TECHNICAL WORKING PAPER SERIES

USING WEIGHTS TO ADJUST FOR SAMPLE SELECTION WHEN AUXILIARY INFORMATION IS AVAILABLE

Aviv Nevo

Technical Working Paper 275
http://www.nber.org/papers/T0275

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2001

I wish to thank Josh Angrist, Moshe Buchinsky, Francesco Caselli, Gary Chamberlain, Zvi Eckstein, Zvi Griliches, Jim Heckman, Kei Hirano, Guido Imbens, Jim Powell, as well as participants in the Econometrics in Tel-Aviv workshop and Camp Econometrics for useful discussions and comments on earlier versions and Geert Ridder for making his data available. The views expressed in this paper are those of the author and not necessarily those of the National Bureau of Economic Research.

Using Weights to Adjust for Sample Selection When Auxiliary Information is Available
Aviv Nevo
NBER Technical Working Paper No. 275
November 2001
JEL No. C23

ABSTRACT

In this paper I analyze GMM estimation when the sample is not a random draw from the population of interest. I exploit auxiliary information, in the form of moments from the population of interest, in order to compute weights that are proportional to the inverse probability of selection. The essential idea is to construct weights, for each observation in the primary data, such that the moments of the weighted data are set equal to the additional moments. The estimator is applied to the Dutch Transportation Panel, in which refreshment draws were taken from the population of interest in order to deal with heavy attrition of the original panel. I show how these additional samples can be used to adjust for sample selection.

Aviv Nevo
549 Evans Hall #3880
Department of Economics
University of California, Berkeley
Berkeley, CA 94720-3880
and NBER
nevo@econ.berkeley.edu

1. INTRODUCTION

Sample selection arises when the observed 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 paper 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, we want to study a panel of individuals in which we believe there is an unobserved individual-specific effect that is correlated with the independent variables. Common solutions include "within" or first difference estimators. However, in order to implement these types of estimators we need at least two observations for each individual. If the panel also suffers from non-random attrition, then the subset of individuals observed for more than one period is no longer a random draw from the population of interest. Assume we have at our disposal additional samples, which are representative of each of the cross-sections, but also suffer from attrition and therefore do not allow consistent estimation of the parameters of interest. An example of such a data set is the Dutch Transportation Panel (DTP) considered by Ridder (1992) and used below.

The procedure proposed in this paper suggests one possible way to use these additional samples, the refreshment samples. I use the refreshment samples to attach a weight to each observation in the balanced sub-panel such that moments in the weighted sample are set equal to corresponding moments in the refreshment samples. I propose estimating the parameters of interest using standard panel estimators and the weighted balanced panel.

The proposed method exploits additional data to adjust for the selection bias by using it to estimate the selection probability, the propensity score. Here I use the propensity score 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; or Ridder, 1992, for the data structure discussed below), to match observations (Ahn and Powell, 1993) or to impute the missing observations (Hirano et al, 1998, for the data used below).

The advantage of the weighting method proposed here, over the alternative methods of using the additional information mentioned above, is its simplicity. Once the weights have been computed the researcher can conduct the same analysis, 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 GMM problem (as below), or by solving a linear programming problem. The latter approach connects the method proposed here to information–based alternatives to GMM (see below).

Although the refreshment samples found in the DTP are not typical of economic data, I claim that the method proposed here is general. For example, Census and Annual Survey of Manufacturers data are sources of random draws from the population of interest and can be matched with smaller data sets 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 that is not frequently done.

1.1 *Previous Literature*

A survey of the literature on sample selection is beyond the scope of this paper.[1] However, I want to relate the method prosed here to previous work. The use of weights to correct for sample selection is not new to this paper. For example, many data collecting agencies provide sampling

[1]See, for example, Heckman (1979, 1987, 1990), Heckman and Robb (1986), Little and Rubin (1987), Newey, Powell and Walker (1990), Manski (1994), Angrist (1995), Kyriazidou (1997), Heckman, Ichimura, Smith and Todd (1998) and references therein.

weights to correct for the sampling procedure (the U.S. Census, for example, provides such weights.) Manski and Lerman (1977) propose a method to compute weights for choice-based samples.[2] In both these examples the selection probability is known, while here I estimate the probability of selection, and the implied weights, jointly with the parameters of interest. Methods that estimate the weights jointly with the parameters of interest also exist. For example, Cassel, Sarndal and Wretman (1979), Koul, Susarla and Van Ryzin (1981) and 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 data set, but for which some information is known through the additional moments. I also allow the weights to be a function of the dependent variable. In other words, I deal with a non-ignorable 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 Scheruan, 1983, Section 3; Little and Wu, 1991). I extend these methods by examining a logistic selection equation and by looking at regression functions rather than contingency tables.

The estimator I use here can be related to information-theoretic alternatives to GMM (Back and Brown, 1990; Qin and Lawless, 1994; Imbens, 1997; Kitamura and Stutzer, 1997; Imbens, Spady and Johnson, 1998; Hellerstein and Imbens, 1999). Standard GMM methods implicitly estimate the distribution of the data by the empirical distribution (i.e., giving each observation a density of $1/N$). The information-theoretic alternatives use over-identifying moment conditions to improve on the estimates of the data distribution, by estimating the distribution jointly with the original parameters of interest. The focus of this literature has been on improving efficiency. In a

[2]See also Cosslett (1981) and Wooldridge (1999).

related paper (Nevo, forthcoming) I show that the estimator proposed here can be related to these methods. I make the difference between the sampled population, from which the sample is drawn, and the population of interest explicit. Therefore, I can deal with bias in the estimation of the distribution. It turns out that when the sampled population equals the target population the estimator proposed here is equivalent to the one proposed by Imbens, Spady and Johnson (1998).

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

2. THE MODEL AND PROPOSED ESTIMATOR

2.1 *The Setup*

Suppose we observe N independent realizations $\{z_1, z_2, ..., z_N\}$ of a (multi-variate) random variable Z, with its support, χ, a compact subset of $\Re^P$. In the population Z has a pdf $f(Z)$. Let θ_1^* denote the true value of the parameters of interest, and $\theta_1^* \in \Theta_1$ where Θ_1 is a compact subset of $\Re^K$.

Assumption A0: θ_1^* uniquely set $E\left[\psi(Z,\theta_1^*)\right]=0$. Furthermore, the moment function, $\psi:\chi\times\Theta_1\to\Re^K$, is twice continuously differentiable with respect to θ, measurable in Z, and $E\left[\psi(Z,\theta_1^*)\psi(Z,\theta_1^*)'\right]$ and $E\left[\partial\psi(Z,\theta_1)/\partial\theta_1'\right]$ are of full rank.

A way of estimating θ_1^* would be to follow the analogy principle by choosing the estimate

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

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 above estimate is biased and inconsistent.

Let D_i=1 if and only if z_i is fully observed.[3] From Bayes' 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 zero.*

This assumption requires that any point in the support has a (strictly) positive probability of being observed, which 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. Since the conditioning vector, Z, includes the dependent variable, the selection probability, $P(D=1|Z)$, will equal either zero or one, depending on the value of *y*. Our method will not be operational in such a case. If the correlation between the selection rule and the variable of interest is less than one this assumption will be satisfied (for example the model considered by Heckman, 1974).

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

$$f(Z) = f(Z|D=1)\cdot P(D=1)/P(D=1|Z). \quad \textbf{(1)}$$

Therefore,

$$E\left[\psi(Z,\theta_1^*)\right] = \int \psi(z,\theta_1^*) f(z)dz = \int \psi(z,\theta_1^*) f(z|D=1)\cdot P(D=1)/P(D=1|z)dz = 0$$

and the correct analog estimator becomes

$$\hat{\theta}_1 \text{ s.t. } \frac{P(D=1)}{N}\sum_{i=1}^{N}\frac{1}{P(D=1|z_i)}\psi(z_i,\hat{\theta}_1) = 0. \quad \textbf{(2)}$$

The empirical distribution of the selected sample can be used to consistently estimate $f(Z|D=1)$.

[3] In a cross-sectional context *D*=0 could imply, for example, that the covariates are observed while the outcome variable is not. In a panel example, *D*=1 for the balanced sub-sample, while *D*=0 for individuals who are present only in some periods.

Therefore, if we know $P(D=1|Z)$ then equation (2) can be used to consistently estimate θ_1.[4]

In general the selection probability is unknown and will have to be estimated. In order to estimate the selection probability, $P(D=1|Z)$, I assume exact knowledge of the expectation, h^*, in the population, of an R-dimensional function of Z, denoted $\bar{h}(Z)$. Formally, $h^* = E\left[\bar{h}(Z)\right] = \int \bar{h}(z) f(z) dz$. Examples include $\bar{h}(Z) = Y$, where the researcher knows the mean of the dependent variable, or $\bar{h}(Z) = Y \cdot X$, where the researcher knows the (non-centered) 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 below will demonstrate this.

Denote $h(Z) = \bar{h}(Z) - h^*$. Note, that $E\left[h(Z)\right] = 0$. For now I assume these moments are known and defer to later discussion of where they come from and the possibility that they are known with error.

Assume, $P(D=1|Z) = P(D=1|Z,\theta_2)$, where $\theta_2^* \in \Theta_2$ and Θ_2 is a compact subset of $\Re^R$. Define

$$\bar{\psi}(Z,\theta) = \begin{pmatrix} \dfrac{\psi(Z,\theta_1)}{P(D=1|Z,\theta_2)} \\ \dfrac{h(Z)}{P(D=1|Z,\theta_2)} \end{pmatrix} \qquad \textbf{(3)}$$

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

$$E\left[\bar{\psi}(Z,\theta^*)\,|\,D=1\right] = \int \bar{\psi}(z,\theta) f(z\,|\,D=1) dz = 0. \qquad \textbf{(4)}$$

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.

[4]This is the fundamental idea in using sampling weights to correct for non-random sampling in surveys and estimation in choice based sampling (Manski and Lerman, 1977; and Cosslett, 1981). See also Wooldridge (1999).

Let $Z_i = (Z_{i1}, Z_{i2})$ be a bivariate binary random variable, where Z_{it} 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 non-random, i.e., $f(Z_{i1}, Z_{i2}) \neq f(Z_{i1}, Z_{i2}|D_1 = 1)$, analysis based on the individuals that are 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. I now ask 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 sub-sample of individuals that are observed in both periods. Assuming the original sample is a random sample from the population then the probability $P(Z_{i1})$ can be identified. Suppose we have additional information on $P(Z_{i2})$. For example, this information can be obtained by taking a random draw from the population at t=2. This additional information in not enough 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 then identification is not trivial. If the additional information is in the form of marginal moments this is even more so. Hirano et al. (1998) prove that by assuming a particular functional form for the selection probability, the probability $P(D_i|Z_i)$ is identified (Hirano et al, 1998, Theorem 2). Since I assume that the additional information comes in the form of moments, I require slightly different conditions for identification.

Assumption A2: The matrix $E\left[h(Z)\cdot h(Z)'|D = 1\right]$ 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 logisitic model of selection used below, i.e., $g(a) = \exp(a)/(1 + \exp(a))$.

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

Proof: Assumptions A2-A3 promise that the equations $E\left[h(Z)/g(h(Z)'\theta_2)|D=1\right]=0$ have a unique solution for θ_2 such that $\theta_2=\theta_2^*$. Therefore, under assumptions A0 and A1 the system of equations $E\left[\psi(Z,\theta_1)/g(h(Z)'\theta_2)|D=1\right]=0$ has a unique solution for θ_1 such that $\theta_1=\theta_1^*$. This and standard GMM theory (Hansen, 1982; Newey and McFadden, 1994) proves 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, i.e., the dimension of θ_2. Assumption A3 requires more than this. The selection probability depends on the functions $h(Z)$, for which we know the expectation in the population.

2.3 *Estimation*

I propose the following three step procedure for estimating the parameters of interest. In the first step, the probability of selection is modeled as

$$P(D=1|Z,\theta_2)=\frac{e^{g^*(Z,\theta_2)}}{1+e^{g^*(Z,\theta_2)}} \qquad \textbf{(5)}$$

where $g^*(Z, \theta_2)$ is an unknown function. At this point no restrictions have been imposed since by defining $g^*(Z, \theta_2)$ appropriately equation (5) can fit any selection model. I approximate the unknown function $g^*(Z, \theta_2)$ by a polynomial, $h(Z)'\theta_2$, with unknown coefficients, θ_2. The model of selection is largely driven by data availability, the presence of the functions $h(\cdot)$. Assuming we have a rich set of moments available to create $h(\cdot)$ the model can be derived from economic modeling. In this case the estimates of θ_2 might be of independent interest.

The conditioning vector in equation (5) may include also the dependent variable, in the

estimation equation. Therefore, this selection model is non-ignorable. One could write the selection probability as a function of observed variables as well as an individual specific unobserved effect. Since the conditioning vector includes the dependent variable in the main estimation equation this type of selection model is covered by our 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 $1/P(D=1|Z,\theta_2)$. Formally, the weights are computed by solving the following set of equations

$$\sum_{i=1}^{N} w_i(z_i,\theta_2)h(z_i) = 0$$
$$\sum_{i=1}^{N} w_i(z_i,\theta_2) = 1 \qquad \textbf{(6)}$$

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

By solving equation (6) the weighted sample counterpart of E[$h(Z)$] is set to zero. Thus, we exploit the additional moments, h^*, in order to estimate the parameters of the selection process. An alternative interpretation of the equation (6) comes from an information-based criterion. This interpretation relates the method proposed here to information-theoretic alternatives to GMM (see references given in Section 1.1), as well as Little and Wu (1991). See Nevo (forthcoming) for details.

Using $g^*(Z,\theta_2) = h(Z)'\theta_2$ and substituting equation (5) into equation (6) we obtain the following system of equations

$$\sum_{i=1}^{N} w(z_i,\theta_2)\cdot h(z_i) = \sum_{i=1}^{N} \frac{\tau}{N\cdot P(D=1|z_i,\theta_2)}\cdot h(z_i) = \sum_{i=1}^{N} \frac{\tau}{N}\cdot\left(1+\frac{1}{e^{h(z_i)'\theta_2}}\right)\cdot h(z_i) = 0$$
$$\sum_{i=1}^{N} w(z_i,\theta_2) = \sum_{i=1}^{N} \frac{\tau}{N}\cdot\left(1+\frac{1}{e^{h(z_i)'\theta_2}}\right) = 1 \qquad \textbf{(7)}$$

where $\tau = P(D=1)$. For some models of selection, 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, τ, will not be identified separately from a scale parameter. The logistic selection probability I propose in this paper does not have this property and this last equation will be more than just a normalization, it will actually change the relative weights.[5]

In the final step, the weights that solve equation (6) are used to obtain analog estimates of the parameters of interest, θ_1^*. Formally, the estimate is given by

$$\hat{\theta}_1^W \quad s.t. \quad \sum_{i=1}^{N} w_i \psi(z_i, \hat{\theta}_1^W) = 0 \qquad \textbf{(8)}$$

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

Proposition 2 *Suppose that* z_i (i=1,2,...) *are i.i.d with the empirical distribution converging to* $f(z|D=1)$, *and* (*i*) *Assumptions A0-A3 are satisfied;* (*ii*) $\Theta_1 \times \Theta_2$ *is compact;* (*iii*) *the moment functions,* $\bar{\psi}(z,\theta)$, *defined by equation* (*3*) *are twice continuously differentiable in* θ*;* (*iv*) $\mathrm{E}\left[\bar{\psi}(Z,\theta)'\bar{\psi}(Z,\theta)\right] < \infty$. *Then* $\hat{\theta}_1^W \to \theta^*$ *and*

$$\sqrt{N}(\hat{\theta}_1^W - \theta_1^*) \to \mathrm{N}\left(0, E_s[\partial\tilde{\psi}/\partial\theta']^{-1}\left(E_s[\tilde{\psi}\tilde{\psi}'] - E_s[\tilde{\psi}\tilde{h}']E_s[\tilde{h}\tilde{h}']^{-1}E_s[\tilde{h}\tilde{\psi}']\right)E_s[\partial\tilde{\psi}/\partial\theta']^{-1}\right)$$

where

$$\tilde{\psi}(Z,\theta) = \psi(Z,\theta_1)\,w(Z,\theta_2), \quad \tilde{h}(y,x) = \begin{pmatrix} h(Z)\,w(Z,\theta_2) \\ w(Z,\theta_2) - 1 \end{pmatrix}$$

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

---

[5] Consider, for example, a linear probability model, i.e., $P(D=1|Z,\theta_2) = \alpha + Z'\beta$. The weights in such a case will be proportional to $[\alpha(1+\alpha^{-1}z_i'\beta)]^{-1}$. Therefore, normalizing the weights to sum up to one will influence only the estimate of the constant, but not the relative weights. However, if the probability of selection is logistic, i.e., $P(D=1|Z,\theta_2) = \exp(h(Z)'\theta_2)/(1+\exp(h(Z)'\theta_2))$ the weights will be proportional to $1+\exp-(\alpha+z_i'\beta)$. Now a normalization will not be fully absorbed in the constant term, and will influence the relative weights as well as their absolute value.

Proof: The expectation of the stacked moment conditions is set to zero at the true parameter value (equation (4)). The assumptions of the proposition and standard GMM theory provide the result (Newey, 1984, Pagan, 1986 and Newey and McFadden, 1994).■

The proof of the proposition stacks the moments as in equation (3) and considers them as just-identified estimation equations. The actual estimation can be obtained by solving these equations in one step (therefore combining the second and third steps in the above discussion.) However, as a computational issue it is simpler to obtain the solution by first solving equation (6), and plugging the solution into equation (8), as proposed in the above algorithm.[6]

The standard errors can alternatively be estimated by bootstrapping. This will work in the following way. First, we generate a bootstrap sample by sampling from the original sample. Next, we solve for the value of the parameter that sets the moments in equation (3) to zero. We repeat these two steps to obtain the bootstrap distribution of the parameters.

The proposition assumes that the value of the population moments is known exactly. This seems an adequate assumption if the second data set is much larger than the primary data set. However, in many interesting problems this will not be the case. Section 2.5 extends the results to take account of 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 Dutch Transportation Panel (DTP) used by Ridder (1992) and to which the proposed estimator is applied below. The unique design of this data set called for draws of refreshment samples in order to deal with the heavy (seemingly non-random) attrition of the

[6]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.

original panel. These refreshment samples are not characteristic of economic data. But as the rest of this section demonstrates, under reasonable assumptions familiar economic data sets can fit into the proposed framework.

Suppose we have a combination of census data and smaller data sets (as in Imbens and Lancaster, 1994; or Hellerstein and Imbens, 1999.) We can think of the larger data set as a random draw from the population, which provides estimates of population moments for certain variables. The smaller data set is not a random draw from the population of interest, however, it is much richer. For example, Gottschalk and Moffitt (1992) document the differences between the NLS (which is a small and rich data set, but suffers from attrition) and the CPS (which is not as rich, but is much larger and representative of the population.)[7] Hellerstein and Imbens (1999) exploit this in order to correct for ability in the NLS. The method proposed here, which is similar to their approach,[8] 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 data set.

The data structure which is perhaps best suited for our method is a panel structure. Consider the setup described in the Introduction. 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, we 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 sub-sample 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

[7] See also MaCurdy, Mroz and Gritz (1998).

[8] The main difference is that I explicitly model the selection probability and therefore the weights are computed to fit this selection model. The weights Hellerstein and Imbens compute implicitly imply a linear probability selection model.

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 we observe at any given period can be considered a random draw from the population of potential firms. Yet, due to non-random exit and entry, if we restrict analysis to only those firms that existed in more than one period we can potentially bias the results.[9] 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 previous sections I assumed that the additional information was in the form of exactly known moments, h^*. This is an adequate setup when the data set providing these moments is much larger than the primary data set, for example if it is a census. However, in the example considered below, 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 we do not know h^* with certainty. Instead we have an estimate $\hat{h}$ of h^*, based on a random sample of size M, i.e., $\hat{h} = 1/M \sum h(z_i)$. This estimate satisfies $\sqrt{M}(\hat{h} - h^*) \rightarrow N(0, \Delta_h)$, with $\Delta_h = E[h(z) \cdot h(z)]$. I assume $\hat{h}$ is independent of the primary sample $\{z_1, z_2, ..., z_N\}$.

We can estimate θ by the algorithm described above, except now we use $\hat{h}$ instead of h^* to construct the additional moments. We have to take this additional step into account when investigating the properties of the estimator as the number of observations in both data sets, N and M, goes to infinity. If N/M converges to zero then in large data sets the variance in the second data set can be neglected, and we are back in the case considered above. On the other hand, if M/N

[9] 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.

converges to zero than the second data set cannot help us adjust for sample selection. Therefore, I consider only the case where the ratio M/N converges to a constant k.

We can obtain the estimate of θ by using the additional moments and standard GMM results. With out loss of generality, and in order to facilitate comparison with the previous exposition, I assume that M/N is exactly equal to some integer k. Let $\tilde{z}_i$ consist of $\{z_i, h_{n1}, ..., h_{ik}\}$. We can think of having N observations $\tilde{z}$. Using the logistic selection probability the estimating equations are

$$0 = \sum_{i=1}^{N} \begin{pmatrix} \psi(z_i, \theta_1) * \tau\left(1 + e^{h(z_i)'\theta_2}\right) \\ \left(\bar{h}(z_i) - \hat{h}\right) * \tau\left(1 + e^{h(z_i)'\theta_2}\right) \\ 1 - \tau\left(1 + e^{h(z_i)'\theta_2}\right) \\ 1/k \sum_{j=1}^{k} (h_{ij} - \hat{h}) \end{pmatrix} \qquad \textbf{(9)}$$

Solving this leads to $\hat{h} = 1/M \sum_{i=1}^{N} \sum_{j=1}^{k} h_{ij}$, while 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 the conditions of Proposition 2 hold, then the estimator* θ_1^W *for* θ_1^* *has the following asymptotic properties:*

$$\sqrt{N}(\hat{\theta}_1^W - \theta_1^*) \rightharpoonup N\left(0, E_s[\partial\tilde{\psi}/\partial\theta_1']^{-1}\left(E_s[\tilde{\psi}\tilde{\psi}'] - E_s[\tilde{\psi}\tilde{h}']E_s[\tilde{h}\tilde{h}']^{-1}E_s[\tilde{h}\tilde{\psi}']\right)E_s[\partial\tilde{\psi}/\partial\theta_1']^{-1} + V\frac{\Delta_h}{k}V'\right)$$

where:

$$V = E_s[\partial\tilde{\psi}/\partial\theta_1']^{-1}E_s[\tilde{\psi}\tilde{h}']E_s[\tilde{h}\tilde{h}']^{-1}E_s[w]I_R + E_s[\partial\tilde{\psi}/\partial\theta_1']^{-1}E_s[\tilde{\psi}\theta_2'w],$$

$$\tilde{\psi}(Z,\theta) = \psi(Z,\theta_1)\cdot w(Z,\theta_2) \quad , \quad \tilde{h}(Z,\theta_2) = \begin{pmatrix} \left(\bar{h}(Z) - \hat{h}\right)\cdot w(Z,\theta_2) \\ w(Z,\theta_2) - 1 \end{pmatrix} \quad , \quad w(Z,\theta_2) = \tau\left(1 + e^{(\bar{h}(Z) - \hat{h})'\theta_2}\right)$$

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

Proof: The same as the proof of Proposition 2, for details see Hellerstein and Imbens (1999).

3. AN EMPIRICAL APPLICATION

In this section I apply the method described in the previous sections to the Dutch Transportation Panel used by Ridder (1992). I start by presenting the data and initial analysis, and continue to present the results based on the procedure proposed here.

3.1 *The Data and Preliminary Analysis*

The data I use is taken from the Dutch Transportation Panel (DTP). The purpose of this panel 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, more than 11 years old, 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 she returns. Several questions were asked about each trip, but this information was not used below. 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, which 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 therefore were somewhat different than the March interviews. I follow Ridder and examine only the March interviews (thus also avoiding any seasonal effects). The original panel suffered from heavy attrition, as seen in Table 1. In order to keep the number of participants constant additional refreshment samples were taken from the population.[10]

[10]For the rest of this paper I will assume 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 in order to compensate for the heavier attrition in some strata. The methods used here can easily deal with this case, however, for simplicity of presentation I ignore this aspect.

For the purpose of demonstrating the method proposed here I concentrate on explaining the number of trips taken as a function of household characteristics.

The average number of trips in different sub-samples of the panel is given in Table 1. By examining the bottom row we conclude that there is virtually no change in the average number of trips over the different waves. However, if we examine any one of the other rows we find 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 totally offset each other.

The data also contain information on various explanatory variables, described in the Appendix. From the summary statistics we see that the distribution of the variables is different 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 non-random attrition there are (non-random) differences in the distribution of the unobserved determinants of the total number of trips. Non-random 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 is part of the pattern due to a real change in mobility.

In order to answer this question I compute a series of regressions presented in Table 2. The following conclusions can be reached from the results in the table. First, if we assume no correlation between the unobserved determinates of the number of trips and the independent variables then we can conclude that there was no drop in mobility. We can see this 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.[11]

Second, using the panel structure we can examine the assumption that unobserved household-specific effects are not correlated with the independent variables. 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 and 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 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 towards a large downward trend in mobility. If we were confident that either the unbalanced or balanced panels were representative of the population we could conclude that there was a downward trend in mobility. However, since we have already concluded that the attrition from the sample is non-random we can not reach this conclusion based on the results presented in any of the columns of Table 2.

Therefore, in order to answer 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 below. An additional estimator that could potentially deal with these issues is the one suggested by Hausman and Wise (1979). Ridder explores this estimator and finds that it fails to alert of non-random attrition, hence, also fails to treat it.[12]

3.2 *Results Using the Proposed Procedure*

---

[11]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, we might conclude that there was a drop in mobility during the second period, but not overall.

[12]Ridder attributes this failure to an implicit restriction, which forces the covariance of the individual effects in the selection and regression equations to have the same sign as the covariance of the random shocks in the two equations (see section 5 in Ridder's paper for details).

In order to evaluate the performance of the procedure proposed in this paper I examine two measures. First, I study out-of-sample prediction of the model. This 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, I compute estimates of the regression coefficients, 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.

In order to test the out-of-sample predictive power of the methods, I compute weights by solving equation (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: First, using moments on only the explanatory variables. This assumes a (particular) ignorable, conditional on observable variables, model of selection. If selection is a linear function of only the explanatory variables then these weights should fully control for selection. The second set of weights were computed using only the first moments of the dependent variable (TOTRIP). Finally, all the variables were used. In all cases I used only the first moments computed from the first and last waves.

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. Since 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 fully control for selection should render the differences, between the moments of Table 3 and the appropriate moments in Table A2, as statistically insignificant. The logic behind this is the same as that of the usual test of over-identification.

The results in Table 3 lead to the following conclusions. First, the weighted samples are more representative of the refreshment population, and therefore the population of interest. For all three selection models the fit is much better for the second wave than the third wave. The ignorable selection model, which uses only the moments of the explanatory variables, is quite strongly rejected. The third model that uses both the dependant 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 non-ignorable models of selection. Under the model where 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 dependant 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) finds that the model of Hausman and Wise (1978), which makes these assumptions about the individual-specific effects, does not fit this data set. In order to deal with the poor fit of the third wave moments I allow the selection probability 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 and a random-effects estimator is computed. The models differ in the 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. Since the analysis of the results in Table 3 suggests that this selection model is not fully capturing 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 OLS results from the repeated cross-section and the (un-weighetd) balanced panel results. Finally, and most importantly, even after controlling for selection 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 I also repeated the above analysis using a Poisson model for the number of trips. The only change is the moments are now non-linear in the parameters. The qualitative effects are similar to the above. 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 the second selection model, the probability of taking a trip is reduced by about 5 percent in the second and fourth waves, relative to the first wave. While this probability is reduced by 15 percent in the third wave. Since the average number of trips is roughly fifty this is close to what the results of Table 4 imply.

4. Final Remarks

This paper proposes a weighting method that takes advantage of additional information to treat sample selection bias. I exploit moments that are available from other sources 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 I compute the selection probability, which is used to inflate the data. The estimator can deal with ignorable as well as non-ignorable

selection mechanisms.

I outlined a few applications where I believe these additional moments are available. One application was presented in detail. This application is characterized by refreshment samples from the target population, which were taken in order to deal with attrition of the original sample. This is not typical of economic data. But maybe it should be. Maybe 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 paper. However, some idea can be obtained by building on other work. Ridder (1992) uses the same data as above to report results from a control function approach, which were discussed above. The conclusions regarding mobility are similar to those obtained for the above analysis. The same is true for a different approach, taken by Hirano et al. (1998), which involves imputing the missing data. The method proposed here has one advantage over these two alternatives: it is much easier to implement. Computing the weights involves solving a simple system of equations, or alternatively a linear programming problem. It only has to be done once, and not repeatedly each time a new specification of the main equation is examined. The actual analysis can be performed with the weighted data, applying standard methods and using standard software packets. The last class of alternative methods are matching methods, in the spirit of Ahn and Powell (1993). This method does not use the auxiliary information, discussed in this paper, but can be extended to do so. Such an extension would require different assumptions than those made in this paper, and is an interesting topic for future work.

Many public use data sets are accompanied by weights which 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 (since 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 to estimating the weights. Hirano, Imbens and Ridder (2000) show 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 relate their result to the result in Wooldridge (1999), which shows that in the context of stratified sampling it is more efficient to use estimated weights rather than known sampling probabilities. The Monte Carlo results in Nevo (forthcoming) seem to suggest that a similar result might be applicable here.

REFERENCES

Ahn, H., and J. Powell (1993), "Semi-parametric Estimation of Censored Selection Models with a Non-parametric Selection Mechanism," *Journal of Econometrics*, 58, 3-29.

Angrist, J. (1995), "Conditioning on the Probability of Selection to Control Selection Bias," National Bureau of Economic Research, Technical Working Paper no. 181.

Back, K., and D. Brown (1993), "Implied Probabilities in GMM Estimators," *Econometrica*, 61(4), 971-6.

Cassel, C., C. Sarndal, and J. Wretman (1979), "Some Use of Statistical Models in Connection With the Non-Response Problem," *Symposium on Incomplete Data: Preliminary Proceedings.*

Cosslett, S. (1981), "Maximum Likelihood Estimator for Choice-Based Samples", *Econometrica*, 49(5),1289-316.

Griliches, Z., and J. Mairesse (1998), "Production Functions: The Search for Identification," in *Practicing Econometrics: Essays in Method and Application*, Cheltenham, UK: Elgar. Also in Steinar Strom (ed.), *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*, Economic Society Monograph Series, Cambridge University Press.

MaCurdy, T., T. Mroz and M. Gritz (1998), "An Evaluation of the National Longitudinal Survey on Youth,"*Journal of Human Resources*, 33(2),345-436.

Gottschalk, P., and R. Moffitt (1992), "Earnings and Wage Distributions in the NLS, CPS, and PSID," Part I of Final Report to the U.S. Department of Labor, May 1992.

Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-54.

Hausman, J. and D. Wise (1979), "Attrition Bias in Experimental and Panel Data: The Gray Income

Maintenance Experience," *Econometrica*, 47, 455-73.

Heckman, J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679-93.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 931-61.

Heckman, J. (1987), "Selection Bias and Self-Selection," in Eatwell, J., M. Milgate and P. Newman (eds.), *The New Palgrave : a dictionary of economics*, pp. 287-97, London : Macmillan Press Ltd.

Heckman, J. (1990), "Varieties of Selection Bias," *American Economic Review*, 80, 313-8.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66(5) 1017-98.

Heckman, J. and R. Robb (1986), "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in H. Wainer (ed.), *Drawing Inferences from Self-Selected Samples*, pp. 63-107, New-York: Springer-Verlag.

Hellerstein, J., and G. Imbens, (1999), "Imposing Moment Restrictions from Auxiliary Data by Weighting," *Review of Economics and Statistics*, 81(1), 1-14.

Hirano, K., G. Imbens, G. Ridder and D. Rubin (1998), "Combining Panel Data Sets with Attrition and Refreshment Samples," National Bureau of Economic Research, Technical Working Paper no. 230.

Hirano, K., G. Imbens, and G. Ridder (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," National Bureau of Economic Research, Technical Working Paper no. 251.

Imbens, G. (1997), "One-Step Estimators for Over-Identified Generalized Method of Moments Models," *Review of Economic Studies*, 64, 359-83.

Imbens, G., and T. Lancaster (1994), "Combining Micro and Macro Data in Microeconomic Models,"

Review of Economic Studies, 61, 655-80.

Imbens, G., R. Spady and P. Johnson(1998), "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 61, 655-80.

Kitamura, Y., and M. Stutzer (1997), "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861-74.

Koul, H., V. Susarla and J. Van Ryzin (1981), "Regression Analysis with Randomly Right-Censored Data," *The Ann*als *of Statistics*, 9(6), 1276-88.

Kyriazidou, E. (1997), "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335-64.

Little, R. and D. B. Rubin (1987), *Statistical Analysis with Missing Data*, John Wiley & Sons.

Little, R. and M. Wu (1991), "Models for Contingency Tables with Known Margins When Target and Sampled Populations Differ," *Journal of the American Statistical Association*, 86 (413), 87-95.

Manski, C. F. (1994), "The Selection Problem," in C. Sims (ed.), *Advances in Econometrics: Sixth World Congress*, Vol. I.

Manski, C. F. and Lerman, S. (1977), "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica*, 45(8), 1977-88.

Nevo, A. (forthcoming), "Sample Selection and Information-Theoretic Alternatives to GMM," forthcoming *Journal of Econometrics.*

Newey, W. (1984), "A Method of Moments Interpretation of Sequential Estimators," Economics Letters, 14, 201-206.

Newey, W., and D. McFadden (1994), "Estimation in Large Samples," in McFadden, D. and R. Engle (eds.), *The Handbook of Econometrics*, Vol. 4.

Newey, W., J. Powell, and J. Walker (1990), “Semiparmetric Estimation of Selection Models: Some Empirical Results,” *American Economic Review*, 80, 324-28.

Oh, H., and F. Scheuren (1983) "Weighting Adjustments for Unit Non-response," in Madow, W., I. Olkin, and D. Rubin (eds.), *Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography*. New York; Academic Press.

Qin, J., and J. Lawless (1994) “Generalized Estimating Equations,” *Annals of Statistics*, 22, 300-25.

Olley, S. and A. Pakes (1996) “The Dynamics of Productivity in the Telecommunications Equipment Industry,” *Econometrica*, 64(6), 1263-97.

Pagan, A. (1986), “Two Stage and Related Estimators and Their Applications,” *Review of Economic Studies*, 53, 517-538.

Ridder, G. (1992), “An Empirical Evaluation of Some Models for Non- Random Attrition in Panel Data,” *Structural Change and Economic Dynamics*, 3(2), 337-55.

Wissen, van L., and H. Meurs (1990), “The Dutch Mobility Panel: Experiences Evaluation,” *Transportation*, 16, 99-119.

Wooldridge, J. (1999), “Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples,” *Econometrica* 67(6), 1385-1406.

TABLE 1
NUMBER OF HOUSEHOLDS AND AVERAGE NUMBER OF TRIPS
BY WAVE AND WAVE OF ATTRITION

	Wave							
Drops out in wave:	1		2		3		4	
1	731	45.4 (33.2)	–	–	–	–	–	–
2	178	57.7 (33.5)	178	48.2 (27.1)	–	–	–	–
3	185	62.2 (37.2)	185	54.9 (32.4)	185	52.3 (28.5)	–	–
4	666	62.8 (34.6)	666	56.7 (31.0)	666	55.9 (30.8)	666	55.1 (31.1)
Total	1760	55.0 (35.1)	1029	54.9 (30.8)	851	55.1 (30.3)	666	55.1 (31.0)

For each wave the left column presents the number of households, while the left column presents the average number of trips and the standard deviation in parentheses.

TABLE 2
REGRESSION RESULTS

	Repeated CS		Unbalanced Panel			Balanced Panel		
Variable	OLS		Total	Within	RE	Total	Within	RE
Constant	55.02 (0.80)	1.91 (1.39)	3.82 (1.35)	–	4.66 (1.49)	7.35 (1.77)	–	8.93 (2.02)
WAVE 2	-3.72 (1.54)	-2.92 (0.90)	-4.13 (0.79)	-6.57 (0.61)	-5.54 (0.57)	-6.25 (1.07)	-5.87 (0.74)	-6.07 (0.73)
WAVE 3	-8.39 (1.68)	-0.64 (0.98)	-5.29 (0.84)	-8.18 (0.67)	-7.03 (0.61)	-7.81 (1.07)	-7.62 (0.76)	-7.72 (0.73)
WAVE 4	-1.72 (1.66)	-1.14 (0.96)	-5.04 (0.92)	-8.75 (0.75)	-7.41 (0.68)	-8.55 (1.08)	-8.37 (0.80)	-8.45 (0.74)
Demographics included:	no	yes	yes	yes	yes	yes	yes	yes

Dependent variable is total number of trips. White-robust standard errors in parentheses. Except the first column, all regressions include as controls the demographic variables described in Table A1.

TABLE 3
WEIGHTED SAMPLE AVERAGES

	Second Wave			Third Wave		
Variable	(i)	(ii)	(iii)	(iv)	(v)	(vi)
TOTRIP	51.63	52.36	51.95	52.47	52.62	53.64
NPER	2.17	2.14	2.17	2.24	2.21	2.24
N1218	0.28	0.26	0.28	0.30	0.29	0.29
N1938	0.95	1.04	0.95	0.97	1.00	0.96
FAMT1	0.13	0.08	0.13	0.14	0.07	0.13
FAMT2	0.25	0.33	0.25	0.22	0.31	0.23
FAMT3	0.14	0.12	0.14	0.15	0.13	0.15
INC1	0.15	0.12	0.14	0.13	0.10	0.13
INC2	0.32	0.39	0.31	0.33	0.37	0.34
INC3	0.29	0.26	0.30	0.31	0.30	0.30
EDLO	0.39	0.35	0.38	0.40	0.35	0.40
EDHI	0.20	0.27	0.20	0.20	0.28	0.20
CITY	0.06	0.06	0.07	0.06	0.06	0.07
NCAR	0.83	0.87	0.83	0.81	0.87	0.82
NLIC	1.31	1.40	1.31	1.33	1.44	1.34

Weights are computed using:
In columns 1 and 4 first moments of explanatory variables in first and fourth waves.
In columns 2 and 5 first moments of TOTRIP in first and fourth waves.
In columns 3 and 6 first moments of TOTRIP and all the explanatory variables in first and fourth waves.

TABLE 4
WEIGHTED REGRESSION RESULTS

	Model 1				Model 2			
	Within		Random Effects		Within		Random Effects	
Variable	est	se	est	se	est	se	est	se
CONSTANT	–		5.17	1.85	–		8.11	1.77
WAVE 2	-2.12	0.69	-2.40	0.69	-3.42	0.72	-2.99	0.71
WAVE 3	-2.10	0.69	-2.11	0.69	-8.01	0.74	-6.85	0.72
WAVE 4	-1.77	0.71	-1.36	0.69	-3.41	0.77	-2.36	0.73

Dependent variable is total number of trips. In Model 1 the weights are computed as a function of the dependant and independent variables in first and fourth waves (as in columns 3 and 6 of Table 3). In Model 2 the weights are computed as a function of all variables in wave 3 and the dependant variable in the other waves. All regressions include as controls the demographic variables described in Table A1.

APPENDIX

This appendix describes, in Table A1, the variables available in the data and provides, in Table A2, their sample statistics in the different waves and sub-samples.

TABLE A1
THE EXPLANATORY VARIABLES

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

Source: Ridder (1992).

TABLE A2
SAMPLE AVERAGES OF VARIABLES

	First Wave		Second Wave			Third Wave			Fourth Wave	
Variable	Unbal	Bal	Unbal	Bal	Refr	Unbal	Bal	Refr	Bal	Refr
N=	1760	666	1029	666	656	851	666	516	666	535
TOTRIP	55.02	62.80	54.90	56.70	51.80	55.12	55.90	46.63	55.13	53.30
NPER	2.19	2.28	2.28	2.28	2.23	2.34	2.32	1.92	2.33	2.21
N1218	0.29	0.33	0.31	0.33	0.36	0.34	0.34	0.20	0.33	0.26
N1938	1.00	1.12	1.05	1.05	0.85	1.03	1.02	0.85	0.97	0.97
FAMT1	0.12	0.12	0.10	0.09	0.09	0.08	0.07	0.12	0.06	0.13
FAMT2	0.24	0.29	0.29	0.32	0.20	0.29	0.30	0.12	0.28	0.20
FAMT3	0.14	0.09	0.10	0.11	0.15	0.11	0.11	0.16	0.12	0.16
INC1	0.19	0.13	0.11	0.11	0.15	0.10	0.09	0.23	0.82	0.13
INC2	0.32	0.36	0.38	0.40	0.35	0.38	0.37	0.29	0.37	0.36
INC3	0.27	0.29	0.30	0.28	0.25	0.33	0.33	0.21	0.35	0.29
EDLO	0.44	0.37	0.35	0.35	0.47	0.35	0.35	0.37	0.34	0.41
EDHI	0.18	0.23	0.26	0.27	0.18	0.27	0.28	0.19	0.30	0.17
CITY	0.10	0.06	0.07	0.05	0.27	0.06	0.05	0.09	0.05	0.19
NCAR	0.85	0.89	0.92	0.90	0.84	0.92	0.90	0.71	0.90	0.79
NLIC	1.35	1.48	1.49	1.47	1.28	1.51	1.50	1.10	1.50	1.31

Columns labeled *Unbal*, *Bal*, and *Refr* present, respectively, averages for: the cross-section of the original panel, the balanced sun panel and the refreshment samples.