Review of Economic Studies (1994) 61, 655-680
0034-6527/94/00320655$02.00

Combining Micro and Macro Data in Microeconometric Models

GUIDO W. IMBENS
Harvard University

and

TONY LANCASTER
Brown University

First version received January 1993; *final version accepted February* 1994 (*Eds*)

Census reports can be interpreted as providing nearly exact knowledge of moments of the marginal distribution of economic variables. This information can be combined with cross-sectional or panel samples to improve accuracy of estimation. In this paper we show how to do this efficiently. We show that the gains from use of marginal information can be substantial. We also discuss how to test the compatibility of sample and marginal information.

1. INTRODUCTION

In this paper we investigate how those macro data that can be viewed as aggregates of micro data can be used in combination with micro or survey data to improve estimates of microeconometric models. We are interested in the parameters characterizing the relationship between the micro variables, and view the aggregate data as useful only as far as they can tell us something about that relationship. We show that in general macro data are indeed useful, even in complex micro models. In fact we will mainly work with general nonlinear models, only occasionally using linear examples to make a particular point more transparent. The key assumption linking the two types of data is that the aggregate variables give information with little or no sampling error about the average of the micro variables.[1] Examples of such macro data are the national employment rate, possibly for different age categories, data on the frequency of unemployment spells of particular length, and aggregate consumer expenditure for various goods. Many of these are published on a yearly basis by the U.S. Bureau of the Census in the Statistical Abstract of the U.S. We will argue that such data are therefore widely available and that they can be of considerable use to micro-econometricians.

About the simplest example of what we shall study is provided by the following. Suppose that, given a regressor x (say income), a dependent variable y (say expenditure on food), is normally distributed with mean $E(y|x)=\beta_0+\beta_1 \cdot x$ and variance σ^2. From a random sample of size N from the population of U.S. households β can be estimated consistently and efficiently by ordinary least squares. Suppose that from the

1. Stoker (1985) makes the same assumption in an analysis of the sources of movements in aggregate dependent variables. See also Jorgenson, Lau and Stoker (1982).

Statistical Abstract we can learn that the national average household expenditure on food in the U.S. is, say, h^*. We will interpret this as exact knowledge of $E(y)$. This is a reasonable interpretation if N is much smaller than the total number of households in the U.S. Knowledge of h^* gives us a restriction in the form of the following orthogonality condition:

$$0 = h^* - E(y) = h^* - E[E(y|x)] = E[h^* - \beta_0 - \beta_1 \cdot x] \tag{1}$$

The sample analogue of this relation can be combined with the scores from the sample data in an efficient generalized method of moments procedure to provide improved estimates of the parameters of interest, β_0 and β_1.

This particular example is a little misleading. In this case the gain here is confined to estimation of the intercept, β_0, presumably not a very interesting parameter. A richer but still simple example would be where the census information provided the mean food expenditures for households with incomes above and below income c. Let these conditional means be denoted by h_1^* and h_2^*. This leads to the pair of orthogonality conditions

$$0 = \Pr(x > c) \cdot (h_1^* - E(y|x > c)) = E[1_{x>c} \cdot (h_1^* - \beta_0 - \beta_1 \cdot x)],$$

$$0 = \Pr(x \leqq c) \cdot (h_2^* - E(y|x \leqq c)) = E[1_{x \leqq c} \cdot (h_2^* - \beta_0 - \beta_1 \cdot x)],$$

where 1_A is the indicator function, equal to 1 if A is true, and equal to zero otherwise. Again these can be replaced by their sample analogues and efficiently combined with the scores from the sample data. This type of (conditional) marginal information yields much more information about β than knowledge of the mean of y alone.

The present work is quite closely related to the literature on choice-based or, in general, endogenously stratified random sampling which dates back at least to Manski and Lerman (1977), and on which related recent papers are Imbens (1992) and Imbens and Lancaster (1991) and Lancaster and Imbens (1991). In that literature it is often assumed that the marginal stratum probabilities are known to the investigator. This knowledge can provide non-parametric identification and, for given parametric assumptions, it can improve efficiency of estimation. By contrast, in the present paper we focus mainly on random sampling and consider rather general forms of marginal information.

A second strand of the literature that is relevant to the current problem is the recent literature on combining datasets. Arellano and Meghir (1992), Angrist and Krueger (1992) and Lusardi (1992) estimate models using information from two different micro data sets. Our work differs in two respects. First, we assume that the model is identified using only the micro dataset, whereas in the afore-mentioned studies identification is only obtained by combining the datasets. Second, we do not require that the second dataset consists of individual-level data. In our framework all that is observed from the second data source is a vector of moments. This allows us to use census tabulations without having access to the raw data on which these tabulations are based.

Our work is also related to, but essentially different from, the literature on the use of prior information in econometrics which dates back to Theil and Goldberger (1961) and which is, of course, a major theme of Bayesian econometrics. In this approach there exists extra sample information, possibly stochastic, about known functions of the unknown parameters. In our case we consider exact information about unknown functions of the parameters. An example would be the right-hand side of equation (1). Our information is relevant because the unknown function can be consistently estimated from the

sample data but it is useless without sample data. In the Bayesian case the prior information is useful even without sample data.

Finally, a strand of literature which is related to the problem analysed here is that of missing regressors. A generic example is the following: a researcher has two data sets, one containing observations on y, x_1 and x_2, and the second containing observations on y and x_1. An alternative interpretation of this is as single data set where observations on x_2 are missing at random. This case has been analysed for the linear model by Griliches, Hall and Hausman (1978) and Gourieroux and Monfort (1981), and for the general case by Robins, Hsieh and Newey (1991). The case we investigate differs in two aspects. First, we assume that the second data set is not completely observed. Only some moments from the joint distribution of the variables in the second data set are observed. Second, we will assume throughout most of the paper that the second data set that these moments come from is so much larger than the first data set that sampling eror in the moments can be ignored. This is not an essential difference, but one that explains why the gains that we will report are much larger than those reported by Griliches, Hall and Hausman (1978) who investigate missing data problems where the proportion of missing data is of the same order of magnitude as that of complete data.

A cautionary remark concerning the combination of micro and aggregate data should be made. In most of the analysis we assume that the aggregate variables are known exactly. We discuss how this can be extended to the case where the aggregate information has a known measure of uncertainty associated with it. However, a more fundamental problem is that some of the aggregate data may have been estimated using the same survey data that the investigator is attempting to combine the aggregate data with. For example, national unemployment statistics in the U.S. are based on the Current Population survey (CPS). Combining CPS data with aggregate unemployment statistics would therefore be futile, because essentially all the information in the latter comes from the former. However, routine application of the techniques developed in this paper might lead the investigator to conclude that there is a substantial gain from combining these data sources. In particular the tests for compatibility of aggregate and micro data discussed in this paper would be unlikely to reject the hypothesis that the data are compatible. One should therefore always investigate which sources the aggregate statistics are based on, and in particular whether the survey data to be combined with these aggregate statistics were used in constructing them.

The paper is organized as follows. Section 2.1 contains the general estimation theory. In Section 2.2 we show that the approach leads to efficient estimators in the class of regular estimators. In addition we find that for a leading example efficiency can be achieved with a smaller number of moments than is necessary in general. In Section 2.3 we show how one can test whether the two sources of information, micro and macro, are compatible. Section 2.4 extends the results of Section 2.1 to the case where the macro data do not provide exact information on population moments but instead provided estimates with some measure of uncertainty. In Section 2.5 we discuss the issues that arise when the micro or survey data form a stratified sample. Section 3 contains an application of the techniques developed in this paper. We use micro data from The Netherlands on employment status, age and education to estimate a probit model. We then combine this with aggregate data published in statistical yearbooks to obtain estimates that, for some parameters, are 50 times as accurate as the original estimates that did not utilize the aggregate information. Section 4 analyses the efficiency gains for probit models in a Monte Carlo study to put the results in Section 3 in perspective. Section 5 contains the conclusion and some suggestions for future research.

2. ESTIMATION

2.1. *Theory*

Let y and x be random vectors with joint probability density function

$$f(y, x)=f(y|x; \theta^*) \cdot r(x)\ \theta^* \in \Theta,\ (x, y) \in X \times Y,$$

where $f(\cdot|\cdot;\cdot)$ is known function, θ^* is an unknown parameter and $r(\cdot)$ is the unknown density function of x not involving θ^*. Let $\{(y_n, x_n)\}_{n=1}^N$ be a random sample of independent observations on y and x. If this random sample is all the information one has about θ^*, one can, under standard regularity conditions, estimate θ^* efficiently by maximizing the conditional likelihood function:

$$\hat{\theta}_{ML}=\text{argmax} \sum_{n=1}^N \ln f(y_n|x_n; \theta).$$

We assume that $f(\cdot|\cdot;\cdot)$ is twice differentiable in θ, that $f(\cdot|\cdot;\cdot)$ and its derivatives are measurable in y and x and that θ^* is the unique maximand of $E[\ln f(y|x; \theta)]$. Under these assumptions the maximum likelihood estimator $\hat{\theta}_{\text{ML}}$ satisfies

$$\sqrt{N}(\hat{\theta}_{\text{ML}}-\theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}),$$

where $\mathcal{I}$ is the Fisher information matrix:

$$\mathcal{I}=-E\frac{\partial^2 \ln f}{\partial\theta\,\partial\theta'}(y|x; \theta^*)=E\frac{\partial \ln f}{\partial\theta}(y|x; \theta^*) \cdot \frac{\partial \ln f}{\partial\theta'}(y|x; \theta^*).$$

Another way of interpreting this estimator is as a method of moments estimator[2] with the score function as the moment:

$$\psi_1(y, x; \theta)=\frac{1}{f(y|x; \theta)}\frac{\partial f}{\partial\theta}(y|x; \theta). \tag{2}$$

In this approach there is no need to parameterize the density of x, $r(\cdot)$. Even if we knew the distribution of x exactly, we would not be able to estimate θ^* more precisely. In this view the problem is a semi-parametric one, even if only trivially so, where the reluctance to specify $r(\cdot)$ comes at no extra cost, either computationally or in terms of efficiency.

Now suppose that in addition to the random sample of y and x the researcher knows the expectation h^* of a known function of y and x, $h(y, x)$. We will assume that $h(\cdot)$ is a measurable function of y and x. We defer the discussion on the value and sources of such information till later. First we show how we can utilize this type of information. If we knew the marginal distribution of x, knowledge of h^* would lead to the exact restriction

$$h^*=g(\theta)=\int_x \int_y h(y, x) \cdot f(y|x; \theta) \cdot r(x)dydx.$$

In that case we can maximize the likelihood function subject to the restriction $h^*=g(\theta)$. The asymptotic variance of the constrained maximum likelihood estimator is equal to

$$V=\mathcal{I}^{-1}-\mathcal{I}^{-1}\Gamma_g'(\Gamma_g\mathcal{I}^{-1}\Gamma_g')^{-1}\Gamma_g\mathcal{I}^{-1}, \tag{3}$$

where $\mathcal{I}$ is as before and $\Gamma_g=(\partial g/\partial\theta')(\theta)$. However, in general we do not know the

2. i.e. as a solution to $(1/N)\sum_{n=1}^N \psi(y_n, x_n, \theta)=0$ rather than as argmax $\sum_{n=1}^N \ln f(y_n|x_n; \theta)$.

marginal distribution of x, and we cannot impose the restriction directly.[3] We can, however, impose a stochastic version of the restriction, using the conditional expectation of $h(y, x)$:

$$h^* = Eh(y, x) = E\{E[h(y, x)|x]\} = Eg(x; \theta^*),$$

where $g(x; \theta) = \int_z h(z, x) \cdot f(z|x; \theta)dz$. We impose this stochastic restriction in a generalized method of moments (GMM) framework, with $\psi_1(y\,x; \theta)$ as the first moment, and

$$\psi_2(y, x; \theta) = h^* - g(x; \theta), \tag{4}$$

as a second moment. This is not the only possible choice for an additional moment. One might think of using $\psi_3(y, x, \theta) = h^* - h(y, x)$ in addition to, or instead of, ψ_2. In Section 2.2 it will be shown that the optimal GMM estimator based on ψ_1 and ψ_2 is efficient and that therefore adding a moment such as ψ_3 does not lower the asymptotic variance of the GMM estimator for θ^*. Using ψ_1 and ψ_3 rather than the efficient combination ψ_1 and ψ_2 might still be useful in cases where the numerical evaluation of $g(x; \theta)$ is difficult, and especially in cases where the conditional density $f(y|x)$ is not parametrically specified, which would make it impossible to compute the integral $\int h(z, x) f(z|x)dz$.

The generic form of the GMM estimator is (Hansen, 1982):

$$\hat{\theta}_{\text{GMM}} = \text{argmin}\left[\frac{1}{N}\sum_{n=1}^{N} \psi(y_n, x_n; \theta)\right]' \cdot C \cdot \left[\frac{1}{N}\sum_{n=1}^{N} \psi(y_n, x_n; \theta)\right],$$

with C a non-negative definite matrix with rank at least equal to the dimension of θ. The optimal form of the GMM estimator uses a weight matrix C equal to $E[\psi(y, x; \theta^*) \cdot \psi(y, x; \theta^*)']^{-1}$ (or a sequence of weight matrices C_N converging in probability to this expectation). In that case the asymptotic distribution of the GMM estimator is:

$$\sqrt{N}(\hat{\theta}_{\text{GMM}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, V),$$

where

$$\begin{aligned} V &= \left\{\left[E\frac{\partial \psi}{\partial \theta'}(y, x; \theta^*)\right]'\left[E\psi(y, x; \theta^*) \cdot \psi(y, x; \theta^*)'\right]^{-1} \cdot \left[E\frac{\partial \psi}{\partial \theta'}(y, x; \theta^*)\right]\right\}^{-1} \\ &= [\mathcal{I} + \Gamma_g' \Delta_g^{-1} \Gamma_g]^{-1} = \mathcal{I}^{-1} - \mathcal{I}^{-1}\Gamma_g'(\Gamma_g \mathcal{I}^{-1} \Gamma_g' + \Delta_g)^{-1}\Gamma_g \mathcal{I}^{-1}, \end{aligned} \tag{5}$$

with $\Delta_g = E[h^* - g(x; \theta^*)] \cdot [h^* - g(x; \theta^*)]'$ and $\Gamma_g = (\partial g/\partial \theta')(\theta^*) = E(\partial g/\partial \theta')(x; \theta^*)$. We assume that Δ_g is non-singular. If Δ_g is singular, some of the restrictions implied by knowledge of h^* are perfectly correlated, at least asymptotically. For the first-order efficiency calculations such restrictions can be ignored, and the corresponding elements of the moment vector ψ_2 dropped. The gain in precision in the first representation of (5) is the second term $\Gamma_g' \Delta_g^{-1} \Gamma_g$. It is large if the variance of $g(x; \theta^*)$ is small and the derivative with respect to θ large in absolute value. This representation allows us to separate the information from the conditional distribution of y given x, $\mathcal{I}$, and that from h^* and the marginal distribution of x, $\Gamma_g' \Delta_g^{-1} \Gamma_g$. In the last representation of V, Δ_g represents the

3. If we specify a parametric family for the marginal distribution of x an argument similar to the one for known $r(\cdot)$ would go through. We would then have an exact restriction θ and the parameters of the marginal distribution of x.

value of knowledge of the marginal distribution of x. If Δ_g is small, V is close to the asymptotic variance of the constrained maximum likelihood estimator given knowledge of $r(x)$, given in (3).

Let us look at an example of this type of marginal information:

Example 1. Let y conditional on x have an exponential distribution with density function $\exp(-\theta_0-\theta_1 \cdot x)\exp[-y \cdot \exp(-\theta_0-\theta_1 \cdot x)]$. The information matrix is

$$\mathcal{I}=\begin{pmatrix} 1 & \mu_x \\ \mu_x & \mu_x^2+\sigma_x^2 \end{pmatrix},$$

where $\mu_x=E(x)$ and $\sigma_x^2=E(x-\mu_x)^2$. If we know that the expectation of y is equal to h^*, we can construct a second moment

$$\psi_2(y, x, \theta)=h^*-E(y|x)=h^*-\exp(-\theta_0-\theta_1' \cdot x).$$

The gain in precision from adding this moment is $\Gamma_g' \Delta_g^{-1} \Gamma_g$, with

$$\Delta_g=E[h^*-\exp(-\theta_0-\theta_1 \cdot x)]^2,$$

and

$$\Gamma_g'=E\begin{pmatrix} \exp(-\theta_0-\theta_1 \cdot x) \\ x \cdot \exp(-\theta_0-\theta_1 \cdot x) \end{pmatrix}.$$

To ease the interpretation of the results, we will simplify them by looking at the special case where x has a standard normal distribution.[4] In that case

$$\mathcal{I}=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and the normalized variance of the GMM estimator:

$$V=\frac{1}{\exp(\theta_1^2)+\theta_1^2}\begin{pmatrix} \exp(\theta_1^2)+\theta_1^2-1 & \theta_1 \\ \theta_1 & \exp(\theta_1^2) \end{pmatrix}.$$

The asymptotic variance of $\sqrt{N}(\hat{\theta}_1-\theta_1^*)$ is $1/[1+\theta_1^2\exp(-\theta_1^2)]$, compared to 1 if one only uses the conditional likelihood score ψ_1 and not the marginal mean information. If $\theta_1=0$, there is no efficiency gain. The efficiency gain reaches a maximum at $\theta_1=1$. In that case the variance of $\sqrt{N}(\hat{\theta}_1-\theta_1^*)$ is $1/(1+e^{-1})\approx 0.7$. Knowledge of $E(y)$ reduced the variance of the estimator of θ_1 by up to 30%.

2.2. *Efficiency*

The estimators proposed in Section 2.1 use the knowledge of $h^*=E[h(y, x)]$ in a generalized method of moments framework with one set of moments equal to the conditional likelihood score and the second set of moments equal to $h^*-\int h(y, x)f(y|x; \theta)dy$. This does not exhaust the set of potential moments. One could create a limitless number of moments of the form $(\partial \ln f/\partial\theta)(y|x; \theta) \cdot s(x)$ for any function $s(x)$. Another option is to include a moment of the form $h^*-h(y, x)$. To show

4. Note that the researcher does not know that x has a standard normal distribution. He or she only knows the conditional distribution of y given x (in this case exponential with mean $\exp(-\theta_0-\theta_1 \cdot x)$), and the expectation h^* of $h(y, x)=y$.

that the estimators proposed in Section 2.1 are efficient we follow the approach to semi-parametric efficiency bounds developed by Begun, Hall, Huang and Wellner (1983), and Newey (1990), in the same way as it is used in Imbens (1992). We will first give the intuition, using Chamberlain's (1987) approach, and then present the formal result. Suppose that x has a discrete distribution with known points of support $x^1, \ldots, x^L$ and $\pi_l = \Pr(X = x^l)$. In that case the maximum likelihood estimates for θ and π are characterized by the equations:

$$\sum_{n=1}^{N} \frac{\partial \ln f}{\partial \theta}(y_n|x_n; \hat{\theta}) = 0, \tag{6}$$

$$\sum_{n=1}^{N} \hat{\pi}_l - 1_{\{x_n = x^l\}} = 0 \quad \text{for all } l = 1, \ldots, L. \tag{7}$$

In terms of $\hat{\pi}$ and $\hat{\theta}$, the maximum likelihood estimate for h^* is

$$\hat{h} = \sum_{l=1}^{L} \hat{\pi}_l \int h(y, x^l) \cdot f(y|x^l; \hat{\theta}) dy = \frac{1}{N} \sum_{n=1}^{N} \int h(y, x_n) f(y|x_n; \hat{\theta}) dy. \tag{8}$$

Because we can characterize $\hat{\theta}$ and $\hat{h}$ by (6) and the right-hand side of (8) without reference to π, we can use those equations to estimate θ^* and h^* even if x is not a discrete random variable. The estimators retain their efficiency properties under weak conditions, as shown by Chamberlain (1987). To estimate θ^* if h^* is known we can use a GMM procedure with moments

$$\psi_1(y, x, \theta) = \frac{\partial \ln f}{\partial \theta}(y|x; \theta),$$

$$\psi_2(y, x, \theta) = h^* - \int h(z, x) f(z|x; \theta) dz.$$

The argument that the GMM estimator for θ with h^* known is efficient goes as follows. If x is discrete, and h unknown, the GMM estimator is identical to the ML estimator. The variance of the constrained GMM estimator (i.e. with h^* known) is equal to the variance of the constrained ML estimator (cf. Lemma 1 in the Appendix). Therefore the constrained GMM estimator is fully efficient if x is discrete. Then we can use Chamberlain's argument to prove efficiency for continuous x.

We will assume that Θ and X are compact sets, and that $h(y, x)$ is a bounded function.

Theorem 1. *The difference between the asymptotic covariance matrix $\tilde{V}$ of $\sqrt{N}(\tilde{\theta} - \theta^*)$, where $\tilde{\theta}$ is any regular estimator, and V, the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta}_{\text{GMM}} - \theta^*)$, is a positive semi-definite matrix.*

Proof. See Appendix. ||

In some cases it is possible to reduce the number of moments necessary to achieve efficiency. An important example is the case where the aggregate information is in the form of probabilities for a particular partition of the sample space. Suppose there are partitions $\{X_j\}_{j=1}^{J}$ of the set X, and $\{Y_k\}_{k=1}^{K}$ of the set Y, such that the extra information consists of knowledge of the probabilities

$$h_{jk}^* = \Pr[x \in X_j, y \in Y_k] = E[1_{\{x \in X_j, y \in Y_k\}}],$$

for $j=1,\ldots,J$, and $k=1,\ldots,K$. This type of information might be obtained from frequency tables in national statistics, as the example in Section 3 illustrates. From this type of information one can construct $J\times K-1$ moments of the form

$$\psi_2(y, x, \theta)_{(j-1)\times K+k}=h^*_{jk}-E[1_{\{x\in X_j, y\in Y_k\}}|x]=h^*_{jk}-1_{\{x\in X_j\}}\int_{Y_k} f(z|x;\theta)dz.$$

The $(J\times K)$th moment is superfluous because the probabilities add up to one.

We can represent this information in these $J\times K-1$ moments in a different way. Let $\tilde{h}_{j0}=\sum_k h^*_{jk}$ be the probability of the event $x\in X_j$ and let $\tilde{h}_{jk}=h^*_{jk}/\tilde{h}_{j0}$ be the conditional probability of the event $y\in Y_k$ given $x\in X_j$. In this representation the probabilities $\tilde{h}_{j0}$ represent what we know about the marginal distribution of x, and $\tilde{h}_{jk}$, for $k\geqq 1$, what we know about the conditional distribution. One can construct moments that reflect this division:

$$\psi_2(y, x, \theta)_{(j-1)\times(K-1)+k}$$
$$=\left[\tilde{h}_{jk}-\int_{Y_k} f(z|x;\theta)dz\right]\cdot 1_{\{x\in X_j\}} \quad \text{for } j=1,\ldots,J, \quad k=1,\ldots,K-1,$$
$$\psi_2(y, x, \theta)_{J\times(K-1)+j}=\tilde{h}_{j0}-1_{\{x\in X_j\}} \quad \text{for } j=1,\ldots,J-1.$$

The first set of moments represents the information about the conditional distribution of y given x, the second set represents the information about the marginal distribution of x. The first component of the gain is:

$$\Delta_g=\begin{pmatrix}\Delta_{g11} & 0\\ 0 & \Delta_{g22}\end{pmatrix},$$

where Δ_{g11} is a $J\cdot(K-1)$ dimensional square matrix, and Δ_{g22} a $J-1$ dimensional square matrix. The second component of the gain can be partitioned accordingly:

$$\Gamma'_g=(\Gamma'_{g1} \quad 0),$$

with Γ'_{g1} a dim $(\theta)\times(J\cdot(K-1))$ dimensional matrix, and the second matrix of dimension dim $(\theta)\times(J-1)$. The gain is therefore

$$\Gamma'_{g1}\cdot\Delta^{-1}_{g11}\cdot\Gamma_{g1}.$$

This is identical to the efficiency gain that we obtain if we only use the moments $\psi_1(y, x, \theta)$ and $\psi_2(y, x, \theta)_i$ for $i\leqq J\cdot(K-1)$. The moments that represent the marginal probabilities for the partition of the X space are not informative if we also have conditional probabilities for y given that x is in those sets. Other information about the marginal distribution of x, such as the expectation of x, can still be useful, but this particular type of extra information x is not. If J is large relative to K, this approach can significantly reduce the number of moments necessary to achieve efficiency.

2.3. *Testing*

In this paper we investigate how a random sample can be used in combination with exact information about some of the moments of the variables in order to estimate parameters of the conditional distribution. As indicated in the introduction, one could view this as an abstraction from the practical case where the information on the moments would come

in the form of averages over a large but not infinite population. In that view it stands out more clearly that one is combining information from different sources. One should therefore be concerned with the question whether these sources are compatible. Specifically, is the population from which the random sample is drawn the same as the population from which the sample is drawn that provided the averages? We will refer to this issue as the question of compatibility of the two sources of information. This issue is distinct from the question of specification of the conditional distribution. Both are forms of misspecification (in one case of the sample design and in the other case of the conditional distribution) and can lead to incorrect inference. However, their implications are different. If the two pieces of information are not compatible, there is no point in combining them and one should use only one of them. In the second case combining the two pieces of information is justified, but the specification of the conditional distribution should be improved. Because of the different implications it is important to try to distinguish between them.

We can test compatibility directly by simply comparing h^* to the average of $h(y, x)$ over the random sample. Consider

$$T_c = N \cdot (\bar{h} - h^*)' \cdot V_h^{-1} \cdot (\bar{h} - h^*),$$

where

$$\bar{h} = \frac{1}{N} \sum_{n=1}^{N} h(y_n, x_n),$$

and V_h is the variance of $h(y, x)$, consistently estimated by the average $\sum[h(y_n, x_n) - \bar{h}] \cdot [h(y_n, x_n) - \bar{h}]'/N$. The distribution of T_c is, as N goes to infinity, $\chi^2(\dim(h))$. Its distribution does not depend on the correct specification of the conditional density of y given x. If one wants to test compatibility given that the model is correctly specified one can improve on this test. The optimal test in that case would be to compare h^* to the average of conditional expectation of y given x, $(1/N) \sum_{n=1}^{N} g(x; \hat{\theta})$. This comparison is done implicitly in the GMM estimation procedure. The value of the objective function in that procedure can therefore be used for this test.

One can test the specification of the conditional distribution without using the extra information in a variety of classical ways. One approach is to use the information matrix equality to test parameter constancy (Chesher (1984)).

In addition to testing separately compatibility and specification, it is possible to test the joint hypothesis that both the sample design and the conditional distribution are correctly specified. Such tests, while they might be harder to interpret because they test the overall specification, might have more power than the separate tests. One test statistic, mentioned above, that is calculated as a by-product of the estimation procedure, is

$$T_{cs} = N \cdot \bar{\psi} \cdot V_\psi^{-1} \cdot \bar{\psi},$$

with $\bar{\psi} = \sum_{n=1}^{N} \psi(y_n, x_n, \hat{\theta})/N$ and V_ψ the variance of $\psi(y, x, \theta^*)$. Asymptotically T_{cs} has a χ^2 distribution with $\dim(h) = \dim(\psi) - \dim(\theta)$ degrees of freedom.

The score tests that can be used in a maximum likelihood framework can be adapted to give tests in this GMM framework. For instance, one can add a moment of the type

$$\psi_3(y, x, \theta) = \frac{\partial^2 \ln f}{\partial \theta_i \partial \theta_j}(y|x; \theta) + \frac{\partial \ln f}{\partial \theta_i}(y|x; \theta) \cdot \frac{\partial \ln f}{\partial \theta_j}(y|x; \theta)$$

to the moment vector and test the over identifying restrictions to get a test similar to the information matrix test.

A test that is more specific to these models that combine two sources of information can be based on the fact that the optimal weight matrix in the GMM procedure, $E[\psi \cdot \psi']^{-1}$, is block diagonal. The expectation of $\psi_1(y, x, \theta) \cdot \psi_2(y, x, \theta)$ is zero because the expectation of ψ_1 is zero conditional on x, and ψ_2 does not depend on y. One can exploit this result by adding a moment of the form.

$$\psi_3(y, x, \theta) = \psi_1(y, x, \theta)_i \cdot \psi_2(y, x, \theta)_j.$$

Both the information matrix type tests and those based on zeros in the weight matrix $E[\psi\psi']^{-1}$ can be implemented in a GMM framework with over identifying restrictions. See Newey (1985) for a general discussion of such tests. These moments do not add any efficiency because the estimators are already efficient in a semi-parametric sense, but they test the specification in different ways.

Another test that can be employed to test simultaneously specification and compatibility is a Hausman test (Hausman (1978)). Under the null hypothesis that the data are compatible and conditional density is equal to $f(y|x; \theta)$, the conditional maximum likelihood estimator $\hat{\theta}_{ML}$ is consistent but not efficient. The efficient estimator is under these assumptions $\hat{\theta}_{GMM}$. The normalized different $\sqrt{N}(\hat{\theta}_{GMM} - \hat{\theta}_{ML})$ has under the null hypothesis asymptotically a normal distribution with mean zero and variance

$$V_{\hat{\theta}_{ML} - \hat{\theta}_{GMM}} = V_{\hat{\theta}_{ML}} - V_{\hat{\theta}_{GMM}} = \mathscr{I}^{-1} - V = \mathscr{I}^{-1}\Gamma_g'(\Gamma_g \mathscr{I}^{-1}\Gamma_g' + \Delta_g)^{-1}\Gamma_g \mathscr{I}^{-1}.$$

Assume that $V_{\hat{\theta}_{ML} - \hat{\theta}_{GMM}}$ has full rank. Then the test statistic

$$T_h = N \cdot (\hat{\theta}_{GMM} - \hat{\theta}_{ML})' \cdot V^{-1}_{\hat{\theta}_{ML} - \hat{\theta}_{GMM}} \cdot (\hat{\theta}_{GMM} - \hat{\theta}_{ML}),$$

has asymptotically a χ^2 distribution with dim (θ) degrees of freedom.

2.4. *Stochastic information*

In the preceding sub-sections we have assumed that the extra information is in the form of a vector h^*, which is exactly equal to the expectation of a known function $h(\cdot)$ of the variables y and x. In this section we will generalize the procedures suggested for that case to the case where we do not know h^* with certainty. Suppose we have an estimate $\bar{h}$ of h^*, based on an average of $h(y, x)$ over a second random sample of size M. Based on that random sample the estimate for h^* would satisfy

$$\sqrt{M}(\bar{h} - h^*) \xrightarrow{d} \mathcal{N}(0, \Delta_h)$$

with

$$\Delta_h = E[h(y, x) - h^*] \cdot [h(y, x) - h^*]'.$$

We assume that the extra information is in the form of the estimate $\bar{h}$ and the size of the second data set M. In addition we assume that $\bar{h} - h^*$ is independent of the micro sample $\{(y_n, x_n)\}_{n=1}^N$. We implicitly assume that we do not observe the second random sample of size M directly. If we did, we could combine the data sets and estimate the parameters on the combined data set. The only difference with the model from the preceding subsections is that in this case the extra information is uncertain.

First we will develop an estimation procedure that leads to an estimator for h^* as well as for θ^*, and then we will show that if all we are interested in is an estimate for θ, we can simplify the procedure. The asymptotics will be based on N, the number of observations in the main data set, going to infinity, while M/N converges to some constant k.

Note that from the main sample $\{(y_n, x_n)\}_{n=1}^N$ we can consistently estimate the variance Δ_h as

$$\hat{\Delta}_h = \frac{1}{N}\sum_{n=1}^N (h(y_n, x_n) - \hat{h}) \cdot (h(y_n, x_n) - \hat{h})'$$

where $\hat{h} = \sum h(y_n, x_n)/N$.

If we did not have the estimate $\bar{h}$ of h^* we could estimate θ^* and h^* from the first or main data set in a GMM framework using the moments

$$\psi_1(y, x, \theta, h) = \frac{1}{f(y|x;\theta)} \frac{\partial f}{\partial \theta}(y|x;\theta),$$

$$\psi_2(y, x, \theta, h) = h - \int_z h(z, x) f(z|x;\theta) dz.$$

The estimator for θ is equal to the conditional maximum likelihood estimator and the estimator for h^* is equal to the average of $g(x;\hat{\theta}) = \int_z h(z, x) f(z|x;\hat{\theta}) dz$ over this dataset. This estimator is efficient, which follows from the efficiency argument in Section 2.2. The joint asymptotic distribution of the estimator and $\bar{h} - h^*$ is:

$$\sqrt{N}\begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{h} - h^* \\ \bar{h} - h^* \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathcal{I} & -\mathcal{I}^{-1}\Gamma_g & 0 \\ -\Gamma_g'\mathcal{I}^{-1} & +\Gamma_g'\mathcal{I}^{-1}\Gamma_g & 0 \\ 0 & 0 & \Delta_h/k \end{pmatrix}\right).$$

We can combine this with the estimator for h^* from the second data set, $\bar{h}$, in a minimum distance framework (see Chamberlain (1984) for details). The quadratic form to minimize is

$$\begin{pmatrix} \hat{\theta} - \theta \\ \hat{h} - h \\ \bar{h} - h \end{pmatrix}' \begin{pmatrix} \mathcal{I} + \Gamma_g\Delta_g^{-1}\Gamma_g' & \Gamma_g\Delta_g^{-1} & 0 \\ \Delta_g^{-1}\Gamma_g' & \Delta_g^{-1} & 0 \\ 0 & 0 & \Delta_h/k \end{pmatrix} \begin{pmatrix} \hat{\theta} - \theta \\ \hat{h} - h \\ \bar{h} - h \end{pmatrix}.$$

If we minimize this over h and θ we obtain estimators with the following asymptotic distribution.

$$\sqrt{N}\begin{pmatrix} \hat{\theta} - \theta^* \\ \tilde{h} - h^* \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \mathcal{I} + \Gamma_g\Delta_g^{-1}\Gamma_g' & \Gamma_g\Delta_g^{-1} \\ \Delta_g^{-1}\Gamma_g' & \Delta_g^{-1} + \Delta_h^{-1}k \end{pmatrix}^{-1}\right).$$

If $k = \text{plim}\ (M/N)$ is large, the asymptotic variance of $\tilde{\theta}$ is close to that obtained for the case where we know h^* exactly. If k is small, the knowledge of h^* from the second data set is very inaccurate and we estimate θ^* with about the same accuracy as if we estimated it based on the main data set alone.

An alternative procedure if we are only interested in estimating θ^* and not in h^*, is to use the same moments ψ_1 and ψ_2 that we used with exact knowledge of h^*, and adjust the variance of the second moment. We replace Δ_g in the weight by $\Delta_g + \Delta_h/k$, and the true value h^* by the estimate $\bar{h}$. We then estimate θ^* by minimizing

$$\sum_{n=1}^N \psi(y_n, x_n, \theta, \bar{h})' \cdot \begin{pmatrix} \mathcal{I}^{-1} & 0 \\ 0 & (\Delta_g + \Delta_h/k)^{-1} \end{pmatrix} \cdot \sum_{n=1}^N \psi(y_n, x_n, \theta, \bar{h}),$$

with

$$\psi(y, x, \theta, \bar{h}) = \begin{pmatrix} \dfrac{1}{f(y|x;\theta)} \dfrac{\partial f}{\partial \theta}(y|x;\theta) \\ h - \int_z f(z|x;\theta)dz \end{pmatrix}.$$

Asymptotically the variance of $\sqrt{N}(\hat{\theta} - \theta^*)$ is $[\mathcal{I} + \Gamma_g(\Delta_g + \Delta_h/k)^{-1}\Gamma_g']^{-1}$. This is the same as the asymptotic variance for the estimator proposed above. Uncertainty in the extra information has therefore the same effect as decreasing the weight given to it in the optimal method of moments procedure. The implication is that even if we are not completely sure about the uncertainty in the aggregate data, we can reduce the effect of the aggregate statistics on the estimates by lowering the weight given to ψ_2. Such an approach can be interpreted as implicitly associating a variance with the aggregate statistics. This might be useful in cases where the aggregate data are based on a variety of sources, combined in complex ways. In such circumstances one would expect the aggregate data to be informative, but not represent exact population values.

2.5. *Stratified sampling*

In many cases survey data are not drawn randomly from the overall population, but are drawn randomly within pre-determined strata, with the sample proportions in the strata different from the corresponding population proportions. In this section we investigate the implications of such sampling schemes for the methods we propose for incorporating aggregate information. We focus on two special cases that illustrate the issues that arise in this situation. In both cases we assume that the stratification itself is known, but we allow the population probabilities for each of the strata to be unknown. First we analyse an *exogenous* sampling scheme where the stratification is purely on the covariates x. Second we analyse an *endogenous* sampling scheme where the stratification is on the dependent variable, y. For a general discussion of stratified sampling see Imbens and Lancaster (1991).

2.5.1. *Stratification on exogenous variables*

To simplify notation we focus on the case with two strata. With probability p^* an observation is randomly drawn from that part of the population that is characterized by $x>0$. With probability $1-p^*$ an observation is drawn from the part of the population with $x \leqq 0$. In the population the proportion of units with $x>0$ is equal to q^*. The joint of y and x is

$$f(y, x) = f(y|x;\theta^*) \cdot r(x) \cdot \left(\frac{p^*}{q^*}\right)^{\delta(x)} \left(\frac{1-p^*}{1-q^*}\right)^{1-\delta(x)},$$

where $\delta(x)$ is an indicator for the event $x>0$. The estimator $\hat{\theta}_{ML}$, obtained by maximizing the conditional likelihood, $\sum \ln f(y_n|x_n;\theta)$ is still consistent and efficient in the absence of aggregate information.

Now assume that there is aggregate information in the form of the (population) expectation of $h(y, x)$, h^*. This information can be expressed as

$$h^* = E_p h(y, x) = E_s h(y, x) \left(\frac{q^*}{p^*}\right)^{\delta(x)} \left(\frac{1-q^*}{1-p^*}\right)^{1-\delta(x)}$$

where E_p denotes the expectation over population distribution, and E_s the expectation over the sampling distribution. If q^*, the population proportion of units with $x>0$ is known, we can add to $\psi_1(y, x; \theta)$ given in equation (2) the moments

$$\psi(y, x; \theta, p)=\begin{pmatrix} p-\delta(x) \\ h^*-g(x;\theta)(q^*/p)^{\delta(x)}((1-q^*)/(1-p))^{1-\delta(x)} \end{pmatrix}.$$

If q^* is not known *a priori*, knowledge of h^* allows one to estimate q, but there is no gain in precision with which one estimates θ. Another interpretation is that because in this case one cannot estimate h^* from the micro data alone, knowledge of h^* does not help in estimating the parameters of interest, θ. There is potentially a gain if the dimension of h^* is higher than the dimension of the unknown stratum probabilities.

2.5.2. *Stratification on endogenous variables*

With probability p^* an observation is randomly drawn from the sub-population characterized by $y>0$. With probability $1-p^*$ an observation is drawn from the sub-population with $y\leqq 0$. Let q^* be the population proportion of units with $y>0$, and $\delta(y)$ an indicator for the event $y>0$. The joint density of y and x is:

$$f(y, x)=f(y|x;\theta^*)\cdot r(x) \times \frac{(p^*/q^{*\delta(y)}\cdot((1-p^*)/(1-q^*))^{1-\delta(y)}}{(p^*/q^*)\int_0^\infty f(z|x;\theta^*)dz+((1-p^*)/(1-q^*))\int_{-\infty}^0 f(z|x;\theta^*)dz}.$$

In the absence of aggregate information one can estimate θ, p and q efficiently in a GMM framework using the moments

$$\psi_{11}(y, x; \theta, p, q)=\frac{1}{f(y|x;\theta)}\frac{\partial f}{\partial\theta}(y|x;\theta)-\frac{\left(\left[\frac{p}{q}-\frac{1-p}{1-q}\right]\frac{\partial}{\partial\theta}\int_0^\infty f(z|x;\theta)dz\right)}{\left(\frac{p}{q}\int_0^\infty f(z|x;\theta)dz+\frac{1-p}{1-q}\int_{-\infty}^0 f(z|x;\theta)dz\right)},$$

$$\psi_{12}(y, x; \theta, p, q)=q-\int_0^\infty f(z|x;\theta)dz\Big/\left(\frac{p}{q}\int_0^\infty f(z|x;\theta)dz+\frac{1+p}{1-q}\int_{-\infty}^0 f(z|x;\theta)dz\right),$$

$$\psi_{13}(y, x; \theta, p, q)=p-\delta(y).$$

These moments are a special case of the general formula in Imbens and Lancaster (1991).

Unlike exogenous stratification one can, under endogenous stratification, typically estimate q, the size of the stratum with $y>0$. Therefore one can estimate the expectation of $h(y, x)$ in a GMM framework using the moment.

$$\psi_2(y, x; \theta, p, q, h)=h-g(x;\theta)\Big/\left(\frac{p}{q}\int_0^\infty f(z|x;\theta)dz+\frac{1-p}{1-q}\int_{-\infty}^0 f(z|x;\theta)dz\right).$$

If h^* is known, the second moment ψ_2 is used with the known value h^*, and optimally weighted with the other moments.

The key difference between the two cases, stratification on x or stratification on y, is that in the first case one cannot estimate h^* from the survey data alone, while in the

TABLE I

Summary statistics

	L	E	A
Mean	0·908	2·82	35·7
st. dev.	0·289	1·13	6·82
Corr. with L	1·000	0·003	−0·031
Corr. with E	0·003	1·000	−0·174
Corr. with A	−0·031	−0·174	1·000

second case it is possible to estimate h^* from the micro sample alone. This is a direct consequence from the fact that we assume that the conditional density of y given x is specified, and not the conditional density of x given y. If one can estimate h^* from the micro data alone then knowledge of h^* is typically informative, but if one cannot do so, knowledge of h^* cannot be informative about other parameters.

3. AN APPLICATION

In this section we will apply the econometric procedures proposed in the previous sections, combining a micro dataset on Dutch labour market histories with aggregate information from national statistics. The ORIN dataset, collected by NIDI[5] (Nederlands Interuniversitair Demografisch Institut) in cooperation with the Tilburg University, Groningen University and the University of Wageningen, contains among other things labour market histories for a random sample of Dutch men from 1977 to 1983. We will only use the labour market status at one point in time, January 1977. For this date we constructed a dataset containing information on labour market status (employed or not employed), education (in five categories, indicating increasing levels of education) and age (in years). We will denote labour market status by L, equal to 1 if someone is employed in January 1977 and 0 if not. Education will be denoted by E, and age by A. Only men at least 25 years old and not yet 50 years old were retained. This left us with 347 observations. Table I gives summary statistics for this dataset.

We estimated the following probit model for employment using this dataset:

$$\begin{aligned}\Pr[L=1|A,E] &= \Phi(x'\theta) \\ &= \Phi[\theta_0+\theta_1\cdot E+\theta_2\cdot(A-35)+\theta_3\cdot(A-35)^2],\end{aligned} \tag{9}$$

with

$$x=\begin{pmatrix}1\\ E\\ A-35\\ (A-35)^2\end{pmatrix}$$

and $\Phi(\cdot)$ is the standard normal distribution function. The log-likelihood function is:

$$\ln(\theta)=\textstyle\sum_{n=1}^{N} L_n\cdot\ln\Phi(x_n'\theta)+(1-L_n)\cdot\ln(1-\Phi(x_n'\theta)),$$

5. We wish to thank NIDI for making these data available to us.

TABLE II

Maximum likelihood estimates

Regressor	Estimate	S.E.	T-value
Intercept	1·4372	0·3167	4·54
Education	−0·0090	0·0930	−0·10
Age-35	−0·0021	0·0148	−0·14
(Age-35)^2	−0·0017	0·0023	−0·74

TABLE III

Size of male population in thousands by age category in 1977

Age category	Working	Total
25–29	555	609·0
30–34	499	534·8
35–39	407	436·7
40–44	370	396·8
45–49	337	378·0
Total	2168	2355·3

with the first-order condition

$$0=\sum_{n=1}^{N}\frac{L_n-\Phi(x_n'\hat{\theta})}{\Phi(x_n'\hat{\theta})(1-\Phi(x_n'\hat{\theta}))}\phi(x_n'\hat{\theta})\cdot x_n.$$

Table II gives the maximum likelihood estimates for this model. Note that none of the coefficients (apart from the one corresponding to the intercept) was estimated to be significantly different from zero. In fact a formal test indicates that the hypothesis that all three coefficients θ_1, θ_2 and θ_3 are all zero cannot be rejected with this dataset. The test statistic is equal to 0.83, with a limiting $\chi^2(3)$ distribution. The micro dataset was clearly not large enough to detect much of a relation between education and age on the one hand and labour market status on the other hand. We therefore investigate whether we can combine this dataset with aggregate data to obtain more accurate estimates of such effects.

The first step in this procedure was to extract information from the statistical yearbooks for The Netherlands.[6] In the *B* (population) section of the 1978 yearbook we find information giving the size of the male population in various age categories in 1977. Page 120 of the 1980 yearbook gives the corresponding numbers for working males for the same age categories for 1977. Table III gives the relevant numbers from both tables. We use this information to estimate both the probability of being employed given the age category (denoted by p_i where the index for the age category i runs from 1 to 5) and the probability of being in a particular age category (denoted by q_i). Table IV gives the estimated probabilities. We will interpret the 2.355 million observations (all men between 25 and 50 years old in The Netherlands in 1977) that led to Table III as a random sample from an infinitely large population. The p_i and q_i in Table IV are, in this interpretation, estimates of the true population parameters. However, given the size of the sample that was used in constructing Table III, compared with the micro sample of 347 for which we

6. Handboek voor de Statistiek, published by the Centraal Bureau voor de Statistiek. Similar aggregate information for the United States is published in the Statistical Abstract of the U.S., published by the U.S. Bureau of the Census. For instance, in the 1990 volume, Table 625 on page 378 gives the size of the population and labour force by a number of age intervals, and Table 638 on page 386 gives the employment status by age.

TABLE IV

Estimate of age category and conditional employment probabilities

Age category	p_i	q_i
25–29	0·911	0·258
30–34	0·933	0·227
35–39	0·932	0·185
40–44	0·932	0·168
45–49	0·891	0·160
Total	0·921	1·000

have exact measurements, we will treat the p_i and q_i in Table IV as true probabilities. Formally this is not correct. One could take account of the fact that it is an estimate of the true probability by estimating its variance using the raw numbers in Table III, as outlined in Section 2.4. This is unlikely to be a fruitful strategy in this case. Because the population is about 7000 times as large as the sample, the variance of these estimates is so small relative to the variation in estimates based on the sample, that it will not have much effect on the final results. A second argument is that the raw numbers in Table III are also estimates, based on various surveys, using complicated weighting schemes. Variance estimates taking them as true counts are therefore likely to be biased downward. For these reasons we take as a first approximation the frequencies in Table IV as true probabilities. The question is now, what do these probabilities tell us?

From Table IV we can see that the conditional probability is not constant over the age categories. We cannot tell from this table whether this is due to age effects or education effects, but we can tell that at least some combination of them must be affecting the employment probabilities. In terms of equation (9), Table IV implies that it cannot be true that $\theta_1=\theta_2=\theta_3=0$. It may however be true that $\theta_2=\theta_3=0$ and that the relation between age and employment that is seen in Table IV is spurious, due to neglect of dependence of labour market status on education. In other words, Table IV contains information about the model, but not by means of parameter estimates. It does not in itself identify any parameters. What we will do now is combine this information with that in the micro dataset to get estimates for θ that are more accurate than those in Table II.

We incorporate knowledge of the conditional employment probabilities by using a GMM estimator with the moments

$$\psi_{1j}(L, x, \theta)=\frac{L-\Phi(x'\theta)}{\Phi(x'\theta)(1-\Phi(x'\theta))}\phi(x'\theta)\cdot x_j \quad \text{for } j=1,\ldots,4, \tag{10}$$

$$\psi_{2i}(L, x, \theta)=1_{\{A\in[20+5i,25+5i)\}}\cdot[p_i-\Phi(x'\theta)] \quad \text{for } i=1, 2,\ldots,5. \tag{11}$$

ψ_1 is the score for the conditional likelihood that was maximized to get the estimates in Table II. The additional information is contained in the moment ψ_2. Since there are $4+5$ moments and only 4 parameters, we estimate the parameters in two rounds, using a consistent estimate of the optimal weight matrix in the second round. The results are in Table V. All standard deviations decrease considerably. The standard deviations on the coefficients θ_2 and θ_3, which capture the age-dependency of the employment probability decrease by a factor 7. It would require a dataset about 50 times as large (i.e. about 15,000 observations rather than 347) as large to get similar precision from the conditional maximum likelihood estimator for these coefficients.

If we estimate θ using the micro sample it is not informative to know the marginal distribution of age, because age is ancillary. However, if we impose the restrictions implied

TABLE V

Estimates using marginal employment probabilities

Regressor	Estimate	S.E.	T-value
Intercept	1·8574	0·2678	6·93
Education	−0·1085	0·0837	−1.30
Age-35	0·0030	0·0020	1·50
(Age-35)^2	−0·0028	0·0003	−9·33

by knowledge of the conditional employment probabilities for the various age categories, this is no longer the case. Age is not ancillary any more and knowledge about its marginal distribution is informative about θ. However, in Section 2.2 it was shown that knowledge of marginal probabilities like the q_i in Table IV is not informative if we know p_i already. Other information about the marginal distribution of x, like the mean $E(x)$, or correlation of elements of x, would have added information but knowledge of q_i does not.

From the preceding estimation we can conclude that it is potentially very useful to incorporate aggregate information of the type incorporated in ψ_2. There are two caveats. The first is that the marginal information might not be correct for the population from which the sample is drawn. This might be the result of inadvertent selection of the sample or of incorrectly measured aggregate statistics. A second caveat is that the model in equation (9) need not be correct. The way in which not only the standard deviations, but also the estimates change, suggests that this might be a problem. These two problems have very different implications. The consequence of the first is that combining the micro and aggregate data is not meaningful. They essentially refer to different populations. This might be a result of the process that determines the aggregate statistics from an underlying population, such as de-seasonalizing, imputing missing values and other procedures. The second problem inplies that the model is mis-specified. One would have to go back to the theoretical model underlying the specification and come up with a better version. In this case however the combination of micro data and aggregate information is productive, even if the final result is negative.

To examine these two problems we employ the specification tests discussed in Section 2.5. First we investigate whether the probabilities in Table IV correspond to the same population as the sample. Table VI gives the sample equivalents of the numbers that are given in Tables III and IV for the entire population. The question is whether the sample can be regarded as drawn randomly from the population that is represented in Tables III and IV. We will test this in two ways. First we look at the conditional employment probabilities $\hat{p}_i$. If the sample is drawn randomly from the population then Y_i should, conditional on N_i, have a binomial distribution with parameters N_i and p_i. The test statistic is

$$T_{c1}=\sum_{i=1}^{5}\frac{N_i(p_i-\hat{p}_i)^2}{(1-p_i)\cdot p_i}.$$

Asymptotically T_1 should have a $\chi^2(5)$ distribution. The value obtained was 0·94. We therefore do not reject the null hypothesis that the sample is drawn randomly from the population. The second test compares the $\hat{q}_i$ to the population probabilities q_i. The standard χ^2 test has the form

$$T_{c2}=\sum_{i=1}^{5}\frac{(N\cdot q_i-N_i)^2}{N\cdot q_i}.$$

TABLE VI

Sample counts and frequencies

Age category	Working (Y_i)	Total (N_i)	$\hat{p}_i = Y_i/N_i$	$\hat{q}_i = N_i/N$
25–29	84	93	0·903	0·268
30–34	78	85	0·918	0·245
35–39	55	59	0·932	0·170
40–44	56	61	0·918	0·176
45–49	42	49	0·857	0·141
Total	315	347	0·908	1·000

Asymptotically the test statistics should have a $\chi^2(4)$ distribution. The value obtained was 1·92. Again we cannot reject the null hypothesis of random sampling. The advantage of doing the test in two steps rather than doing one comprehensive χ^2 test on all ten cells (employed or not, for the five age categories) is that we can distinguish between two sources of non-randomness. Suppose people in particular age categories were over-sampled, but conditional on age the sampling was random. In that case the information in p_i would still be valid, and our estimator would therefore still be consistent. One would then expect only the second test to reject. If on the other hand the sampling was conditional on employment status, but independent of age, the second type of information on q_i would still be valid. In that case one would expect the first test to reject the null, but not the second.

We can also test the model specification. Testing it separately, using only the data from the random sample is now very insightful, as we cannot even reject the hypothesis that all slope coefficients are zero in the specification (9). We therefore test the specification of the conditional distribution using the aggregate information. We will still interpret these tests as testing mainly the specification of the conditional distribution rather than that of compatibility.

The Hausman test discussed in Section 2.3, based on the difference between the efficient and the conditional maximum likelihood estimates is equal to 10·6. Asymptotically it should have an $\chi^2(4)$ distribution. The 95% and 97.5% percentile of the $\chi^2(4)$ distribution are 9·5 and 11·1 respectively. We can therefore marginally reject the null hypothesis.

The method of moments estimators whose results are reported in Tables IV and V use more moments than parameters. The minimizing value of the objective function has asymptotically a χ^2 distribution with degrees of freedom equal to the number of over-identifying restrictions. The test statistic that resulted when we estimated θ using moments ψ_1 and ψ_1 is equal to 11·4. Asymptotically it should have a $\chi^2(5)$ distribution. The 95% and 97·5% percentiles for this distribution are 11·1 and 12·8 respectively. Again, we can marginally reject the null hypothesis that the model is correctly specified. Testing the model specification in a method of moments framework with over identifying restrictions need not be restricted to the use of ψ_1 and ψ_2. We can also add the restriction of zero correlation between ψ_1 and ψ_2. In Section 2.3 it was shown that this zero correlation between the scores of the conditional likelihood and the moments incorporating the aggregate information is a general characteristic of these methods. The appropriate moment to add for this test is of the form

$$\psi_{3ij}(L, x, \theta) = \psi_{1j}(L, x, \theta) \cdot \psi_{2i}(L, x, \theta)$$

for $i = 1, \ldots, 5$ and $j = 1, \ldots, 4$. Adding these moments will not increase the efficiency of the estimators, as they already achieve a semi-parametric efficiency bound, but it provides

an additional specification test. We applied this test using only the following five of the potential twenty restrictions:

$$\psi_{3i}(L, x, \theta) = \psi_{11}(L, x, \theta) \cdot \psi_{2i}(L, x, \theta)$$
$$= \frac{L - \Phi(x'\theta)}{\Phi(x'\theta)(1-\Phi(x'\theta))} \cdot \phi(x'\theta) \cdot 1_{\{A \in [20+5i, 25+5i)\}} \cdot [p_i - \Phi(x'\theta)]$$

for $i = 1, \ldots, 5$. The test statistic, the minimum value of the objective function, has asymptotically a $\chi^2(10)$ distribution. The value obtained was 73.1. We can therefore reject the specification of the conditional employment probability.

The results of this test should be interpreted with some caution. The construction of the test uses an estimate for the variance of the moments based on the outer product form. Such tests, in the context of information matrix tests, are known to have finite-sample distributions very different from the large-sample distributions. In particular, they tend to reject too often, even if the restrictions on the covariance matrix implied by the model are imposed. For a discussion of these issues see Chesher and Spady (1991).

Another difficulty with the last test is that rejection of the null hypothesis is more difficult to interpret than a rejection would have been with the GMM test using only ψ_2 as the over-identifying moments. Using only ψ_1 and ψ_2 as moments the test clearly has power against alternatives that are characterized by an age profile more general than the quadratic one allowed for in the specification of $f(y|x; \theta)$. This is the type of misspecification that we are interested in given the presence of direct information on the age profile, and the fact that both that GMM test and the Hausman test only marginally reject suggest that the model does not capture the age profile too badly. Additional evidence is the fact that when we included a cubic term in age, $(\text{Age}-35)^3$, the coefficient was not significant with a T-value of 0·83. With the GMM test using the moments ψ_1, ψ_2 and $\psi_{11} \cdot \psi_2$ it is less clear how to interpret the rejection, or identify alternatives against which this test has power.

The conclusion from the tests is therefore mixed. We cannot reject the hypothesis that the marginal information is correct in that it refers to the population from which the sample is drawn. We can however, using this information, reject the model specification, with some reservations. Given that the initial results, where we only used the information in the sample of 347, indicated that we could not reject the hypothesis that none of the regressors had any effect on the employment probabilities, this shows that this type of information can be quite valuable. This is even more surpising given that the extra information in itself does not identify any of the parameters. It is the combination of the two sources of information that makes for a powerful mix.

4. EFFICIENCY GAINS FOR PROBIT MODELS

To put the results from the previous section in perspective we calculate in this section the efficiency gains for a number of different probit models. In particular we try to answer two questions. First, are the considerable efficiency gains reported in the previous section due to the particular parameter values or to the particular distributions of that case? Second, what type of marginal moment information leads to such gains? The basic model we investigate has the probability of $y=1$ equal to $\Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)$, where $\Phi(\cdot)$ is the standard normal distribution function. The regressors x_1 and x_2 have, unknown to the researcher, a bivariate normal distribution with both means equal to zero, both variances equal to one and the correlation coefficient equal to ρ. If we set the values of the

parameters equal to $\theta_0=0$, $\theta_1=0{\cdot}5$, $\theta_2=0{\cdot}5$ and $\rho=0$, the normalized variances for the conditional likelihood estimators for θ_1 and θ_2 are both equal to 6·85. We then calculate the variance for the optimal GMM estimators under four different assumptions about additional information.

I. The researcher knows the expectation of y to be equal to h^*. The estimator uses in addition to the scores from the conditional likelihood function the moment

$$\psi_2(y, x; \theta) = h^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2).$$

This reduces the variance of both $\hat{\theta}_1$ and $\hat{\theta}_2$ by less than 0·01. This type of information is clearly not very useful. Note that in the logit model information on the marginal mean of y is not useful about the slope parameters at all. It turns out that in the probit model this is also approximately true.

II. The researcher knows the expectation of y conditional on $x_1<0$ and conditional on $x_1 \geqq 0$. The estimator uses in addition to the scores from the conditional likelihood function the moments:

$$\psi_2(y, x; \theta)_1 = 1_{x_1<0} \cdot [h_1^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)],$$
$$\psi_2(y, x; \theta)_2 = 1_{x_1 \geqq 0} \cdot [h_2^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)].$$

This reduces the variance of $\hat{\theta}_1$ considerably, from 6·85 to 1·85. It does not decrease the variance of $\hat{\theta}_2$ significantly. If the additional information is about the conditional distribution of y given x_1, then the gains are only in the precision with which one estimates the coefficient of x_1.

III. The researcher knows the expectation of y conditional on $x_1<-1{\cdot}282$, $-1{\cdot}282 \leqq x_1<-0{\cdot}43$, $-0{\cdot}43 \leqq x_1<0{\cdot}43$, $0{\cdot}43 \leqq x_1<1{\cdot}282$ and $1{\cdot}282 \leqq x_1$. The five additional moments are:

$$\psi_2(y, x; \theta)_1 = 1_{x_1<-1{\cdot}282} \cdot [h_1^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)],$$
$$\psi_2(y, x; \theta)_2 = 1_{-1{\cdot}282 \leqq x_1<-0{\cdot}43} \cdot [h_2^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)],$$
$$\psi_2(y, x; \theta)_3 = 1_{-0{\cdot}43 \leqq x_1<0{\cdot}43} \cdot [h_3^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)],$$
$$\psi_2(y, x; \theta)_4 = 1_{0{\cdot}43 \leqq x_1<1{\cdot}282} \cdot [h_4^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \Phi_2)],$$
$$\psi_2(y, x; \theta)_5 = 1_{1{\cdot}282 \leqq x_1} \cdot [h_5^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)].$$

This again reduces the variance $\hat{\theta}_1$ considerably, this time to 1·23. The variance of $\hat{\theta}_2$ again does not decrease. It is interesting to compare these efficiency gains to those under the previous case where the additional information was on just two moments, $E[y|x_1<0]$ and $E[y|x_1 \geqq 0]$. Here we know the conditional expectation of y for a much finer partitioning of the X_1 space, but that does not lead to a much larger efficiency gain than with the coarse partitioning.

IV. The researcher knows the conditional expectation of y conditional on $x_1<0$, $x_1 \geqq 0$, $x_2<0$ and $x_2 \geqq 0$. This leads to the following four moments in addition to the score from the conditional likelihood function.

$$\psi_2(y, x; \theta)_1 = 1_{x_1<0} \cdot [h_1^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)],$$
$$\psi_2(y, x; \theta)_2 = 1_{x_1 \geqq 0} \cdot [h_2^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)],$$
$$\psi_2(y, x; \theta)_3 = 1_{x_2<0} \cdot [h_3^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)],$$
$$\psi_2(y, x; \theta)_4 = 1_{x_2 \geqq 0} \cdot [h_4^* - \Phi(\theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2)].$$

TABLE VII

Changing θ_0

	I		II		III		IV	
θ_0	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)
−2·0	0·96	0·99	0·13	0·98	0·10	0·94	0·12	0·11
−1·0	1·00	1·00	0·23	0·99	0·15	1·00	0·22	0·21
0·0	1·00	1·00	0·27	1·00	0·18	1·00	0·26	0·26
1·0	0·99	0·99	0·24	1·00	0·16	1·00	0·22	0·21
2·0	1·00	1·00	0·16	0·97	0·11	0·96	0·12	0·12

TABLE VIII

Changing θ_1

	I		II		III		IV	
θ_0	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)
−0·0	1·00	1·00	0·19	1·00	0·14	1·00	0·19	0·07
0·1	1·00	1·00	0·19	1·00	0·13	1·00	0·19	0·08
0·5	1·00	1·00	0·27	1·00	0·17	1·00	0·28	0·26
1·0	1·00	1·00	0·36	1·00	0·21	0·99	0·38	0·51

This reduces both the variance of $\hat{\theta}_1$ and that of θ_2 to 1·78. Knowing $E[y|x_1<0]$, $E[y|x_1\geqq 0]$, $E[y|x_2<0]$, and $E[y|x_2\geqq 0]$ is almost equivalent to having a sample four times as large as the original sample. Note that these four expectations are all estimable from data sets that contain information about only two variables at a time. Such data sets do not identify any of the parameters by themselves, but they do make estimation with a random sample much more precise.

In the following tables we compare the variances of the efficient GMM estimators under these four different assumptions about additional information to the variances of the conditional maximum likelihood estimators for different values of the parameters. We report the ratios:

$$\text{eff}(\theta_1)=\{[\mathcal{I}+\Gamma_g'\Delta_g^{-1}\Gamma_g]^{-1}\}_{22}/\{\mathcal{I}^{-1}\}_{22}$$
$$\text{eff}(\theta_2)=\{[\mathcal{I}+\Gamma_g'\Delta_g^{-1}\Gamma_g]^{-1}\}_{33}/\{\mathcal{I}^{-1}\}_{33}$$

where $\{A\}_{ij}$ is the ij-th element of the matrix A. The expected values $\mathcal{I}$, Γ_g and Δ_g are estimated using 10,000 simulated observations, replacing the expectation by an average, evaluated at the true parameter values.

In Table VII we change the value of the intercept. For values of the intercept that make the marginal probability of $y=1$ close to zero or one the efficiency gains are considerably larger. This is true for all four types of extra information.

In Table VIII we change the value of θ_1. This does not seem to have a major impact on the efficiency gains for $\hat{\theta}_2$. If the extra information is about both the conditional distribution of y given x_1 and that of y given x_2, the smaller θ_1, the more valuable the information is for $\hat{\theta}$. In Table IX we change the value of θ_2. This has considerable effect on the value of information about the conditional distribution of y given x_1. The smaller θ_2, the more valuable such information is. The value for estimation of θ_2 is not much affected by the magnitude of θ_2.

In Table X we change both θ_1 and θ_2 at the same time. If both are equal to zero, the covariance matrix becomes singular, and the parameters converge faster than $\sqrt{N}$. The smallest value that we look at is therefore 0·01. In that case the efficiency gains are very

TABLE IX

Changing θ_2

	I		II		III		IV	
θ_2	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)
0·0	1·00	1·00	0·07	1·00	0·01	1·00	0·07	0·19
0·1	1·00	1·00	0·09	1·00	0·02	0·98	0·08	0·18
0·5	1·00	1·00	0·27	0·99	0·17	0·99	0·26	0·26
1·0	1·00	1·00	0·51	1·00	0·40	1·00	0·51	0·38

TABLE X

Changing θ_1, θ_2

	I		II		III		IV	
θ_1, θ_2	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)
0·01	1·00	1·00	0·00	1·00	0·00	0·08	0·00	0·00
0·10	1·00	1·00	0·01	1·00	0·01	0·96	0·01	0·01
0·50	1·00	1·00	0·27	1·00	0·17	0·99	0·26	0·26
1·00	1·00	1·00	0·61	1·00	0·49	0·99	0·62	0·61

TABLE XI

Changing ρ

	I		II		III		IV	
ρ	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)
−0·99	1·00	1·00	0·98	1·00	0·94	0·96	0·08	0·08
−0·90	1·00	1·00	0·82	1·00	0·81	0·98	0·20	0·20
−0·50	1·00	1·00	0·45	1·00	0·42	1·00	0·27	0·28
0·00	1·00	1·00	0·25	1·00	0·17	1·00	0·25	0·25
0·50	1·00	1·00	0·30	1·00	0·17	1·00	0·19	0·19
0·90	1·00	1·00	0·72	1·00	0·68	1·00	0·08	0·08
0·99	1·00	1·00	0·97	1·00	0·96	1·00	0·03	0·03

large. In general, the larger the values of the parameters, the less valuable the additional information becomes. This is true for both $\hat{\theta}_1$ and $\hat{\theta}_2$ and for all four types of information.

In Table XI we change the correlation between the regressors. If the additional information is only about the conditional distribution of y given one of the regressors (cases 2 and 3), then a correlation coefficient close to one in absolute value decreases the value of the additional information. This does not happen if the information is about both the conditional distribution of y given x_1 and about the conditional distribution of y given x_2. One should keep in mind when interpreting these results that with the correlation coefficient close to one in absolute value the ratio reported here is the ratio of two variances that are both much larger than the corresponding variances when the correlation coefficient is close to zero.

Table XII shows the effect of different distributions of the regressors. We look, in addition to the normal distribution used so far at the uniform distribution and discrete distributions with two and four points of support (all points of support have equal probability). In each case the mean and variance are fixed at zero and one respectively. The fewer the points of support, the more valuable any given amount of aggregate information. For example, knowing the conditional probabilities $P(y=1|x_1<0)$ and $P(y=1|x_1\geqq 0)$ is much more valuable if the distribution of x_1 is binary than if the distribution of x_1 has four points of support, or is uniform or normal. Intuitively the reason is that if x_1 is binary

TABLE XII

Changing the distribution of x_1

	I		II		III		IV	
Distribution	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)	eff (θ_1)	eff (θ_2)
Normal	1·00	1·00	0·27	1·00	0·18	1·00	0·26	0·26
$x_1 \in \{-1, 1\}$	1·00	1·00	0·16	1·00	0·16	1·00	0·22	0·26
$x_1 \in \left\{-\frac{3}{\sqrt{5}}, -\frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}, \frac{3}{\sqrt{5}}\right\}$	1·00	1·00	0·22	1·00	0·16	1·00	0·22	0·26
Uniform $[-\sqrt{3}, \sqrt{3}]$	1·00	1·00	0·21	1·00	0·15	1·00	0·21	0·26

the aggregate information amounts to knowing the entire conditional distribution $P(y=1|x_1)$ rather than just two probabilities as in the continuous case.

5. CONCLUSION

In this paper we develop a method for using aggregate data in the estimation of micro-econometric models. The estimator that efficiently incorporates this type of information is a generalized method of moments estimator. The efficiency gains that result are considerable. Since this type of information is both freely and widely available there seems to be a strong case for using this type of information, in the framework outlined in this paper.

The typical example where this approach may be useful is one where the investigator has a small survey with all relevant variables, in combination with aggregate data on some of the variables. One qualification is that one should be sure that the aggregate data are not based on exactly the same survey data. In that case one could erroneously conclude that bringing in the aggregate data leads to substantial gains while any gains would be spurious.

Another example where this approach might be particularly useful is that where one wishes to use micro-level data to predict or track patterns of aggregate data. See for instance Blundell, Pashardes and Weber (1989) and Heckman and Walker (1989). Using the aggregate data available to link the model more closely to the past pattern of these macro data might lead to a considerable improvement in predictions.

There are of course many questions unanswered. We have only looked at a very limited set of parametric models. What are the efficiency gains for models other than the ones analysed in this paper? Can these procedures be adapted to deal with general semi-parametric models where the conditional distribution of y given x is not parametrically specified? What if there are measurement errors destroying the direct link between the micro and the macro variables?

All of these questions deserve further attention, but the conclusion based on the results in this paper is that using aggregate data to improve on estimates based on micro data alone is both a feasible and a fruitful approach.

APPENDIX

First we give a useful lemma from Imbens (1992):

Lemma 1. *Suppose the maximum likelihood estimator of a vector* θ *with* $\theta = (\theta_1' \theta_2' \theta_3')'$ *can be characterized*

by

$$\sum_{n=1}^{N} h_1(\hat{\theta}_1, \hat{\theta}_2, x_n) = 0$$

$$\sum_{n=1}^{N} h_2(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, x_n) = 0$$

with $\dim(h_1) = \dim(\theta_1) + \dim(\theta_2)$ *and* $\dim(h_2) = \dim(\theta_3)$. *Then, the optimal constrained method of moments estimator for* θ_1 *given* $\theta_2 = 0$ *based on minimization of*

$$\left[\frac{1}{N}\sum_{n=1}^{N} h_1(\theta_1, 0, x_n)\right]' \cdot A_N \cdot \left[\frac{1}{N}\sum_{n=1}^{N} h_1(\theta_1, 0, x_n)\right]$$

where $A_N \xrightarrow{a.s.} [Eh_1 \cdot h_1']^{-1}$, *has the same asymptotic covariance matrix as the constrained maximum likelihood estimator. In other words, it achieves the Cramér–Rao lower bound.*

Proof. See Imbens (1992). ‖

Proof of Theorem 1. The proof is similar to that of Theorem 2 in Imbens and Lancaster (1991) and based on work on semi-parametric efficiency bounds by Begun, Hall, Huang and Wellner (1983) Chamberlain (1987) and Newey (1990). We take a sequence of parameterizations of the marginal density of x, $r(x)$. We then show that the sequence of Cramér–Rao bounds associated with this sequence of parameterizations converges to V, the asymptotic covariance matrix of $\hat{\theta}$. Therefore any estimator that has an asymptotic covariance matrix that is less than V, must be more efficient than a maximum likelihood estimator. This implies that it cannot be a regular estimator.

The sequence of parameterizations is indexed by ε. For any $\varepsilon > 0$ partition X into L_ε subsets X_l such that $X_l \cap X_m = \emptyset$ for all $l \neq m$, and for all x^0 and $x^1 \in X_l \| x^0 - x^1 \| < \varepsilon$. Let $\phi_{lx} = 1$ if $x \in X_l$ and 0 otherwise. Define

$$r_\varepsilon(x) = r(x) \Big/ \left[\sum_l \phi_{lx} \int_{X_l} r(z)dz\right].$$

The parameterization we will use is

$$f(y|x; \theta) \cdot \sum_l \delta_l \phi_{lx} r_\varepsilon(x)$$

with parameters δ and θ. The true value of δ_l is $\int_{X_l} r(z)dz$. The maximum likelihood estimator for θ can be characterized by

$$0 = \frac{1}{N}\sum_{n=1}^{N} \frac{\ln f}{\partial\theta}(y_n|x_n; \hat{\theta}) \tag{12}$$

and that for δ_l by

$$0 = \frac{1}{N}\sum_{n=1}^{N} \phi_{lx} - \hat{\delta}_l.$$

Given the maximum likelihood estimators $\hat{\theta}$ and $\hat{\delta}$, the maximum likelihood estimator for $h^* = Eh(y, x)$ can be written as

$$\hat{h} = \int_z \int_v h(v, z) f(v|z; \hat{\theta}) \sum_l \hat{\delta}_l \phi_{lz} r_\varepsilon(z) dv dz$$

or

$$0 = \frac{1}{N}\sum_{n=1}^{N} \left[\hat{h} - \sum_l \phi_{lx_n} \int_{X_l} \int_Y h(v, z) f(v|z; \hat{\theta}) r_\varepsilon(z) dv dz\right]. \tag{13}$$

The fact that we can characterize $\hat{h}$ and $\hat{\theta}$ as solutions to moment equations (12) and (13) that do not involve δ implies that we can invoke the above lemma. This lemma implies that the GMM estimator for θ^* given h^* based on the optimally weighted combination of the two moments.

$$\psi_{\varepsilon 1}(y, x; \theta, h) = \frac{\partial \ln f}{\partial\theta}(y|x; \theta)$$

and

$$\psi_{\varepsilon 2}(y, x; \theta, h) = h - \sum_l \phi_{lx} \int_{X_l} \int_Y h(v, z) f(v|z; \theta) r_\varepsilon(z) dv dz$$

is as efficient as the constrained maximum likelihood estimator. Therefore the sequence of Cramér–Rao bounds is identical to the sequence of asymptotic covariance matrices for the GMM estimator of θ^*. This sequence is equal to

$$V_\varepsilon = [\mathcal{I} + \Gamma_\varepsilon' \Delta_\varepsilon^{-1} \Gamma_\varepsilon]^{-1}$$

with

$$\Delta_\varepsilon = E\psi_{\varepsilon 2}(y, x; \theta^*, h^*) \cdot \psi_{\varepsilon 2}(y, x; \theta^*, h^*)'$$

and

$$\Gamma_{\varepsilon 2} = E\frac{\partial \psi_\varepsilon}{\partial \theta'}(y, x; \theta^*, h^*).$$

Sufficient for convergence of V_ε to V is that Γ_ε and Δ_ε converge to Γ and Δ. For this to hold it is sufficient that $\psi_{\varepsilon 2}(y, x | \theta, h)$ and $\partial \psi_{\varepsilon 2}/\partial \theta$ $(y, x | \theta, h)$ converge uniformly to

$$h - \int_Y h(v, x) f(v|x; \theta) dv$$

and

$$-\int_Y h(v, x) \frac{\partial f}{\partial \theta}(v|x; \theta) dv$$

respectively. This follows from the assumption that f is twice differentiable, the compactness of X, Θ, and the boundedness of $h(y, x)$. ‖

Acknowledgements. An earlier version of this paper circulated under the title "Uses of Marginal Information in Econometrics", Brown University, November 1989. We are grateful for comments by Gary Chamberlain, Ken Small, Josh Angrist, participants in seminars at the Center for Economic Research (Tilburg University), Harvard University, Hebrew University of Jerusalem, Tel Aviv University, Yale University and Princeton University, and two anonymous referees. We also acknowledge financial support from the NSF under grant SES 9122477.

REFERENCES

ANGRIST, J. D. and KRUEGER, A. B. (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*, **87**, 328–336.

ARELLANO, M. and MEGHIR, C. (1992), "Female Labour Supply and on the Job Search: an empirical model estimated using Complementary Data sets", *Review of Economic Studies*, **59**, 537–559.

BEGUN, J. M., HALL, W. J., HUANG, W-M. and WELLNER, J. A. (1983), "Information and Asymptotic Efficiency in Parametric–Nonparametric models", *Annals of Statistics*, **11**, 432–452.

BLUNDELL, R., PASHARDES, P. and WEBER, G. (1989), "What do we learn about consumer demand patterns from micro-data?" (Discussion Paper 89-18, University College London).

CENTRAAL BUREAU VOOR DE STATISTIEK, (1978, 1980) *Handboek voor de Statistiek* (Voorburg: The Netherlands).

CHAMBERLAIN, G. (1984), "Panel Data", in Z. Griliches and M. Intrilligator (eds.), *Handbook of Econometrics, Volume* 2 (Amsterdam: North-Holland).

CHAMBERLAIN, G. (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics*, **34**, 305–334.

CHESHER, A. (1984), "Testing for Neglected Heterogeneity", *Econometrica*, **52**, 865–872.

CHESHER, A. and SPADY, R. (1991), "Asymptotic Expansions of the Information Matrix Test Statistic", *Econometrica*, **59**, 787–815.

GOURIEROUX, C. and MONFORT, A. (1981), "On the Problem of Missing Data in Linear Models", *Review of Economic Studies*, **48**, 597–596.

GRILICHES, Z., HALL, B. and HAUSMAN, J. (1978), "Missing Data and Self-Selection in Large Panels", *Annales de l'INSEE*, No. 30–31, 137–176.

HANSEN, L. P. (1982), "Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, **50**, 1029–1054.
HAUSMAN, J. A. (1978), "Specification Tests in Econometrics", *Econometrica*, **46**, 1251–72.
HECKMAN, J. and WALKER, J. (1989), "Forecasting Aggregate Period-Specific Birth Rates: The Time Series Properties of a Microdynamic Neoclassical Model of Fertility", *Journal of the American Statistical Association*, **84**, 958–965.
IMBENS, G. W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-based Sampling", *Econometrica*, **60**, 1187–1214.
IMBENS, G. and LANCASTER T. (1991), "Efficient Estimation and Stratified Sampling" (Discussion Paper 1545, Harvard Institute of Economic Research).
JORGENSON, D., LAU, L. and STOKER, T. (1982), "The Transcendental Logarithmic Model of Aggregate Consumer Behavior", in R. Basmann and G. Rhodes (eds.), *Advances in Econometrics* (Greenwich, Conn: JAI Press).
LANCASTER, T. and IMBENS, G. (1991), "Choice-Based Sampling: Inference and Optimality" (Department of Economics Working Paper Brown University).
LUSARDI, A. (1992), "Euler Equations in Micro Data: Merging Data from Two Samples" (Department of Economics, Dartmouth College).
MANSKI, C. F. and LERMAN, S. R. (1977), "The Estimation of Choice Probabilities from Choice-Based Samples", *Econometrica*, **45**, 1977–1988.
NEWEY, W. (1985), "Maximum Likelihood Specification Testing and Conditional Moment Tests", *Econometrica*, **53**, 1047–1069.
NEWEY, W. (1990), "Semiparametric Efficiency Bounds", *Journal of Applied Econometrics*, **5**, 99–136.
ROBINS, J., HSIEH, F. and NEWEY, W. (1991), "Semiparametric Efficient Estimation of a Conditional Density with Missing or Mismeasured Covariates" (mimeo, School of Public Health, Harvard University).
STOKER, T. (1985), "Aggregation, Structural Change, and Cross-Section Estimation", *Journal of the American Statistical Association*, **80**, 720–729.
THEIL, H. and GOLDBERGER, A. (1961), "On Pure and Mixed Statistical Estimation in Economics", *International Economic Review*, **2**, 65–78.
U.S. BUREAU OF THE CENSUS, (1990) *Statistical Abstract of the U.S.* (Washington, D.C.: U.S. Government Printing Office).