1 = H'$, we see that $H'_1H_1 = HH' = I_G$, so that

$$\hat{\gamma} = (P'P)^{-1}P'Hx_{\cdot k}, \tag{53}$$

which is **precisely the estimator that would have been obtained** had we implemented the first stage using grouped data! Thus, as claimed earlier, unless individual based information enters the model in the form of instruments, i.e. variables that are independent of (or uncorrelated with) the structural error(s), no efficiency gain is obtained by using individual based data in the estimation of the first stage. A claim to that effect (without demonstration) is also noted in Prais and Aitchison (1954).

6.4 Data available from different sources

In case all requisite data are not available from the same source, can we combine data from different sources to estimate the parameters of the problem? The answer is yes, provided these diverse sources pertain to the same universe, i.e. the data generating function for all relevant sources pertains to the same model. To see why this is so, revert to the equation defining the Mixed 2SLS estimator, i.e.

$$\hat{\beta} = (\hat{X}'H'H\hat{X}')^{-1}\hat{X}'H'Hy$$

Thus, for example, if the constituent data in matrix $\hat{X}$ are available from one source, and Hy only is available from another source, combining these two sources enables us to obtain the mixed 2SLS, as we did earlier, provided the data in Hy refer to the same universe, or data generating function. Indeed, since the properties of estimators depend on the limits of data moment matrices, it matters little, in principle, whether all moments come from the same sample, or from different samples, provided the constituent data in the moments refer to the same process. Thus, for example, if an estimator is of the form, say $(X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$, it matters little if the moment $Z'y$ comes from a sample consisting of (y, Z), while the moments $X'Z$, $Z'Z$ come from another sample, say (X, Z), because the properties of the estimators depend only on the limits of these moments, which will be the same whether obtained from sample 1 or sample 2, provided both samples refer to the same data generating function or universe. Angrist and Krueger (1992) refer to this estimator to as a **two sample IV estimator**. It is precisely the same estimator we would have gotten were it possible to use only one sample. In a sense, we are not producing a different estimator, we are merely obtaining the moments required by this (same) estimator from two different sources.

Indeed, if need be, we could utilize more than two sources!

Note that one can think of the mixed 2SLS estimator as a two-sample IV estimator where the first stage uses individual level data and the second stage uses aggregate data. The contribution of this paper is simply to have shown that even when all the data are available from the same sample at the aggregate level, one might gain efficiency by utilizing more dissagregated data in the first stage, even if the latter comes from a different sample.

7 Conclusions

This paper has derived the properties of an IV estimator that can be obtained when the dependent variable is only available for groups, whereas the endogenous regressor(s) is available at the individual level. In this situation it might be possible to gain efficiency by estimating the first stage using the available individual data, and then estimating the second stage at the aggregate level. This estimation procedure yields a consistent and asymptotically normal estimator that we refer to as Mixed-2SLS. Depending on the parametric configuration of the model, the Mixed-2SLS estimator can be more or less efficient than standard "2SLS, which uses only aggregate data. In fact, given the standard assumptions on an individual based model, the "2SLS using only aggregate (group) data is asymptotically equivalent to the OLS estimator (using only grouped data).

Simulation results confirm our theoretical findings.

REFERENCES

Angrist, Joshua D. (1991), "Grouped Data Estimation and Testing in Simple Labor Supply Models," *Journal of Econometrics*, vol. 47, pp. 243-66

Angrist, Joshua D. (1990),"Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, vol. 80, Issue 3, pp. 313-36

Angrist, Joshua D. and Alan B. Krueger (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples", *Journal of the American Statistical Association*, vol. 87, pp. 328-36.

Blackburn, McKinley (1997), "Misspecified skedastic functions in grouped-data models," *Economic Letters* vol. 55, pp. 1-8

Cramer, J.S. (1964), "Efficient Grouping, Regression and Correlation in Engel Curve Analysis", *Journal of the American Statistical Association*, vol. 59, pp. 233-250.

Dhrymes, P. J. (1970) *Econometrics: Statistical Foundations and Applications*, New York, Harper and Row.

Dhrymes, P.J. (1989), *Topics in Advanced Econometrics: vol. I, Probability Foundations*, New York, Springer-Verlag.

Dhrymes, P.J. (1994), *Topics in Advanced Econometrics: vol. II Linear and Nonlinear Simultaneous Equations*, New York, Springer-Verlag.

Dhrymes, P. J. (2000), *Mathematics for Econometrics*, third edition, New York, Springer-Verlag.

Dickens, William T. (1990), "Error Component in grouped data: Is it ever Worth Weighting?", *Review of Economics and Statistics*, vol. 72, pp. 328-333.

Dee, Thomas S. and William N. Evans (1999), "State Alcohol Policies, Teen Drinking and Traffic Fatalities", *Journal of Public Economics*, Vol. 72, pp. 289-315.

Hoxby, Caroline and M. Daniele Paserman (1998), "Overidentification Test with Grouped Data," NBER Technical Working Paper No. 223.

Im, Kyung So (1998), "Efficient Estimation with Grouped Data", *Economics Letters*, vol. 59, pp. 169-74.

Lleras-Muney, A. (2001),"The effect of Education on Mortality: An Analysis for the US using a Unique Social Experiment," Dissertation Chapter, Columbia University.

Moffit, Robert A. (1995), "Selection Bias Adjustment in Treatment-Effect Models as a method of Aggregation," NBER Technical Working Paper No. 187.

Prais S.J. and J. Aitchison (1954), "The Grouping of Observations in Regression Analysis," *Journal of the International Statistical Institute*, No. 22.

Prichett, Lant and Lawrence H. Summers (1996), "Wealthier is Healthier," *Journal of Human Resources*, 31(4) pp. 841-68.

Shore-Sheppard, Lara D. (1996) "The Precision of Instrumental Variables Estimates with Grouped Data", Industrial Relations Section Working Paper No. 374, Princeton University.

Victoria-Feser, Maria Pia and Elvezio Rochetti (1997), "Robust Estimation for Grouped Data", *Journal of the American Statistical Association*, vol. 92 pp. 333-40

Winter Ebmer, Rudolph and Raphael Steven (1999), "Identifying the Effect of Unemployment on Crime," Center for Economic Policy Research Discussion Paper 2129.

APPENDIX 1

TABLE 1.1

Mixed efficient ($2\beta_k\sigma_{12} + \beta_k^2\sigma_{22}) = -33.75 < 0$)

Number of observations is 50,000; number of replications is 1,000.

β	Mixed ($\widehat{\beta}$)		2SLS ($\tilde{\beta}$)		λ
	Mean	sd	Mean	sd	V($\tilde{\beta}$)-V($\widehat{\beta}$)
0.03	0.033	0.043	0.033	0.046	0.0007
0.02	0.022	0.037	0.024	0.045	0.0007
1	0.999	0.031	0.999	0.035	0.0002
-0.5	-0.496	0.048	-0.495	0.054	0.00006
-0.8	-0.797	0.048	-0.797	0.053	0.0004
0.5	0.499	0.007	0.499	0.008	0.000

TABLE 1.2

2SLS efficient ($2\beta_k\sigma_{12} + \beta_k^2\sigma_{22}) = 44.75 > 0$)

Number of observations is 50,000; number of replications is 1,000.

β	Mixed ($\widehat{\beta}$)		2SLS ($\tilde{\beta}$)		λ
	Mean	sd	Mean	sd	V($\widehat{\beta}$)-V($\tilde{\beta}$)
0.03	0.027	0.053	0.027	0.049	0.0009
0.02	0.0021	0.052	0.020	0.044	0.0008
1	0.999	0.039	0.999	0.034	0.0004
-0.5	-0.500	0.062	-0.500	0.055	0.00008
-0.8	-0.796	0.056	-0.796	0.051	0.00000
0.5	0.499	0.009	0.499	0.008	0.0006

TABLE 1.3

Small Sample Properties:

Number of observations is 500; number of groups is 100

Mixed is efficient. Number of replications is 1,000.

β	Mixed ($\widehat{\beta}$)		2SLS ($\tilde{\beta}$)		λ
	Mean	sd	Mean	sd	V($\tilde{\beta}$)-V($\widehat{\beta}$)
0.03	0.014	0.419	0.015	0.438	0.040
0.02	0.022	0.248	0.023	0.279	0.026
1	0.996	0.254	0.998	0.280	0.020
-0.5	-0.498	0.433	-0.496	0.469	0.005
-0.8	-0.790	0.462	-0.790	0.494	0.000
0.5	0.500	0.072	0.499	0.077	0.015

APPENDIX 2

TABLE 2.1

Group Means of x_1 when Observations are Grouped by the magnitude of x_2.

Number of observations is 10,000; number of replications is 1,000.

High Correlation Case, i.e. $Ex_1 = Ex_2 = 0\,, \quad Ex_1x_2 \simeq .89$

	x_2			x_1		
Group	Means	sd	t	Means	sd	t
1	-7.849	0.084	-92.968	-1.570	0.023	-67.219
2	-4.673	0.066	-70.427	-0.935	0.019	-49.021
3	-3.031	0.060	-50.835	-0.606	0.019	-32.466
4	-1.731	0.056	-30.642	-0.346	0.019	-18.450
5	-0.565	0.057	-9.994	-0.113	0.019	-6.086
6	0.562	0.057	0.057	9.914	0.112	6.131
7	1.725	0.057	30.464	0.346	0.018	19.123
8	3.027	0.059	51.205	0.605	0.019	32.284
9	4.669	0.067	69.900	0.933	0.020	46.676
10	7.845	0.088	88.809	1.569	0.023	68.216

Low Correlation Case, i.e. $Ex_1 = Ex_2 = 0\,, \quad Ex_1x_2 \simeq .1$

	x_2			x_1		
Group	Means	sd	t	Means	sd	t
1	-17.636	0.198	-89.290	-0.175	0.032	-5.533
2	-10.499	0.144	-73.089	-0.105	0.032	-3.232
3	-6.806	0.133	-51.230	-0.067	0.032	-2.095
4	-3.87	0.129	-30.046	-0.037	0.032	-1.128
5	-1.259	0.122	-10.324	-0.016	0.032	-0.510
6	1.263	0.121	10.431	0.014	0.030	0.476
7	3.883	0.124	31.344	0.041	0.033	1.254
8	6.807	0.131	51.920	0.067	0.030	2.197
9	10.497	0.143	73.599	0.103	0.032	3.220
10	17.629	0.193	91.199	0.175	0.032	5.470

TABLE 2.2

Aggregation Trade-off:
Group size versus number of groups

1,000 replications-Mixed is efficient

Mixed Estimator

β	n=10,000 g=100			n=10,000 g=200			n=20,000 g=200		
	Mean	sd	MSE	Mean	sd	MSE	Mean	sd	MSE
0.03	0.032	0.102	0.010	0.037	0.099	0.010	0.033	0.068	0.005
0.02	0.021	0.084	0.007	0.024	0.078	0.006	0.022	0.058	0.003
1	0.998	0.068	0.005	0.997	0.063	0.004	1.002	0.046	0.002
-0.5	-0.496	0.111	0.012	-0.496	0.109	0.012	-0.501	0.079	0.006
-0.8	-0.8	0.107	0.012	-0.791	0.103	0.011	-0.798	0.074	0.006
0.5	0.499	0.017	0.000	0.499	0.016	0.000	0.499	0.012	0.000

2SLS Estimator

β	n=10,000 g=100			n=10,000 g=200			n=20,000 g=200		
	Mean	sd	MSE	Mean	sd	MSE	Mean	sd	MSE
0.03	0.032	0.107	0.012	0.033	0.105	0.011	0.033	0.072	0.005
0.02	0.032	0.098	0.010	0.025	0.091	0.008	0.022	0.068	0.005
1	0.998	0.077	0.006	0.997	0.071	0.005	1.003	0.052	0.003
-0.5	-0.495	0.124	0.016	-0.495	0.121	0.015	-0.501	0.087	0.008
-0.8	-0.800	0.117	0.014	-0.791	0.112	0.013	-0.798	0.080	0.006
0.5	0.499	0.019	0.000	0.499	0.018	0.000	0.499	0.013	0.000