Using Weights to Adjust for Sample Selection When Auxiliary Information Is Available

Aviv Nevo
University of California, Berkeley and the National Bureau of Economic Research (nevo@econ.berkeley.edu)

This article analyzes generalized method of moments estimation when the sample is not a random draw from the population of interest. Failure to take this selection into account can potentially lead to inconsistent and biased estimates of the parameters of interest. This article develops a method that uses auxiliary information to inflate the observed data so it will be representative of the population of interest; thus standard estimators applied to the weighted data will yield consistent and unbiased estimates.

Suppose, for example, that one wants to study a panel of individuals in which there is thought to be an unobserved individual-specific effect correlated with the independent variables. Common solutions include “within” and first-difference estimators. However, implementing these types of estimators requires at least two observations for each individual. If the panel also suffers from nonrandom attrition, then the subset of individuals observed for more than one period is no longer a random draw from the population of interest. Assume that additional samples are at our disposal that are representative of each of the cross-sections, but also suffer from attrition and thus do not allow consistent estimation of the parameters of interest. An example of such a dataset is the Dutch Transportation Panel (DTP) considered by Ridder (1992) and used herein.

The procedure proposed in this article suggests one possible way to use these additional samples, the refreshment samples. The refreshment samples are used to attach a weight to each observation in the balanced subpanel such that moments in the weighted sample are set equal to corresponding moments in the refreshment samples. Estimating the parameters of interest using standard panel estimators and the weighted balanced panel is proposed.

The proposed method exploits additional data to adjust for the selection bias by using it to estimate the selection probability, the propensity score. Here the propensity score is used to inflate the observations. This is not the only way to exploit the additional information. Alternatively, it can be used to construct a control function (Heckman 1979; Ridder 1992, for the data structure discussed later), to match observations (Ahn and Powell 1993), or to impute the missing observations (Hirano, Imbens, Ridder, and Rubin 2001 for the data used herein).

The advantage of the proposed weighting method over the alternative methods of using the aforementioned additional information is its simplicity. Once the weights have been computed, the researcher can conduct the same analysis that he or she would perform if the data were randomly drawn from the population of interest, using the weighted data. Thus any estimator can be computed with little added complexity due to the selection. The weights can be computed either by setting up the problem as a standard generalized method of moments (GMM) problem, or by solving a linear programming problem. The latter approach connects the proposed method to information-based alternatives to GMM.

Although the refreshment samples found in the DTP are not typical of economic data, the method proposed here is general. For example, census and annual surveys of manufacturers data are sources of random draws from the population of interest that can be matched with smaller datasets that suffer from attrition but have information on economic agents over time. Furthermore, the use of refreshment samples to adjust for selection bias motivates collection of such samples, a procedure not often done.

1.1 Previous Literature

A survey of the literature on sample selection is beyond the scope of this article. (For this, see, e.g., Heckman 1979, 1987, 1990; Heckman and Robb 1986; Little and Rubin 1987; Newey, Powell, and Walker 1990; Manski 1994; Angrist 1995; Heckman, Ichimura, Smith, and Todd 1998; and references therein.) Nonetheless the aim is to relate the method proposed here to previous work. The use of weights to correct for sample selection is not new to this article. For example, many data-collecting agencies (including the U.S. Census Bureau)
provide sampling weights to correct for the sampling procedure. Manski and Lerman (1977) proposed a method for computing weights for choice-based samples (see also Cosslett 1981; Wooldridge 1999). In both of these examples, the selection probability is known; in contrast, in this article the probability of selection and the implied weights are estimated jointly with the parameters of interest. Methods that estimate the weights jointly with the parameters of interest also exist (e.g., Cassel, Sarndal, and Wretman 1979; Koul, Susarla, and Van Ryzin 1981; Heckman 1987).

The method proposed here differs from these methods in two ways. First, the weights are allowed to be a function of variables that are not fully observed in the main dataset, but for which some information is known through the additional moments. The weights are also allowed to be a function of the dependent variable. In other words, this is a nonignorable selection mechanism (Little and Rubin 1987). Second, the weights are computed by exploiting auxiliary information. This connects the method proposed here to weighting techniques for contingency tables with known marginal distributions (Oh and Scheren 1983, sec. 3; Little and Wu 1991). These methods are extended by examining a logistic selection equation and by looking at regression functions rather than contingency tables.

The estimator used here can be related to information-theoretic alternatives to GMM (Back and Brown 1990; Qin and Lawless 1994; Imbens 1997; Kitamura and Stutzer 1981; Wooldridge 1999). In both of these examples, the selection probability is known through the additional information of some factors present only in some periods. From this, it is possible to consistently estimate the density of the data by the empirical distribution (i.e., giving each observation a density of 1/N). The information-theoretic alternatives use overidentifying moment conditions to improve on the estimates of the data distribution, by estimating the distribution jointly with the parameters of interest. The focus of this literature has been on improving efficiency. A related work (Nevo 2002) has shown that the estimator proposed here can be related to these methods. The differences between the sampled population, from which the sample is drawn, and the population of interest are made explicit. Therefore, bias in the estimation of the distribution can be dealt with. It turns out that when the sampled population equals the target population, the estimator proposed here is equivalent to the one proposed by Imbens et al. (1998).

This article is organized as follows. Section 2 presents the proposed estimator, and Section 3 applies this estimator to the data used by Ridder (1992). Section 4 concludes.

2. THE MODEL AND PROPOSED ESTIMATOR

2.1 The Setup

Suppose that N independent realizations \( \{z_1, z_2, \ldots, z_N\} \) of a (multivariate) random variable Z with support \( \chi \), a compact subset of \( \mathbb{R}^k \), are observed. In the population, Z has a pdf, \( f(Z) \). Let \( \theta_i^* \) denote the true value of the parameters of interest and let \( \theta_i \in \Theta_i \), where \( \Theta_i \) is a compact subset of \( \mathbb{R}^k \).

**Assumption A0.** \( \theta_i^* \) uniquely set \( E[\psi(Z, \theta_i^*)] = 0 \). Furthermore, the moment function, \( \psi: \chi^\ast \Theta_i \to \mathbb{R}^k \), is twice continuously differentiable with respect to \( \theta \) and measurable in \( Z \), and \( E[\psi(Z, \theta_i^*)] \psi(Z, \theta_i^*)' \) and \( E[\partial \psi(Z, \theta_i^*)/\partial \theta_i] \) are of full rank.

A way of estimating \( \theta_i^* \) would be to follow the analogy principle by choosing the estimate

\[
\hat{\theta}_i \text{ s.t. } \frac{1}{N} \sum_{i=1}^{N} \psi(z_i, \hat{\theta}_i) = 0.
\]

Implicitly this assumes that the empirical distribution of the data is a consistent estimate of \( f(Z) \). However, if the observed sample is not a random draw from the population of interest, then the foregoing estimate is biased and inconsistent.

Let \( D_i = 1 \) if and only if \( z_i \) is fully observed. In a cross-sectional context, \( D = 0 \) could imply, for example, that the covariates are observed but the outcome variable is not. In a panel example, \( D = 1 \) for the balanced subsample, whereas \( D = 0 \) for individuals present only in some periods. From Bayes’s rule, the distribution of the observed \( z_i \) is

\[
f(Z|D = 1) = f(Z) \cdot P(D = 1|Z)/P(D = 1).
\]

**Assumption A1.** \( P(D = 1|Z) \) is bounded away from 0.

This assumption requires that any point in the support have a (strictly) positive probability of being observed. This is not a trivial requirement. Consider, for example, a truncation problem, a type 1 Tobit model. A unit will be observed if and only if the dependent variable, \( y \), exceeds a certain value. Because the conditioning vector, \( Z \), includes the dependent variable, the selection probability, \( P(D = 1|Z) \), will equal either 0 or 1, depending on the value of \( y \). The proposed method will not be operational in such a case. If the correlation between the selection rule and the variable of interest is less than 1, then this assumption will be satisfied (e.g., the model considered in Heckman 1974).

Assumption A1 allows one to recover the pdf, \( f(Z) \), given knowledge of \( f(Z|D = 1) \) and \( P(D = 1|Z) \) by

\[
f(Z) = f(Z|D = 1) \cdot P(D = 1)/P(D = 1|Z). \tag{1}
\]

Therefore,

\[
E[\psi(Z, \theta_i^*)] = \int \psi(z, \theta_i^*) f(z) dz = \int \psi(z, \theta_i^*) f(z|D = 1) \cdot P(D = 1)/P(D = 1|z) dz = 0,
\]

and the correct analog estimator becomes

\[
\hat{\theta}_i \text{ s.t. } \frac{1}{N} \sum_{i=1}^{N} \frac{1}{P(D = 1|z_i)} \psi(z_i, \hat{\theta}_i) = 0. \tag{2}
\]

The empirical distribution of the selected sample can be used to consistently estimate \( f(Z|D = 1) \). Therefore, if \( P(D = 1|Z) \) is known, then (2) can be used to consistently estimate \( \theta_i \). This is the fundamental idea behind using sampling weights to correct for nonrandom sampling in surveys and estimation in choice-based sampling (Manski and Lerman 1977; Cosslett 1981; Wooldridge 1999).

In general, the selection probability is unknown and must be estimated. To estimate the selection probability, \( P(D = 1|Z) \),
2.2 Identification

exact knowledge of the expectation, \( h^* \), is assumed in the population of an \( R \)-dimensional function of \( Z \), denoted by \( h(Z) \). Formally, \( h^* = E[h(Z)] = \int h(z)f(z)dz \). Examples include \( h(Z) = Y \), where the researcher knows the mean of the dependent variable, or \( h(Z) = Y - X \), where the researcher knows the (noncentered) covariance between the dependent variable and some of the independent variables. In the context of panel data, the moments \( h^* \) can come from the moments of the marginal (cross-sectional) distributional of the unbalanced panel. The application that follows demonstrates this.

Denote \( h(Z) = h(Z) - h^* \). Note that \( E[h(Z)] = 0 \). For now, assume that these moments are known and defer a discussion of where they come from and the possibility that they are known with error.

Assume that \( P(D = 1|Z) = P(D = 1|Z, \theta_2) \), where \( \theta_2 \in \Theta_2 \) and \( \Theta_2 \) is a compact subset of \( \Re^k \). Define

\[
\hat{\psi}(Z, \theta) = \begin{pmatrix}
\int \psi(Z, \theta_1)P(D = 1|Z, \theta_1)h(Z) & \int h(Z)P(D = 1|Z, \theta_2)
\end{pmatrix},
\]

(3)

where \( \theta = (\theta_1, \theta_2) \). Let \( \theta^* = (\theta^*_1, \theta^*_2) \) be the true value of the parameters. Note that by construction,

\[
E[\hat{\psi}(Z, \theta^*)|D = 1] = \int \hat{\psi}(z, \theta)f(z|D = 1)dz = 0.
\]

2.2 Identification

This section examines under what conditions the additional moments are sufficient to identify the selection probability. To see that the identification is not trivial, consider the following example. Let \( Z_i = (Z_{i1}, Z_{i2}) \) be a bivariate binary random variable, where \( Z_{i1} \) measures a characteristic (outcome) of individual \( i \) in time \( t (t = 1, 2) \). Also, let \( D_i = 1 \) if \( i \) is observed in both periods. If the attrition is nonrandom, that is, \( f(Z_{i1}, Z_{i2}) \neq f(Z_{i1}, Z_{i2}|D_i = 1) \), then analysis based on the individuals observed in both periods will yield biased and inconsistent estimates. Equation (1) corrects this bias by using \( P(D_i|Z_i) \) to weight the observations. Under what conditions is this probability identified?

In a large sample, the probabilities \( P(Z_{i1} = z_1, Z_{i2} = z_2|D_i = 1), \) \( z_1, z_2 \in \{0, 1\} \) can be estimated from the subsample of individuals observed in both periods. Assuming that the original sample is a random sample from the population, the probability \( P(Z_{i1}) \) can be identified. Suppose that additional information on \( P(Z_{i2}) \) is available, for example, by taking a random draw from the population at \( t = 2 \). This additional information is not sufficient to identify \( P(D_i|Z_i) \) in general. Without information on either the joint probability \( P(Z_{i1}, Z_{i2}) \) or additional restrictions, the selection probability is not identified.

This example demonstrates that even if the additional information comes in the form of the complete marginal distribution, identification is not trivial. If the additional information is in the form of marginal moments, this is even more true. Hirano, Imbens, Ridder, and Rubin (2001) proved that by assuming a particular functional form for the selection probability, the probability \( P(D_i|Z_i) \) is identified (Hirano et al. 2001, thm. 2). Because it is assumed that the additional information comes in the form of moments, slightly different conditions are required for identification.

Assumption A2. The matrix \( E[h(Z) \cdot h(Z)|D = 1] \) is of full rank.

Assumption A3. \( P(D = 1|Z) = g(h(Z)\theta_2) \), where \( g: \Re \to \Re \) is a known, differentiable, strictly increasing function such that \( \lim_{a \to \infty} g(a) = 0 \) and \( \lim_{a \to -\infty} g(a) = 1 \).

Most of the standard probability models satisfy assumption A3—in particular, the logistic model of selection used later, that is \( g(a) = \exp(a)/(1 + \exp(a)) \).

Proposition 1. Under assumptions A0–A3 and assuming that the functions \( h(\cdot) \) can be constructed, the parameters \( \theta = (\theta_1, \theta_2) \) are identified using a sample \( \{z_1, z_2, \ldots, z_N\} \), such that \( z_i \) for all \( i \) are iid with the empirical distribution converging to \( f(z|D = 1) \).

Proof. Assumptions A2 and A3 promise that the equations \( E[h(Z)\cdot g(h(Z)\theta_2)|D = 1] = 0 \) have a unique solution for \( \theta_2 \) such that \( \theta_2 = \theta_2^* \). Therefore, under assumptions A0 and A1, the system of equations \( E[\psi(Z, \theta_1)/g(h(Z)\theta_2)|D = 1] = 0 \) has a unique solution for \( \theta_1 \) such that \( \theta_1 = \theta_1^* \). This is standard GMM theory (Hansen 1982; Newey and McFadden 1994) prove the proposition.

A necessary condition for identification is that the number of (linearly independent) additional moments is at least as large as the number of parameters governing the selection, that is, the dimension of \( \theta_2 \). Assumption A3 requires more than this. The selection probability depends on the functions \( h(Z) \), for which the expectation in the population is known.

2.3 Estimation

A three-step procedure is proposed for estimating the parameters of interest. In the first step, the probability of selection is modeled as

\[
P(D = 1|Z, \theta_2) = \exp(g^*(Z, \theta_2))/(1 + \exp(g^*(Z, \theta_2))),
\]

(5)

where \( g^*(Z, \theta_2) \) is an unknown function. At this point no restrictions have been imposed, because if \( g^*(Z, \theta_2) \) is defined appropriately, then (5) can fit any selection model. The unknown function \( g^*(Z, \theta_2) \) is approximated by a polynomial, \( h(Z)\theta_2 \), with unknown coefficients, \( \theta_2 \). The model of selection is driven largely by data availability, the presence of the functions \( h(\cdot) \). Assuming that a rich set of moments is available for creating \( h(\cdot) \), the model can be derived from economic modeling. In this case the estimates of \( \theta_2 \) might be of independent interest.

The conditioning vector in (5) may also include the dependent variable in the estimation equation. Thus this selection model is nonignorable. The selection probability can be written as a function of observed variables as well as an individual-specific unobserved effect. Because the conditioning vector includes the dependent variable in the main estimation equation, this type of selection model is covered by the setup in some cases, depending on the exact assumptions governing the setup in some cases, depending on the exact assumptions governing the distribution of the individual-specific effects.
The second step is to estimate weights that are proportional to $1/P(D = 1|Z, \theta)$. Formally, the weights are computed by solving

$$
\sum_{i=1}^{N} w_i(z_i, \theta_2) h(z_i) = 0, \\
\sum_{i=1}^{N} w_i(z_i, \theta_2) = 1,
$$

(6)

where $w_i(z_i, \theta_2) = 1/P(D = 1|z_i, \theta_2)$.

By solving (6), the weighted sample counterpart of $E[h(Z)]$ is set to 0. Thus the additional moments, $h^*$, are exploited to estimate the parameters of the selection process. An alternative interpretation of (6) comes from an information-based criterion. This interpretation relates the proposed method to information-theoretic alternatives to GMM (see the references listed in Sec. 1.1 and Little and Wu 1991). Details have been provided in earlier work (see Nevo 2002) for details.

Using $g^*(Z, \theta_2) = h(Z)\theta_2$ and substituting (5) into (6) yields the system of equations

$$
\sum_{i=1}^{N} w_i(z_i, \theta_2) \cdot h(z_i) = \sum_{i=1}^{N} \frac{\tau}{N \cdot P(D = 1|z_i, \theta_2)} \cdot h(z_i) = 0, \\
\sum_{i=1}^{N} w_i(z_i, \theta_2) = \sum_{i=1}^{N} \frac{\tau}{N} \left( 1 + \frac{1}{\exp(h(z_i)/\theta_2)} \right) \cdot h(z_i) = 1,
$$

(7)

where $\tau = P(D = 1)$. For some selection models, the last equation in the system defined by (7) will just be a normalization. Therefore, it can be ignored in the solution and imposed later by dividing all the weights by their sum. In such cases the parameter $\tau$ will not be identified separately from a scale parameter. The logistic selection probability proposed herein does not have this property. Thus this last equation will be more than just a normalization—it will actually change the relative weights. Consider, for example, a linear probability model, $P(D = 1|Z, \theta_2) = \alpha + Z^\prime \beta$. The weights in such a case will be proportional to $[\alpha(1 + \alpha^{-1}Z^\prime \beta)]^{-1}$. Therefore, normalizing the weights to sum up to 1 will influence only the estimate of the constant, not the relative weights. But if the probability of selection is logistic [i.e., $P(D = 1|Z, \theta_2) = \exp(h(Z)/\theta_2)/(1 + \exp(h(Z)/\theta_2))]$ then, the weights will be proportional to $1 + \exp(-(\alpha + Z^\prime \beta))$. Now a normalization will not be fully absorbed in the constant term and will influence both the relative weights and their absolute value.

In the final step, the weights that solve (6) are used to obtain analog estimates of the parameters of interest, $\theta_1^\star$. Formally, the estimate is given by

$$
\hat{\theta}_1^w \text{ s.t. } \sum_{i=1}^{W} w_i \psi(z_i, \hat{\theta}_1^w) = 0,
$$

(8)

and the weights, $w_i$, solve (6). The asymptotic properties of this estimator are given by the following proposition.

**Proposition 2.** Suppose that $z_i$ ($i = 1, 2, \ldots$) are iid with the empirical distribution converging to $f(z|D = 1)$, and (a) assumptions A0–A3 are satisfied; (b) $\Theta_1 \times \Theta_2$ is compact; (c) the moment functions, $\psi(z, \theta)$, defined by (3), are twice continuously differentiable in $\theta$; and (d) $E[\psi(Z, \theta)\psi(Z, \theta)] < \infty$. Then $\hat{\theta}_1^w \rightarrow \theta^*$ and

$$
\sqrt{N}(\hat{\theta}_1^w - \theta^*) \rightarrow N(0, E_\psi(\hat{\psi}_\theta^{-1})^{-1} E_\psi(\hat{\psi}_{\theta}\hat{\psi}_\theta^{-1}^{-1} E_\psi(\hat{\psi}_{\theta}\hat{\psi}_\theta^{-1})^{-1} E_\psi(\hat{\psi}_\theta^{-1}^{-1} E_\psi(\hat{\psi}_{\theta}\hat{\psi}_\theta^{-1}))),
$$

where

$$
\hat{\psi}(Z, \theta) = \psi(Z, \theta_1)w(Z, \theta_2), \\
\tilde{h}(y, x) = \frac{h(Z)w(Z, \theta_2)}{w(Z, \theta_2) - 1},
$$

and $E_\psi[\cdot]$ denotes expectations taken with respect to $f(z|D = 1)$.

**Proof.** The expectation of the stacked moment conditions is set to 0 at the true parameter value; see (4). The assumptions of the proposition and standard GMM theory provide the result (Newey 1984; Pagan 1986; Newey and McFadden 1994).

The proof of the proposition stacks the moments as in (3) and considers them as just-identified estimation equations. The actual estimation can be obtained by solving these equations in one step (thus combining the second and third steps in the foregoing discussion.) However, as a computational issue, it is simpler to obtain the solution by first solving (6), then plugging the solution into (8), as proposed in the foregoing algorithm. Given the assumptions and that the set of moments is just identified solving all the moments jointly or in two steps yields the same unique solution.

Alternatively, the standard errors can be estimated by bootstrapping. This works as follows. First, a bootstrap sample is generated by sampling from the original sample. Next, the value of the parameter that sets the moments in (3) to 0 is determined. These two steps are repeated to obtain the bootstrap distribution of the parameters.

The proposition assumes that the value of the population moments is known exactly. This seems to be an adequate assumption if the second dataset is much larger than the primary dataset. However, in many interesting problems this will not be the case. Section 2.5 extends the results to take into account sampling error in the additional moments.

### 2.4 Examples of Data Structures

The proposed procedure assumes knowledge of population moments. A leading example of a source for these moments is the DTP used by Ridder (1992) and to which the proposed estimator is applied in Section 3. The unique design of this dataset called for draws of refreshment samples to deal with the heavy (seemingly nonrandom) attrition of the original panel. These refreshment samples are not characteristic of economic data. But, as the rest of this section demonstrates, under reasonable assumptions familiar economic datasets can fit into the proposed framework.

Assume a combination of census data and smaller datasets (as in Imbens and Lancaster 1994; Hellerstein and Imbens 1999). The larger dataset can be viewed as a random draw from the population, which provides estimates of
population moments for certain variables. The smaller dataset is not a random draw from the population of interest, however, it is much richer. For example, Gottschalk and Moffitt (1992) documented the differences between the National Longitudinal Survey (NLS), a small and rich dataset, that suffers from attrition, and the Current Population Survey (CPS), which is not as rich, but is much larger and representative of the population. Hellerstein and Imbens (1999) exploited this to correct for ability in the NLS. The method proposed in this article, which is similar to their approach, suggests using the additional data available from the CPS to treat the attrition in the NLS. The moments needed for the second step of the algorithm can be obtained from the CPS. The computed weights are then attached to the NLS, and the analysis is performed using the weighted dataset.

The data structure perhaps best suited for the proposed method is a panel structure. Consider the setup described in Section 1. The required moments, \( h^* \), can be obtained from census data, which does not have a time dimension to it, as in the previous example. Alternatively, one might be willing to assume that each cross-section is a random draw from the marginal distribution of the population of interest, yet the balanced subsample is not a random draw from the joint distribution. As long as the probability of selection is a separable function of cross-sectional variables, one can use the cross-sections to construct \( h(Z) \). The parameters of interest can be estimated using standard panel estimators applied to the weighted balanced panel.

An example is the estimation of a production function. The firms observed at any given period can be considered a random draw from the population of potential firms. Yet, due to nonrandom exit and entry, restricting analysis to only those firms that existed in more than one period can potentially bias the results. (See Olley and Pakes 1996 for an example of the importance of accounting for entry and exit, and Griliches and Mairesse 1998 for a survey of the literature dealing with this problem.) The proposed method combines the full information contained in the unbalanced panel, while still controlling for the unobserved individual effects.

### 2.5 Taking Into Account Sampling Error in the Moment Restrictions

In the preceding sections, it was assumed that the additional information was in the form of exactly known moments, \( h^* \). This is an adequate setup when the dataset providing these moments is much larger than the primary dataset; for example, if it is a census. However, in the example considered in Section 3, as in many other examples, this will not be the case. The results previously given can be generalized, as shown by Hellerstein and Imbens (1999), to the case where \( h^* \) is not known with certainty but instead there is an estimate, \( \tilde{h} \), of \( h^* \) based on a random sample of size \( M \), that is, \( \tilde{h} = 1/M \sum_i h(z_i) \). This estimate satisfies \( \sqrt{M}(\tilde{h} - h^*) \rightarrow \N(0, \Delta_h) \), with \( \Delta_h = E[h(z) \cdot h(z)] \). Here \( \tilde{h} \) is assumed to be independent of the primary sample \( \{z_1, z_2, \ldots, z_N\} \).

Now \( \theta \) can be estimated by the algorithm described earlier, except now \( \tilde{h} \) instead of \( h^* \) is used to construct the additional moments. This additional step must be taken into account when investigating the properties of the estimator as the number of observations in both datasets, \( N \) and \( M \), goes to infinity. If \( N/M \) converges to 0, then in large datasets the variance in the second dataset can be neglected, as in the case considered earlier. On the other hand, if \( M/N \) converges to 0, then the second dataset cannot help adjust for sample selection. Therefore, this work considers only the case where the ratio \( M/N \) converges to a constant \( k \).

The estimate of \( \theta \) can be obtained by using the additional moments and standard GMM results. Without loss of generality, and to facilitate comparison with the previous exposition, assume that \( M/N \) is exactly equal to some integer \( k \). Let \( \tilde{z}_i \) consist of \( \{z_i, h_{i1}, \ldots, h_{ik}\} \). Assume \( N \) observations \( \tilde{z} \). Using the logistic selection probability, the estimating equations are

\[
0 = \sum_{i=1}^N \left( \frac{\psi(z_i, \theta_1) \cdot \tau(1 + \exp(h(z_i) \cdot \theta_2))}{(\tilde{h}(z_i) - \tilde{h}) \cdot \tau(1 + \exp(h(z_i) \cdot \theta_2))} - \frac{1 - \tau(1 + \exp(h(z_i) \cdot \theta_2))}{1/k \sum_{j=1}^k (h_{ij} - \tilde{h})} \right).
\]

Solving this leads to \( \tilde{h} = 1/M \sum_{i=1}^N \sum_{j=1}^k h_{ij} \), whereas the rest of the estimators are the same as those produced by solving (7). The following proposition describes the asymptotic behavior of the estimator.

**Proposition 3.** Suppose that the conditions of Proposition 2 hold; then the estimator \( \theta^w \) for \( \theta_1^w \) has the following asymptotic properties:

\[
\sqrt{N}(\hat{\theta}_1^w - \theta_1^w) \rightarrow N(0, E[\partial \psi/\partial \theta_1^w]^{-1} E[\psi \psi]) - E[\partial \psi/\partial \theta_1^w]^{-1} E[\psi \psi]
\]

\[
- E[\partial \psi/\partial \theta_1^w]^{-1} E[\psi \psi] + V\sum_{k} \Delta_k V^w,
\]

where

\[
V = E[\partial \psi/\partial \theta_1^w]^{-1} E[\partial \psi/\partial \theta_1^w]^{-1} E[\psi w] V^w + E[\partial \psi/\partial \theta_1^w]^{-1} E[\psi \psi] V^w.
\]

\[
\tilde{\psi}(Z, \theta) = \psi(Z, \theta_1), w(Z, \theta_2),
\]

\[
\tilde{h}(Z, \theta_2) = \frac{(\tilde{h}(Z) - \tilde{h}) \cdot w(Z, \theta_2)}{w(Z, \theta_2) - 1},
\]

and \( E_z[] \) denotes expectations taken with respect to \( f(z|D = 1) \).

**Proof.** This is the same as the proof of Proposition 2 (see Hellerstein and Imbens 1999 for details).

### 3. An Empirical Application

In this section the method described in the previous sections is applied to the DTP data used by Riddler (1992). Section 3.1 presents the data and initial analysis, and Section 3.2 gives the results based on the proposed procedure.
3.1 Data and Preliminary Analysis

The data are taken from the DTP, the purpose of which was to evaluate the change in the use of public transportation over time as price was increased. The first wave of the panel consists of a stratified random sample of households in 20 towns interviewed in March 1984. Each member of the household, older than 11 years was asked to keep a travel log of all the trips taken during a particular week. A trip starts when the household member leaves the home and ends when he or she returns. Several questions were asked about each trip, but this information was not used here. (For a detailed description of the DTP, and a survey of research conducted with it, see van Wissen and Meurs 1990.)

After the initial interview in March 1984, each participant who did not drop out was subsequently interviewed twice a year, in September and March. The September interviews did not ask the participants to fill out a detailed log and thus were somewhat different than the March interviews. This study follows that of Ridder and examines only the March interviews (thus also avoiding any seasonal effects). The original panel suffered from heavy attrition, as seen in Table 1. To keep the number of participants constant, additional refreshment samples were taken from the population. For the rest of this article it is assumed that the refreshments samples were taken as random samples from the population. In reality, the refreshment samples were sampled randomly with the same stratification as the original sample but with different weights, to compensate for the heavier attrition in some strata. The methods used here can easily deal with this case; however, for simplicity of presentation, this aspect is ignored. To demonstrate the proposed method, the focus is on explaining the number of trips taken as a function of household characteristics.

The average number of trips in different subsamples of the panel is given in Table 1. Examining the bottom row reveals that there is virtually no change in the average number of trips over the different waves. However, examining any one of the other rows shows a clear downward trend. This is not seen in the total, because the number of trips increases with the number of waves of participation. Surprisingly, these two effects completely offset each other.

Table 1. Number of Households and Average Number of Trips by Wave and Wave of Attrition

<table>
<thead>
<tr>
<th>Dropsouts</th>
<th>Wave</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>731</td>
<td>45.4</td>
<td>(33.2)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>178</td>
<td>57.7</td>
<td>(33.5)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>185</td>
<td>62.2</td>
<td>(37.2)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>666</td>
<td>62.8</td>
<td>(34.6)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1,760</td>
<td>55.0</td>
<td>(35.1)</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: For each wave, the left column presents the number of households and the right column presents the average number of trips and the standard deviation (in parentheses).

The data also contain information on various explanatory variables, described in the Appendix. The summary statistics demonstrate that the distribution of the variables differs in the different waves of the original panel. These differences in the distribution of the explanatory variables do not explain the pattern observed in the number of trips (Ridder 1992, table 5). This leaves (a combination of) the following as possible explanations to the pattern in the number of trips observed in Table 1. Either there is a real downward trend in mobility or, due to nonrandom attrition, there are (nonrandom) differences in the distribution of the unobserved determinants of the total number of trips. Nonrandom attrition is a real concern given that the sample means of the variables in the waves of the original panel differ from the means in the refreshment panel. The question is whether this attrition completely explains the patterns of Table 1 or rather is part of the pattern due to a real change in mobility.

To answer this question, a series of regressions presented in Table 2 is computed. The following conclusions can be reached from the results in the table. First, if no correlation is assumed between the unobserved determinates of the number of trips and the independent variables, then it can be concluded that there was no drop in mobility. This can be seen from the results in the first two columns of the table, which are based on ordinary least squares regressions in the original sample plus the three refreshment samples. The three wave dummy variables are jointly statistically significant at a 5% level. However, the dummy variables for the third and fourth waves are not jointly significant. Therefore, it might be concluded that a drop in mobility occurred during the second period, but not overall.

Second, using the panel structure the assumption that unobserved household-specific effects are not correlated with the independent variables can be examined. The fixed- and random-effects results can be combined to compute the standard Hausman test, which is strongly rejected for both the balanced and unbalanced panels (85.7 for the unbalanced panel, 28.7 for the balanced panel). Therefore, it seems that the assumption made by the regression based on the repeated cross-sections is not valid, and that the estimates presented in the first two columns are inconsistent. This also suggests that the only results in the table that are consistent are the within estimates, which point toward a large downward trend in mobility. If either the unbalanced or balanced panels are representative of the population, then it could be concluded that there was a downward trend in mobility. However, because it has already been demonstrated that the attrition from the sample is nonrandom, this conclusion cannot be reached based on the results presented in any of the columns of Table 2.

Therefore, to determine whether there was a change in mobility we require an estimator that deals both with the attrition from the sample and the potential correlation between the explanatory variables and the error terms. The estimator introduced in the previous sections has these properties and is used here. An additional estimator that could potentially deal with these issues is the one suggested by Hausman and Wise (1979). Ridder (1992) explored this estimator and found that it fails to alert of nonrandom attrition and hence also fails to treat it. Ridder attributed this failure to an implicit restriction that forces the covariance of the individual effects in the
Nevo: Using Weights to Adjust for Sample Selection

selection and regression equations to have the same sign as the covariance of the random shocks in the two equations (see Ridder 1992, sec. 5, for details).

3.2 Results Using the Proposed Procedure

To evaluate the performance of the proposed procedure, two measures are examined. First, out-of-sample prediction of the model is studied by testing the ability of the weights, computed based on only the first and last cross-sections, to match the weighted balanced panel moments with the moments from the refreshment panel. Next, estimates of the regression coefficients are computed, similar to those presented in Table 2, which answer the question of whether or not there was a real change in mobility in the Netherlands.

To test the out-of-sample predictive power of the methods, weights are computed by solving (7) using moments from the first (unbalanced) wave of the panel and the last wave of the refreshment samples. Table 3 demonstrates the effects of weights on the sample statistics. Three different sets of weights were computed. The first set was computed using moments on only the explanatory variables. This assumes a (particular) ignorable (conditional on observable variables) selection model. If selection is a linear function of only the explanatory variables, then these weights should control fully for selection. The second set of weights was computed using only the first moments of the dependent variable (TOTRIP).

Finally, all of the variables were used. In all cases, only the first moments computed from the first and last waves were used.

These weights were attached to the balanced observations, and the sample statistics for this weighted sample were computed. Table 3 presents the weighted sample averages for the second and third waves. Because the weights were computed using only the moments from the first and fourth waves, these can be considered out-of-sample predictions. These moments can be used to construct a formal or informal test of the selection model. Weights that control fully for selection should render the differences between the moments of Table 3 and the appropriate moments in Table A.2 as statistically insignificant. The logic behind this is the same as that of the usual test of overidentification.

The results in Table 3 lead to the following conclusions. First, the weighted samples are more representative of the refreshment population, and thus of the population of interest. For all three selection models, the fit is much better for the second wave than for the third wave. The ignorable selection model, which uses only the moments of the explanatory variables, is quite strongly rejected. The third model, which uses both the dependent and independent variables, fits the second wave but not the third, suggesting that the third wave is somewhat different.

One explanation of this last result can be seen by examining different nonignorable models of selection. Under the model in which both the regression and selection equations are a function of fixed (over time) individual-specific effects, selection should be fully controlled for by conditioning on the dependent variable. The difficulty in predicting the moments in the third wave suggest that this model is wrong. Therefore,
it is not surprising that Ridder (1992) found that the model of Hausman and Wise (1978), which makes these assumptions about the individual-specific effects, does not fit this dataset. To deal with the poor fit of the third-wave moments, the selection probability is allowed to depend also on second- and third-wave variables.

Table 4 presents the weighted regression coefficients computed using the balanced panel. For each model, both a fixed-effects estimator and a random-effects estimator are computed. The models differ in selection probability. Model 1 models the selection probability as a function of the dependent and independent variables in the first and fourth waves. It is equivalent to the selection model used to produce the results in columns 3 and 6 of Table 3. Because the analysis of the results in Table 3 suggests that this selection model does not fully capture the selection in the third wave, in model 2 the weights are computed as a function of all variables in the third wave and the dependent variable in the other waves.

The following conclusions can be drawn from the results. First, a Hausman test of the equality of the fixed- and random-effects estimates is rejected. Despite this, the coefficients on the wave dummy variables are similar in both the fixed- and random-effects models. Second, in general the weighted coefficients are between the ordinary least squares results from the repeated cross-section and the (unweighted) balanced panel results. Finally, and most importantly, even after selection is controlled for in several different ways, the negative trend in mobility is still present. It is true that attrition makes this trend seem larger than it really is; however, it still exists. The drop in mobility is particularly large during the third wave. This is especially true in model 2, which allows for a more general model of selection in the third wave.

Given the count nature of the data, the foregoing analysis was repeated using a Poisson model for the number of trips. The only change was that the moments were now nonlinear in the parameters. The qualitative effects were similar to those found earlier. In particular, the estimates from a fixed-effects (conditional) Poisson model suggest a downward trend in mobility, with a larger drop in the third wave. The estimates suggest that using model 2, the probability of taking a trip is reduced by about 5% in the second and fourth waves relative to the first wave and is reduced by 15% in the third wave. Because the average number of trips is roughly 50, this is close to what the results of Table 4 imply.

4. FINAL REMARKS

This article proposes a weighting method that takes advantage of additional information to treat sample selection bias. Moments available from other sources are exploited to adjust for sample selection in the primary data. The method is applicable only in cases where these moments are available or can be estimated. Using these additional moments, the selection probability is computed, and this is used to inflate the data. The estimator can deal with both ignorable and nonignorable selection mechanisms.

A few applications in which these additional moments are thought to be available have been out-lined. The application presented in detail is characterized by refreshment samples from the target population, which were taken to deal with attrition of the original sample. This is not typical of economic data, but perhaps it should be. Perhaps rather than putting great effort into maintaining panels that follow individuals or firms over a long period, more attention should be focused on obtaining additional cross-sectional draws from the population of interest.

An area for future work is a comparison of the method proposed here to alternatives. A full comparison, either theoretical or empirical, is beyond the scope of this article. However, some idea can be obtained by building on other work. Ridder (1992) used the same data reported here to report results in the spirit of Ahn and Powell (1993). This method does not use the auxiliary information discussed in this article, but it can be extended to do so. Such an extension would require different assumptions than those made in this article and is an interesting topic for future work.

Many public use datasets are accompanied by weights that are treated as known. The method proposed here allows the
researcher to compute weights even if these are not available or to compare the weights to the ones provided (because in some cases it is not clear how the provided weights are computed). However, even if weights are provided and the researcher is satisfied with how they were computed, there still might be an efficiency argument for estimating the weights. Hirano, Imbens, and Ridder (2000) showed that in some cases estimators that weight observations by the inverse estimate of the selection probability are more efficient than estimators that use the true selection score. Furthermore, they related their result to the result of Wooldridge (1999) that shows that in the context of stratified sampling, it is more efficient to use estimated weights instead of known sampling probabilities. The Monte Carlo results of Nevo (2002) seem to suggest that a similar result might be applicable here.

ACKNOWLEDGMENTS

The author thanks Josh Angrist, Moshe Buchinsky, Francesco Caselli, Gary Chamberlain, Zvi Eckstein, Zvi Griliches, Jim Heckman, Kei Hirano, Guido Imbens, Jim Powell, and participants in the Econometrics in Tel-Aviv 1995 workshop and Camp Econometrics 1998 for useful discussions and comments on earlier versions of the manuscript, and Geert Ridder for making his data available.

APPENDIX: VARIABLES AND SAMPLE STATISTICS

This appendix describes the variables available in the data (Table A.1) and provides, their sample statistics in the different waves and subsamples (Table A.2).

### Table A.1. The Explanatory Variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPER</td>
<td>Number of persons over age 11</td>
</tr>
<tr>
<td>N1218</td>
<td>Number of persons age 12–18</td>
</tr>
<tr>
<td>N1938</td>
<td>Number of persons age 19–38</td>
</tr>
<tr>
<td>INC1</td>
<td>Annual net family income $&lt; 17,000 guilders</td>
</tr>
<tr>
<td>INC2</td>
<td>Annual family income 24,000–37,999 guilders</td>
</tr>
<tr>
<td>INC3</td>
<td>Yearly net family income 38,000 guilders</td>
</tr>
<tr>
<td>CITY</td>
<td>Inhabitant of large city (&gt;$500,000)</td>
</tr>
<tr>
<td>NCAR</td>
<td>Total number of cars in household</td>
</tr>
<tr>
<td>NLIC</td>
<td>Total number of driving licenses in household</td>
</tr>
<tr>
<td>FAMT1</td>
<td>Household with head under age 35 and no children</td>
</tr>
<tr>
<td>FAMT2</td>
<td>Household with children younger than age 12</td>
</tr>
<tr>
<td>FAMT3</td>
<td>Household with head over age 65</td>
</tr>
<tr>
<td>EDLO</td>
<td>Highest education of head primary school or lower</td>
</tr>
<tr>
<td>EDHI</td>
<td>Highest education of head university or higher</td>
</tr>
</tbody>
</table>


### Table A.2. Sample Averages of Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Wave 1</th>
<th>Wave 2</th>
<th>Wave 3</th>
<th>Wave 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unbal</td>
<td>Bal</td>
<td>Unbal</td>
<td>Bal</td>
</tr>
<tr>
<td>TOTRIP</td>
<td>55.02</td>
<td>62.80</td>
<td>54.90</td>
<td>56.70</td>
</tr>
<tr>
<td>NPER</td>
<td>2.19</td>
<td>2.28</td>
<td>2.23</td>
<td>2.28</td>
</tr>
<tr>
<td>N1218</td>
<td>0.29</td>
<td>0.33</td>
<td>0.36</td>
<td>0.33</td>
</tr>
<tr>
<td>N1938</td>
<td>1.00</td>
<td>1.12</td>
<td>1.05</td>
<td>1.05</td>
</tr>
<tr>
<td>FAMT1</td>
<td>0.12</td>
<td>0.12</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>FAMT2</td>
<td>0.24</td>
<td>0.29</td>
<td>0.29</td>
<td>0.32</td>
</tr>
<tr>
<td>FAMT3</td>
<td>0.14</td>
<td>0.09</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>INC1</td>
<td>0.19</td>
<td>0.13</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>INC2</td>
<td>0.32</td>
<td>0.36</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>INC3</td>
<td>0.27</td>
<td>0.29</td>
<td>0.30</td>
<td>0.28</td>
</tr>
<tr>
<td>EDLO</td>
<td>0.44</td>
<td>0.37</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>EDHI</td>
<td>0.18</td>
<td>0.23</td>
<td>0.26</td>
<td>0.27</td>
</tr>
<tr>
<td>CITY</td>
<td>0.10</td>
<td>0.06</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>NCAR</td>
<td>0.85</td>
<td>0.89</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>NLIC</td>
<td>1.35</td>
<td>1.48</td>
<td>1.49</td>
<td>1.47</td>
</tr>
</tbody>
</table>

NOTE: The columns labeled Unbal, Bal, and Refr present averages for the cross-section of the original panel, the balanced subpanel, and the refreshment samples.