Econometric Theory, **22**, 2006, 537–541. Printed in the United States of America.
DOI: 10.1017/S0266466606060270

A NOTE ON IDENTIFICATION WITH AVERAGED DATA

José A.F. Machado
Faculdade de Economia, Universidade Nova de Lisboa

J.M.C. Santos Silva
ISEG, Universidade Técnica de Lisboa

In most cases where estimation with averaged data is performed, interest lies on the parameters of a model at the individual level, but grouped data are used because disaggregate data are not observed. In this note we study the conditions under which it is possible to consistently estimate the parameters of the individual data model using averaged data, giving particular attention to the case of endogenous selection into groups.

1. MOTIVATION

Estimation with averaged data is often used in practice, and the subject is covered in many standard textbooks (e.g., Kmenta, 1986). In this situation, typically, interest lies in the parameters of a model at the individual level, but estimation is performed using data averaged across different groups because individual data are not observed. An issue that has apparently been neglected in this literature is that selection into groups can be endogenous and in this case consistency can only be achieved under very special conditions.

The problems that arise with estimation using averaged data are akin to those encountered when using stratified samples. In both cases, the way the data are obtained generally leads to some groups or strata in the population being over- or underrepresented in the sample, and standard estimation methods can only be employed if the selection into the groups is exogenous. The problem addressed in this note is, therefore, the following: under which conditions is it possible to consistently estimate the parameters of the individual data model using averaged data?

We are grateful to an anonymous referee for helpful comments and suggestions. We also thank Pedro Duarte Neves, Les Godfrey, Paulo Parente, Maximiano Pinheiro, and Pedro Portugal for helpful discussions. The usual disclaimer applies. Machado is consultant for the Research Department of Banco de Portugal. Santos Silva is grateful for the hospitality, working conditions, and financial support provided by Banco de Portugal, which made this work possible. The authors also gratefully acknowledge partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER. Address correspondence to João Santos Silva, ISEG, R. do Quelhas 6, 1200 Lisboa, Portugal; e-mail: jmcss@iseg.utl.pt.

© 2006 Cambridge University Press 0266-4666/06 $12.00

https://doi.org/10.1017/S0266466606060270 Published online by Cambridge University Press

2. MAIN RESULTS

Consider the case in which the researcher is interested in estimating a vector of parameters whose value in the population is defined by

$$E[f(x)u(y,x,\beta_0)] = 0, \tag{1}$$

where y is the variate of interest, x is a vector of explanatory variables, β_0 is a vector of parameters, and $u(\cdot,\cdot,\cdot)$ and $f(\cdot)$ are known functions. With individual data, $\{y_i,\ x_i\}_{i=1}^n$, β_0 could be consistently estimated by the method of moments under very general conditions on the conditional distribution of y given x, which are now assumed to hold.

Assume, however, that data on $i = 1,\ldots,n$ individuals are only available in the form of averages across G groups or strata. That is, defining the g-group mean of any random variable z as

$$\bar{z}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} z_{gi} = \frac{1}{p_g}\frac{1}{n}\sum_{i=1}^{n} z_i S_{gi}, \qquad g = 1,\ldots,G,$$

where S_{gi} denotes an indicator variable that equals 1 if observation i belongs to group g, being 0 otherwise, $n_g = \sum_{i=1}^n S_{gi}$, $p_g = n_g/n$, and the researcher observes $\bar{y}_g$ and $\bar{x}_g$. The splitting of the population into G groups can explicitly be taken into account by writing equation (1) as

$$\sum_{g=1}^{G} \pi_g E[f(x)u(y,x,\beta_0)|S_g = 1] = 0, \tag{2}$$

where $\pi_g = \Pr(S_g = 1)$. Clearly, (2) does not imply that the orthogonality condition (1) holds for every strata.

When $E[u(y,x,\beta_0)|x,S_g = 1] \neq E[u(y,x,\beta_0)|x]$,[1] the strata indicators S_g are endogenous. In this case it is possible that, for some g, $E[f(x)u(y,x,\beta_0)|S_g = 1] \neq 0$. When the strata indicators are exogenous, (2) implies that $E[f(x)u(y,x,\beta_0)|S_g = 1] = 0$ for every group.

In what follows, we study the conditions under which it is possible to estimate β_0 when only $\bar{y}_g$ and $\bar{x}_g$ are observed, giving special attention to the case in which the strata indicators are endogenous. The analysis proceeds under the following set of additional assumptions.

A1. Let $u(y,x,\beta_0) = y - x'\beta_0$ and $f(x) = x$.[2]

A2. The sample space is partitioned into G groups, and the means $(\bar{y}_g,\bar{x}_g)$ are computed from random samples of size n_g of the pair (y,x), drawn from each group.

A3. $E(y|S_g = 1) \equiv \mu_{y_g}$ and $E(x|S_g = 1) \equiv \mu_{x_g}$ are finite.

A4. $\sum_{i=1}^G \mu_{x_g}\mu'_{x_g}$ and $\sum_{i=1}^G \pi_g \mu_{x_g}\mu'_{x_g}$ are both nonsingular.

https://doi.org/10.1017/S0266466606060270 Published online by Cambridge University Press

Under A2, and given the existence of moments, the sample means $\bar{z}_g$ converge a.s. ($n_g \to \infty$, $\forall g$, G fixed) to

$$\frac{1}{\pi_g} E(zS_g) = \frac{1}{\pi_g} \pi_g E(z|S_g = 1) = \mu_{z_g}.$$

With this notation, the frequently used least squares estimator on group means can be represented as

$$\hat{\beta} - \beta = \left(\sum_{g=1}^{G} \bar{x}_g \bar{x}_g' \right)^{-1} \sum_{g=1}^{G} \bar{x}_g \bar{u}_g$$

and converges a.s. as $n_g \to \infty$, $\forall g$, to

$$\left(\sum_{g=1}^{G} \mu_{x_g} \mu_{x_g}' \right)^{-1} \sum_{g=1}^{G} \mu_{x_g} E(u|S_g = 1).$$

A sufficient condition for this to be zero is that $E(u|S_g = 1) = 0$, $\forall g$.

That is, ordinary least squares (OLS) on group means is consistent when S_g is exogenous. When that is not the case, generally it is not possible to estimate β_0 consistently. This is because averaged data always depend on the strata indicators and it is not possible to construct moment conditions identifying β_0 because all the variables are then endogenous.

It is now interesting to consider the special case in which the covariates are constant within groups. Formally, this means that the conditional distribution of the covariates given the strata indicator is degenerate and $\bar{x}_g = x_g (= \mu_{x_g})$ $\forall g$. The crucial term in the convergence analysis can now be written as

$$\sum_{g=1}^{G} x_g \bar{u}_g,$$

which converges to

$$\sum_{g=1}^{G} x_g E(u|S_g = 1) = \sum_{g=1}^{G} E(xu|S_g = 1).$$

Although this is zero under exogenous strata indicators, that is not the case with endogenous stratification. Nevertheless, it is easy to show that the often recommended weighted least squares estimator, defined by[3]

$$\hat{\beta}_W - \beta = \left(\sum_{g=1}^{G} p_g x_g x_g' \right)^{-1} \sum_{g=1}^{G} p_g x_g \bar{u}_g,$$

is consistent in this case. Indeed, by the law of large numbers,

$$\sum_{g=1}^{G} p_g x_g x_g' \to \sum_{g=1}^{G} \pi_g x_g x_g' \quad \text{a.s.,} \qquad n_g \to \infty,$$

https://doi.org/10.1017/S0266466606060270 Published online by Cambridge University Press

and

$$\sum_{g=1}^{G} p_g x_g E(u|S_g = 1) \to \sum_{g=1}^{G} \pi_g x_g E(u|S_g = 1) = E(xu) = 0 \quad \text{a.s.}, \qquad n_g \to \infty,$$

where the last equality follows from (1). Consequently, the estimator is consistent, whether or not the strata indicators are exogenous.

It is important to notice that, when the S_g's are endogenous, the weighting does not work when the covariates are not constant within groups because then

$$\sum_{g=1}^{G} p_g \bar{x}_g \bar{u}_g \to \sum_{g=1}^{G} \pi_g E(x|S_g = 1) E(u|S_g = 1),$$

which in general differs from

$$0 = \sum_{g=1}^{G} \pi_g E(xu|S_g = 1) = \sum_{g=1}^{G} \pi_g E[xE(u|x, S_g = 1)|S_g = 1].$$

Therefore, under endogenous selection into groups, consistency can be achieved only when weighted least squares are used and the covariates are constant within groups. When the strata indicators are exogenous, both least squares and weighted least squares are consistent, whether or not the covariates are constant within groups. The choice between the two estimators for the case of exogenous strata indicators has been discussed at length in the literature (e.g., Kmenta, 1986; Dickens, 1990).

3. DISCUSSION

Averaged data are often used in hedonic models where estimation is typically performed by regressing the average price of a good (or its logarithm) on a set of its characteristics (Berndt, 1991, Ch. 4). The number of observations in each group depends on the quantity sold of each of the products considered. Therefore, basic economics suggests that the size of the groups can be endogenous. If group sizes are indeed endogenous, our results show that consistent estimation of the parameters in a hedonic regression requires three conditions: (i) the characteristics of the good are identical within groups; (ii) $u(y, x, \beta_0)$ is linear in y; (iii) weighted least squares is used.[4]

Related examples are some of the wage curves estimated by Blanchflower and Oswald (1994), whose Table 4.26 (p. 168) reports results for different models in which "all variables, including the dependent variable, are measured as the mean of all observations in a year/region cell." The dependent variable of these models "is the log of the average annual wage in each cell." If the number of workers in each group is not conditionally independent of the wage, estimation of the parameters of the equation at the individual level is not possible because the regressors are not constant within groups. Moreover, even if the

https://doi.org/10.1017/S0266466606060270 Published online by Cambridge University Press

group sizes are exogenous, this inconsistency will persist because of the nonlinear transformation of the dependent variable. Therefore, it is not surprising that when comparing the results obtained using individual data with those of the averaged data model the authors find important differences. Specifically, the authors conclude that using averaged data "reduces the coefficient on unemployment by approximately one half" (Blanchflower and Oswald, 1994, p. 169).

Finally, it is important to note that whether the selection into the groups is endogenous or not is an empirical question. A formal test for the hypothesis that the strata indicators are exogenous can be performed by comparing weighted and ordinary least squares estimators, as in White (1980), with the appropriate weights being the group sizes. Further details on the construction of the test and an empirical illustration of the results can be found in Machado and Santos Silva (2001).

NOTES

1. This condition is equivalent to $\Pr(S_g = 1|y, x) \neq \Pr(S_g = 1|x)$.

2. More generally, $u(y, x, \beta_0)$ may be assumed to be nonlinear in all random variables with degenerate distributions. The important case in which the distribution of x is degenerate is addressed in Section 2.

3. This estimator can be interpreted as an instrumental variables estimator with instruments defined by $p_g x_g$. It is also worth noting that this estimator is numerically equal to the one that would be obtained by running a least squares regression of y_i on x_i, if individual data were available.

4. Notice that the second condition is not necessary if y is also constant within groups.

REFERENCES

Berndt, E.R. (1991) *The Practice of Econometrics Classic and Contemporary*. Addison-Wesley.

Blanchflower, D.G. & A.J. Oswald (1994) *The Wage Curve*. MIT Press.

Dickens, W.T. (1990) Error components in grouped data: Is it ever worth weighting?" *Review of Economics and Statistics* 72, 328–333.

Kmenta, J. (1986) *Elements of Econometrics*, 2nd ed. Maxwell Macmillan International Editors.

Machado, J.A.F. & J.M.C. Santos Silva (2001) Identification with Averaged Data and Implications for Hedonic Regression Studies. Banco de Portugal, Economic Research Department, WP10-01 (www.bportugal.pt/publish/wp/2001-10.pdf).

White, H. (1980) Using least squares to approximate unknown regression functions. *International Economic Review* 21, 149–170.

https://doi.org/10.1017/S0266466606060270 Published online by Cambridge University Press