Weighting Methods for Choice Based Panels with Attrition

Ram M. Pendyala
Department of Civil Engineering
University of Southwestern Louisiana
Lafayette, LA 70504-2291

Ryuichi Kitamura
Institute of Transportation Studies
and
Department of Civil Engineering
University of California at Davis
Davis, CA 95616

Prepared for the First U.S. Conference on Panels in Transportation Planning
Lake Arrowhead, CA, October 24-27, 1992

Abstract

This paper develops a weighting method that can be applied to choice-based panel samples. In recent times, rising costs of surveys and the need to study infrequent travel choices have motivated the use of choice-based sampling procedures where sample entities are chosen based on the endogenous choice variable. As a choice-based sample is not representative of the population, suitable weights need to be developed for such samples before population inferences can be drawn. The issue is further complicated if a choice-based sampling technique is employed in a panel survey. Panel surveys allow the measurement of changes in behavior over time and therefore provide valuable information that conventional cross-sectional surveys do not. However, a choice-based panel sample needs to be treated for selective attrition as well. While past research has developed weights to treat for choice-based sampling and attrition separately, this study is the first attempt to account for both issues simultaneously. In this study, weights are developed for a choice-based panel sample from the Puget Sound Transportation Panel to obtain unbiased population estimates of transitions in mode choice. A simultaneous equation system in which attrition and mode choice are treated as discrete endogenous dependent variables is estimated to compute weights for each household in the sample that participated in both contacts of the survey.

1. Introduction

Choice-based samples are derived by selecting sample entities based on the values of the endogenous variable they exhibit. Such a sampling scheme is often preferred over a purely random sampling procedure when the study needs to include a sizable number of behavioral units exhibiting infrequent choices. The sample proportion of these infrequent phenomena is larger than the population proportion. In other words, the choice-based sample is not representative of the population. Drawing unbiased inferences regarding population behavior would require the treatment of choice-based samples using appropriate weighting methods. These methods have been developed previously [Cosslett, 1981; Lancaster and Imbens, 1990; Manski and Lerman, 1977; Manski and McFadden, 1981] and are reviewed in Amemiya [1985], and Ben-Akiva and Lerman (1985].

Choice-based sampling procedures may be employed in the conduct of panel studies. Panel surveys have been gaining increasing acceptance in transportation planning as they offer longitudinal information on the same behavioral units. This allows policy analyses and travel demand forecasting based on measured changes in behavior while controlling for unobserved individual factors that do not change over time (Kitamura, 1990]. These advantages are not offered by conventional cross-sectional surveys. However, panel studies invariably need to be treated for attrition, the non-random dropping out of survey participants over successive contacts (waves) of the survey. Appropriate weights are applied to the stayer sample (portion of the original sample that responds to all waves of the survey) to make it representative of the original sample and the population. The weights can be based on the probability of staying in successive waves of the survey, with those that have higher propensities to drop out receiving larger weights. This weighting method has been developed previously (Kitamura and Bovy, 1987] in an application to the Dutch National Mobility Panel Study [van Wissen and Meurs, 1989].

If a choice-based sampling procedure is employed in a panel study, the resulting data set could offer valuable information regarding the dynamics of infrequent choices that would not have been otherwise available. This motivates the examination of how choice-based panels, which would involve both endogenous sampling biases and attrition biases, need to be treated to be able to draw unbiased population inferences. This paper is aimed at fulfilling this objective by developing a weighting method that would jointly account for endogenous sampling and attrition biases.

The Puget Sound Transportation Panel [Murakami and Watterson, 1990] offers a unique opportunity to examine the relationship between attrition and the endogenous variable upon which sampling is based. This panel study involved the selection of participants based on their mode choice classification to ensure that a sizable number of households using mass transit was included in the overall sample.

In a previous study [Pendyala, et al., 1992], a joint weighting method was developed and applied to the Puget Sound Transportation Panel Data Set to generate population estimates of mode choice transitions. However, the empirical analysis was performed under the assumption that the choice behavior (on which sampling was based) was exogenous to attrition. Specifically, it was assumed that the error terms across the mode choice and attrition equations were independent, and mode choice was a pre-determined variable in the estimation of attrition probabilities. As a consequence, the study offered population estimates of mode choice transitions under highly restrictive assumptions.

The objective of this paper is to test the veracity of these assumptions by extending the previous study to treat mode choice as endogenous to attrition and examine the significance of the error covariance. This is accomplished by estimating a simultaneous equation system with individual specific error components. A logit formulation is adopted to estimate choice and attrition probabilities. Both an independent model system and a model system incorporating endogeneity are estimated and compared for the Puget Sound Transportation Panel. The methodology developed in this paper is applicable to any choice-based panel which may need the recognition of endogeneity. The estimation procedure is computationally tractable and provides a convenient measure of the significance of error covariance.

In the next section, literature pertaining to the development of choice-based sample weights is reviewed. This is followed by a description of modeling methods to develop a joint choice-based attrition weight while incorporating endogeneity of the choice variable. Section 4 describes the Puget Sound Transportation Panel and its sampling procedure. Section ***5*** develops weights for the Puget Sound Transportation Panel and provides results of the model estimation. Section 6 provides a discussion on the estimation of choice-based panel weights when the choice variable is endogenous to panel participation. Finally, Section 7 presents unweighted and weighted mode choice transitions and key conclusions.

2. Review of Choice-based Sampling

Choice-based sampling falls under the broader scheme of stratified sampling. In stratified sampling, the population is divided into groups according to a set of measured variables, and sample units are then drawn at random from each group. If the population is divided based on the endogenous variable of the study, then the division is referred to as endogenous stratification, and the resulting sample is a choice-based sample.

This section will review the mathematical formulations of weights for choice-based samples. The extension to incorporate attrition in the case of panels will be the focus of the next section. The discussion here closely follows that of Cosslett (1981], Manski and McFadden (1981], and Lancaster and Imbens [1990].

Let $\boldsymbol{C}$ represent a finite choice set consisting of $\boldsymbol{M}$ mutually exclusive discrete alternatives. Let $\boldsymbol{Z}$ represent the space of explanatory attributes characterizing the population. The population is contained in the product space $\boldsymbol{C X Z}$. Then, each sample unit can be described by a value for the choice variable, $j \in C$, and a vector of explanatory variables, $z \in Z$.

The joint probability density of choice $j \in C$ and $z \in Z$**,** is given by,

$$f(j,z|\theta) = p(z)P(j|z,\theta) \tag{1}$$

where

$f(j,z|\theta)$ = joint density function of (j,z) pairs in the population

$p(z)$ = marginal probability density of the distribution of attributes in the population

$$= \sum_{j \in C} f(j,z)$$

$P(j|z,\theta)$ = conditional probability of choice ***j*** given ***z***, the attribute vector, and

θ = the vector of underlying population parameters relating ***z*** and the probability of choice ***j***.

Various sampling schemes can be employed to choose observations of ***(j,z)*** pairs from the population space ***C X Z.*** When a purely simple random sample is selected, the likelihood of observing a ***(j,z)*** pair in the sample is given by the joint probability density of observing events $j \in C$ and $z \in Z$; that is,

$$L_r = f(j,z|\theta) \tag{2}$$

In the case of endogenous or choice-based sampling, the choice set ***C*** is partitioned into measurable subsets C_b, $b \in B$**,** where C_b refers to the ***bth*** subset and there are ***B*** such subsets. Each subset may contain a value or values of the choice variable ***C*** and is referred to as a sampling choice stratum. Then, the population may be considered to be made up of ***B*** strata, and the ***bth*** sampling stratum may be represented as $A_b = C_b X Z$. It is to be noted that sample choice strata may overlap, i.e., the same values for the choice variable may appear in several strata. Then, we can write,

$$\begin{aligned}\sum_{j,z \in C_b XZ} f(j,z|\theta) &= \sum_{j \in C_b} \sum_{z \in Z} P(j|z,\theta)p(z) \\ &= \sum_{j \in C_b} Q(j|\theta) \\ &= Q(b|\theta) \qquad (3)\end{aligned}$$

where $Q(j|\theta)$ is the marginal probability of choice ***j***, and $Q(b|\theta)$ is the marginal probability that $j \in C_b$**.** Then the likelihood of sampling a ***(j,z)*** pair is obtained through the occurrence of two events; the first that stratum A_b is chosen and the second that the pair (j,z) is sampled given that A_b is chosen. The likelihood may then be represented mathematically as follows:

$$L_c \quad = \quad H(b)\, P(j,z \in A_b | b,\theta)$$

$$= \quad \frac{H(b) f(j,z \mid \theta}{\sum_{j,z \in A_b} f(j,z \mid \theta)}$$

where $H(b)$ represents the probability of choosing stratum A_b. Substituting the expression from Equation 3 for the denominator above, the likelihood reduces to,

$$L_c \quad = \quad \frac{H(b) f(j,z \mid \theta)}{Q(b \mid \theta)} \qquad (4)$$

Equation 4 can be reduced to Equation 2 by multiplying it with $[(H(b)/Q(b|\theta)]^{-1}$. This represents a weight which, when applied to the choice-based sample, makes the likelihood of each sample unit equivalent to that of a pure simple random sample. The resulting weighted sample would be representative of the population from which it is drawn, similar to a pure random sample. When overlapping choice strata are present, the factor may be generalized as,

$$\omega(j) \quad = \quad [\sum_{j \in C_b, b \in B} \frac{H(b)}{Q(b \mid \theta)}]^{-1} \qquad (5)$$

where $\omega(j)$represents the weight applied to choice j.

As an illustration, let C= {car, bus, rail) be a choice set of modes available to a population. Let there be two strata, the first consisting of roadway modes and the second consisting of public transit, namely, C_1 = (car, bus) and C_2 = {bus, rail). The two strata overlap because the choice of bus is an element of both sets. Let a fraction $H(1)$ of the observations be drawn from C_1 and a fraction $H(2)$ be drawn from C_2. Then, there are three distinct weights to be calculated, one for each choice of mode. They are derived from Equation *5*.

$$\omega(car) \quad = \quad [\frac{H(1)}{Q(1)}]^{-1} \qquad (6)$$

$$\omega\boldsymbol{(rail)} \;=\; [\frac{\boldsymbol{H(2)}}{\boldsymbol{Q(2)}}]^{-1} \tag{7}$$

$$\omega\boldsymbol{(bus)} \;=\; [\frac{\boldsymbol{H(1)}}{\boldsymbol{Q(1)}}+\frac{\boldsymbol{H(2)}}{\boldsymbol{Q(2)}}]^{-1} \tag{8}$$

where ***Q(1)*** is the population proportion of car and bus users and ***Q(2)*** is the population proportion of bus and rail users.

3. Choice-based Panel Samples

Choice-based sampling in panel studies may take one of two forms. The first is referred to as stock sampling while the second is called flow sampling [Lancaster and Imbens, 1990]. Stock sampling involves the selection of sample units based on the endogenous variable value they exhibit at one time point. The sampling itself is based on a cross sectional observation of choices. Once the sample units are selected, they are repeatedly contacted and their behavior observed. In the case of flow sampling, sample selection is based on transitions in choices exhibited by the population. This sampling process is more complex as it requires the researcher to observe behavior at two time points before recruiting sample entities (in some cases, however, observable behavior may signify a transition between states; for example, an application for new utility service may indicate residential relocation). Choice strata are defined by changes, or the lack thereof, in values of the endogenous variable, and sample units selected randomly from these choice strata.

However, it is noted that adopting stock or flow sampling procedures only changes the definition of strata. The mathematical formulation of choice-based weights is not affected by the definition of strata. As such, the weights derived in the previous section are equally applicable to stock and flow samples as regards the treatment for endogenous sampling. Additional treatment is now needed to account for panel attrition, where sample units cease to participate in the survey in a non-random fashion over successive waves of the survey. This section documents in detail modeling issues and methodologies available for the derivation of joint choice-based attrition weights.

For convenience, let us consider a binary choice variable, i.e., $\boldsymbol{m}=\{0,1\}$. Then initial choice and attrition behavior may be represented by a simultaneous equation system as follows (subscript ***i*** to represent the individual is suppressed for notational convenience):

$$\boldsymbol{C^* = \theta' Z + \psi}$$

$$\boldsymbol{m} = \begin{cases} 1, & \text{if } \boldsymbol{C^*} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$\boldsymbol{A^* = \beta_b' X + \gamma m + \varepsilon}$$

$$\boldsymbol{w} = \begin{cases} 1, & \text{if } \boldsymbol{A^*} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where

$\boldsymbol{C^*}$ = latent variable underlying initial choice behavior
$\boldsymbol{m}$ = observed indicator of initial choice
$\boldsymbol{A^*}$ = latent variable underlying attrition behavior
$\boldsymbol{w}$ = observed indicator of attrition; 1 if continued to participate in panel and 0 otherwise
$\boldsymbol{\theta,\beta}$ = coefficient vectors
$\boldsymbol{\gamma}$ = scalar coefficient
$\boldsymbol{Z}$ = explanatory variables influencing choice behavior
$\boldsymbol{X}$ = explanatory variables influencing attrition behavior
$\boldsymbol{\psi\varepsilon}$ = random error terms

Vectors $\boldsymbol{Z}$ and $\boldsymbol{X}$ may contain common explanatory variables. The subscript $\boldsymbol{b}$ for the coefficient vector $\boldsymbol{\beta}$ allows different strata to exhibit different attrition behavior. In this system of equations, initial choice may be regarded exogenous to attrition if one or both of the following conditions apply:

(a) $\boldsymbol{C^*}$ is independent of $\boldsymbol{X}$ and $\boldsymbol{\gamma} = 0$; and
(b) Error terms are uncorrelated, i.e., $\mathbf{E}[\boldsymbol{\psi\varepsilon}] = 0$.

A discussion regarding the estimation of the simultaneous equation system under conditions (a) or (b) can be found in Pendyala, et al. [1992]. ~When either one of the above assumptions is acceptable, the system of equations can easily be estimated using single equation estimation procedures. In Pendyala, et al. [1992], condition (b) was assumed to be true, and M, the observed choice indicator, was considered exogenous to attrition. Then, a single-equation binary probit estimation yielded attrition probabilities that could be used to derive weights, i.e.,

$$\mathbf{P}(\boldsymbol{w}=\mathbf{1}|\boldsymbol{X,m}) = \boldsymbol{\Phi}(\boldsymbol{\beta' X + \gamma m}) \tag{11}$$

where the left hand side represents the probability of continuing to participate in the panel given the vector $\boldsymbol{X}$ and initial choice $\boldsymbol{m}$ and the right hand side is the standard normal cumulative distribution function evaluated at the parameter estimates.

This paper aims at relaxing the assumption that $\mathbf{E[\psi\varepsilon] = 0}$ and developing a methodology for examining endogeneity of mode choice in the estimation of choice-based panel weights. If $\mathbf{E[\psi\varepsilon] \neq 0}$, then the two-equation system should be estimated simultaneously via full-information maximum likelihood procedures.

If full-information maximum likelihood estimation of a bivariate probit with correlated error terms is to be performed, two-dimensional integrals of the bivariate normal density function need to be computed to evaluate the likelihood function. This procedure is computationally intensive and involves the application of approximate numerical integration techniques as the bivariate normal density does not have an analytical closed form integral.

On the other hand, if a limited information approach is adopted, parameters are estimated. one equation at a time with instrumental variables [see Maddala, 1983] or correction terms [see Heckman, 1976] introduced to account for error correlation. For linear systems, these techniques provide consistent, but inefficient estimates of parameters [Maddala, 1983; Nelson, 1984]. In a system of two binary choice equations as the one in this study, however, these approaches may lead to inconsistent estimates (numerical comparisons of alternative estimators are given in Kitamura, 1992). Predictions of attrition obtained from such model estimates may not be robust.

The full-information approach is the most desirable approach as it offers consistent and efficient estimates, while allowing the researcher to test the significance of error correlation across equations. Considering the difficulty in implementing a full-information bivariate probit estimation procedure, an alternative, error component method is developed in this paper which uses the logit formulation for estimating choice probabilities. This approach is presented below.

Let the error terms in Equations 9 and 10 be composed of an individual specific error component and a random component so that the equations may be written as,

$$\boldsymbol{C^* = \theta' Z + \kappa_1 \xi + u}$$

$$\boldsymbol{m} = \begin{cases} 1, & \text{if } \boldsymbol{C^*} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

$$\boldsymbol{A^* = \beta_b' X + \gamma m + \kappa_2 \xi + v}$$

$$\boldsymbol{w} = \begin{cases} 1, & \text{if } \boldsymbol{A^*} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where $\xi \sim \mathbf{N(0,1)}$. The error terms of Equations 9 and 10 are now replaced as $\boldsymbol{\psi = \kappa_1 \xi + u}$ and $\boldsymbol{\varepsilon = \kappa_2}\xi + v$. ξ is called the individual specific error term and represents unobserved individual attributes that affect behavior. Error correlation is measured through the factors κ_1 and κ_2 because

$$\mathbf{E[\psi\varepsilon] = E[(\kappa_1\xi + u)(\kappa_2\xi + v)] = \kappa_1 \kappa_2}$$

under the assumptions that ~ is standard normal, and E[uv] = E[ξu] = E[ξv] = 0. The above representation, then, offers a convenient mechanism to identify the presence of endogeneity in the estimation of weights.

To express the choice probabilities, distributional assumptions need to be made on random error components ***u*** and ***v***. The logistic function is more convenient than the probit because it has a closed form analytical solution for its integral, while the probit does not. As such, ***u*** and ***v*** are considered logistically distributed with variance $\pi^2/3$ (their components are independent and identically Gumbel distributed with variance $\pi^2/6$). Then the likelihood function, derived from the joint probability of observing a pair of values for ***m*** and ***w*** for an individual ***i***, is

$$P(w_i,m_i|X_i,Z_i) = \int_{-\infty}^{+\infty} \frac{(e^{\theta' Z_i+\kappa_1\xi_i})^{m_i}}{1+e^{\theta' Z_i+\kappa_1\xi_i}} \frac{(e^{\beta' X_i+\kappa_2\xi_i})^{w_i}}{1+e^{\beta' X_i+\kappa_2\xi_i}} \frac{e^{-\xi_i^2/2}}{\sqrt{2\pi}} d\xi_i$$

Now, let

$$TZ_i, = \exp\,(\theta' Z_i + \kappa_1\xi_i)$$
$$BX_i, = \exp\,(\beta' X_i + \kappa_2\xi_i)$$

$$I_i = \int_{-\infty}^{+\infty} \frac{(e^{\theta' Z_i+\kappa_1\xi_i})^{m_i}}{1+e^{\theta' Z_i+\kappa_1\xi_i}} \frac{(e^{\theta' Z_i+\kappa_2\xi_i})^{m_i}}{1+e^{\theta' Z_i+\kappa_2\xi_i}} \frac{e^{-\xi_i^2/2}}{\sqrt{2\pi}} d\xi_i$$

$$f(\xi_i) = \frac{e^{-\xi_i^2/2}}{\sqrt{2\pi}}$$

Then, the analytical first derivatives used for maximum likelihood estimation are,

$$\frac{\partial}{\partial\boldsymbol{\theta}} \text{In } P(w_i,m_i|X_i,Z_i) = \frac{1}{I_i}\int_{-\infty}^{+\infty} \frac{TZ_i}{(1+TZ_i)^2}\,[m_i(1+TZ_i) = TZ_i^{\,m_i}]\, \frac{(B\mathbf{X}_i)^{w_i}}{(1+BX_i)}\, \mathbf{Z}_i f(\xi_i)\, d\xi_i \qquad (15)$$

$$\frac{\partial}{\partial \beta} \mathbf{In}\ P(w_i,m_i|X_i,Z_i) = \frac{1}{I_i} \int_{-\infty}^{+\infty} \frac{(TZ_i)^{m_i}}{(1+TZ_i)} [w_i(1+BX_i) = BX_i^{\ w_i}]\ \frac{B\mathbf{X}_i}{(1+BX_i)^2}\ X_i f(\xi_i)\, d\xi_i \qquad (16)$$

$$\frac{\partial}{\partial \kappa_1} \mathbf{In}\ P(w_i,m_i|X_i,Z_i) = \frac{1}{I_i} \int_{-\infty}^{+\infty} \frac{TZ_i}{(1+TZ_i)^2} [m_i(1+TZ_i) = TZ_i^{\ m_i}]\ \frac{(B\mathbf{X}_i)^{w_i}}{(1+BX_i)}\ f(\xi_i)\, d\xi_i \qquad (17)$$

$$\frac{\partial}{\partial \kappa_2} \mathbf{In}\ P(w_i,m_i|X_i,Z_i) = \frac{1}{I_i} \int_{-\infty}^{+\infty} \frac{(TZ_i)^{m_i}}{(1+TZ_i)} [w_i(1+BX_i) = BX_i^{\ w_i}]\ \frac{B\mathbf{X}_i}{(1+BX_i)^2}\ f(\xi_i)\, d\xi_i \qquad (18)$$

Now, once parameter estimates are obtained, the next step is to derive weights for choice-based panels. We consider the joint probability of three events; the first that stratum ***b*** is chosen, the second that a ***(j,z)*** pair is chosen from stratum ***b,*** and the third that the unit ***(j,z)*** continues to participate in the panel (event represented by γ) . Then the likelihood that a particular sample unit continues to participate in the choice-based panel is given by,

$$\begin{aligned} L_{cp} &= P(b,j,z,\gamma|\theta,\beta \\ &\quad P(b)\ P(j,z|b,\theta,\beta\ P(\gamma|b,j,z,\theta,\beta) \\ &\quad H(b)\ P(j,z \in A_b \mid \theta,\beta)\ P(\gamma \mid b,j,z,\theta,\beta) \end{aligned} \qquad (19)$$

where ***P(γ|b,j,z,*θ,β)** is the probability of participating in successive waves of the panel given the sampling stratum, endogenous variable, exogenous variables, and parameters explaining choice behavior.

But, from Equation 4, we know that $L_c = H(b)\ P(j,z \in A_b \mid \theta,\beta)$. Substituting the expression for L_c into Equation 19, we obtain,

$$L_{cp} = \frac{H(b)\ P(j,z \mid \theta)\ P(\gamma \mid b,j,z,\theta,\beta)}{Q(b \mid \theta)} \qquad (20)$$

Given that consistent estimates of β can be obtained, weights that account for biases arising from endogenous sampling procedures in a panel survey with attrition can be developed as,

$$\omega(j) = [\sum_{j \in C_b, b \in B} \frac{H(b)\ \ P(\gamma \mid b,j,z,\theta,\beta)}{Q(b \mid \theta)}\ ]^{-1} \qquad (21)$$

4. The Puget Sound Transportation Panel

In 1989, the Puget Sound Council of Governments (now Puget Sound Regional Council) commenced the first general purpose transportation panel survey in the country. This survey is being conducted in cooperation with transit agencies of the region and is referred to as the Puget Sound Transportation Panel (hereafter PSTP). It has three main objectives [Murakami and Watterson, 1990]:

i) To be a metropolitan "current population survey" to track changes in employment, work characteristics, household composition, and vehicle ownership

ii) To monitor changes in travel behavior and responses to changes in the transportation environment

iii) To examine changes in attitudes and values as they affect mode choice and travel behavior

The sampling scheme in the PSTP was designed to obtain an enriched sample, which is a special case of the generalized choice-based sample described in Cosslett [1981]. It consists of a mixture of a random sample and a choice-based sample, the random and choice-based samples being collected from overlapping choice strata.

In the PSTP, the population was first exogenously stratified by county of residence. Telephone random digit dialing was employed to first collect a purely random sample of households from each county. This sample served as the primary source for households classified as single-occupant vehicle (SOV) and carpool households. Following this procedure, a choice-based sample of households classified as transit households was collected through special recruiting methods targeted towards households using public transit. These households were recruited through on-board solicitations of randomly selected bus routes, and by recontacting respondents of an earlier Seattle Metro Transit Survey.

The same households were then contacted in the next year (1990) for the second wave of the panel. The entire sample in the PSTP may be considered to be a "stratified enriched stock sample". This sample is then made up of three distinct mode (endogenous) strata:

i) *SOV Households:* Households in which no one made at least four one-way work trips by carpool or transit

ii) *Carpool Households:* Households in which at least one person made at least four one-way work trips by carpool (≥~ 2 licensed vehicle occupants)

iii) *Transit Households:* Households in which at least one person made at least four one-way work trips by public transit[1].

The special choice-based recruitment of transit households made the enriched sample have a larger proportion of transit households than in the population. As such, this sample is not representative of the population, and inferences regarding transitions in mode choice can be drawn only after applying weights.

In the survey, all persons aged 15 years or older in participating households were asked to fill out two day travel diaries recording characteristics of all trips made over the two day period. Table 1 shows the composition of the first wave and stayer samples over two waves of the survey conducted in 1989 and 1990. Initially, 5175 households were contacted for participation in the panel. Of these, 2944 households agreed to participate and were sent survey instruments. In the first wave of data collection, which took place from September through December 1989, 1713 households returned survey instruments, of which 1682 offered complete information with no missing data.

Different monetary incentives were offered to participating households. Households that received $1 per person together with the instrument showed the highest response rate followed by those which were promised $10 per household on return of completed instruments. Households provided with no incentive showed the lowest response rate (see Murakami and Watterson, 1990 for details). Accounting for any self-selectivity bias arising from different response rates by incentive scheme is, however, beyond the scope of this paper.

The first wave of the travel survey was followed by an attitudinal survey in February of 1990. The panel participants were contacted again during the summer of 1990 to inform them of the second Wave of the panel survey. The second wave was administered in the Fall of 1990 and included refreshment households that were added to reflect changes in population characteristics and to compensate for possible attrition.

1391 households returned travel instruments in the second wave also. Of these, 1330 households offered complete information with no missing data. This sample constitutes the stayer sample from which mode choice transitions can be derived. Its composition is also shown in Table 1.

[1] If a household met both carpool and transit criteria, it was classified as a transit household as they were more difficult to recruit.

Table 1

Composition of First Wave and Stayer Samples by Mode Choice

Recruitment Method		Mode			Total~Sample Size
		SOV	Carpool	Transit	
Tele-RDD	First Wave	1132	192	222	1546
	Stayer	886	136	173	1195
On-Bus	First Wave	0	0	75	75
	Stayer	0	0	57	57
Metro Seg.	First Wave	*5*	1	44	50
	Stayer	4	1	39	44
MetroR/NR	FirstWave	1	0	41	42
	Stayer	0	0	34	34
Total	First Wave	1138	193	382	1713
	Stayer	890	137	303	1330

TeIe-RDD: Telephone random digit dialing
On-Bus: On-bus solicitation of volunteer participants
Metro Seg.: volunteers from the Metro Market Segmentation Study
Metro RINR: volunteers from the Metro Rider/Non Rider Survey

5. Derivation of Weights for PSTP

In the Puget Sound Transportation Panel, the sampling unit was the household. The development of weights in this paper will be performed at the household level for this reason. In this section, weights are derived first for stratified choice-based sampling and then combined with attrition weights to develop the joint weight.

The exogenous stratification based on county of residence may be treated as per Kish 11965]. In this method, the disproportionate sample proportions are weighted such that the population proportions are reflected in the sample. As the PSTP was exogenously stratified by county of residence, population figures for these counties were collected and tabulated. Table 2 provides sample and population proportions for different counties of residence. In turn, these proportions can be used to compute exogenous stratification weights. For example, the weight applied to households residing in King county is,

$$\boldsymbol{\omega(\text{King}) = [41.4/57.9]^{-1} = 1.399}$$

Table 2

Households by County of Residence

County	Survey Sample		Population (1989)		Weight
	N	%	N	%	
King	709	41.4	601,960	57.9	1.399
Kitsap	206	12.0	66,920	6.4	0.535
Pierce	363	21.2	208,981	20.1	0.949
Snohomish	435	25.4	161,798	15.6	0.613
Total	1,713	100.0	1,039,659	100.0	

The PSTP enriched sample consists of a purely random sample combined with a choice-based sample of transit households. As such, mode choice "transit" is a member of two strata, while "SOV" and "Carpool" are members of only one strata. These strata can be defined as: C_1 = {SOV, Carpool, Transit} and C_2 = {Transit}. This definition of strata implies that there will be one weight applicable to SOV and carpool households, and a second weight applicable to transit households.

Let a fraction $H(1)$ of the sample be drawn from C_1 and a fraction $H(2)$ be drawn from C_2. Then, from the discussion of Section 3, the weights may be defined as follows:

$$\omega(SOV \ \& \ Carpool) = [H(1)/Q(1)]^{-1}$$

$$= [H(1)]^{-1}$$

because $Q(1)$, the population proportion of SOV, carpool, and transit households is equal to one (under the assumption that C_1, represents the universal choice set). From Table 1, it can be seen that the number of households collected from C_1 is 1546. The total number of households is 1713. Then,

$$\omega(SOV \ \& \ Carpool) = [1546/1713]^{-1} = 1.108$$

For transit households,

$$\omega(transit) = [\frac{H(1)}{Q(1)} + \frac{H(2)}{Q(2)}]^{-1}$$

where Q(1) = 1. Now H(2) represents the fraction of the total sample collected from ***C~.*** From Table 1, it can be seen that the total number of transit households collected is 382, while the total sample size is 1713. Then,

$$\boldsymbol{H}(\mathbf{2}) \;=\; 382/1713 = 0.2230$$

An estimate of ***Q*(2)** can be obtained if it is assumed that the sample collected purely randomly is representative of the population. That is, the fraction of transit households in the population is reflected in the portion of the sample collected through telephone random digit dialing. Table 1 shows that the number of transit households in the random portion of the sample is 222 while the total number of households collected through random digit dialing is 1546. Then,

$$\boldsymbol{Q}(\mathbf{2}) \;=\; 222/1546 = 0.\ 1436$$

The weight to be applied to transit households is, then,

$$\boldsymbol{\omega(transit)} \;=\; \mathbf{[1.108 \;+\; \frac{0.223}{0.1436}\,]^{-1} \;=\; 0.4073} \qquad (23)$$

These weights can be applied to transition tables of mode choice and demographics to account for biases introduced by endogenous sampling procedures. The application of the exogenous weight to account for stratification by county of residence and the choice-based sampling weight resulted in a weighted sample as shown in Table 3.

The unweighted stayer sample has 1330 households of which only 67% are SOV households while 23% are transit households. After the application of weights to account for choice-based sampling, the weighted sample shows a 79% proportion of SOV households and only a 9% proportion of transit households. The proportion of carpool households did not change appreciably, indicating that the sample proportion of carpool households nearly replicates that of the population. The overall weighted sample size is found to be 1574. The next step involves combining the choice-based sampling weight with the attrition weight. This is done next through the estimation of a simultaneous equation model system to compute attrition probabilities.

Table 3

Unweighted and Weighted Sample
(Accounting for Choice-based Sampling only)

Mode Choice	Unweighted		Weighted	
	N	%	N	%
SOV	890	67	1238	79
Carpool	137	10	193	12
Transit	303	23	143	9
Total	1330	100	1574	100

Mode choice and attrition behavior are modeled as per Equations 12 and 13 using the maximum likelihood approach outlined in Equations 14 through 18. The model system was estimated on 1682 first wave respondent households for which complete data was available. The percent attrition in the sample by mode choice is shown in Table 4. There are 1330 stayers and 352 leavers. Transit households showed the lowest attrition rate, presumably because they were specially recruited through choice-based means. Carpool households showed a larger attrition rate and this is partially attributable to the household dynamics that these households experienced [Murakami and Watterson, 1990].

Table 4

Household Attrition by Mode Choice

Mode Choice	Stayers	Leavers	Total	% Attrition
SOV	890	226	1116	20.3
Carpool	137	54	191	28.3
Transit	303	72	375	19.2
Total	1330	352	1682	20.9

Results of the simultaneous equation estimation effort are shown in Table *5.* The variance of ξ is exogenously imposed to be equal to one. The first portion corresponds to the mode choice model of Equation 12, while the latter portion corresponds to the attrition model.

Table *5*

Mode Choice and Attrition Model

Variable (Z, X)	Parameter Estimate (θ, β)	t-statistic
Mode Choice Model (Z. θ)		
Constant	-3.939	-7.52
ONECAR	2.252	5.50
TWOCAR	2.188	5.38
MULTICAR	2.066	5.01
LOTOFCARS	2.348	7.20
YEARHOME	0.129	3.01
LOINCOME	0.280	1.70
HIGHINCOME	-0.294	-2.10
BUSDIST	-0.081	0.70
Attrition Model (X, β)		
Constant	1.062	2.73
ONECAR	0.876	2.65
TWOCAR	1.200	3.43
MULTICAR	1.270	3.41
NWORKERS	0.222	2.26
YEARHOME	0.181	3.37
LOINCOME	-0.361	-2.12
HIGHINCOME	-0.229	-1.37
SNOLADULT	-0.744	-2.40
YNCIADULTS	-0.946	-3.72
MIDADULTS	-0.372	2.12
HHLDSIZE	0.285	-4.35
TELE-RDD	-0.622	-2.41
SOV-MODE	0.103	0.47
κ_1	-0.044	-0.05
κ_2	0.103	0.06
L(0) = -2331.75	-2[L(0)-L(θ, β)]= 2272.37 with 25 d.f.	
L(C) = -1935.80	-2[L(C)-L(θ, β)]= 1480.47 with 22 d.f.	
L(θ, β) = -1935.80	Number of cases=1682	ρ^2 = 1-L(θ, β)/L(0) = 0.49

Table 5 (continued)

Description of Variables

Variable	Definition
ONECAR	Dummy variable = 1 if household owns one car; 0 otherwise
TWOCAR	Dummy variable = 1 if household owns two cars; 0 otherwise
MULTICAR	Dummy variable = 1 if household owns three cars; 0 otherwise
LOTOFCARS	Dummy variable = 1 if household owns more than three cars; 0 otherwise
NWORKERS	Number of employed persons in household
YEARHOME	Number of years in current residence
LOINCOME	Dummy variable=1 if annual household income ≤ $15,000
HIGHINCOME	Dummy variable= 1 if annual household income > $50,000
SNGLADULT	Dummy variable= 1 if household has only one adult less than 35 years and no children; 0 otherwise
YNGADULTS	Dummy variable= 1 if household has two or more adults less than 35 years and no children; 0 otherwise
MIDADULTS	Dummy variable= 1 if household has two or more adults aged 35-64 years and no children; 0 otherwise
HHLDSIZE	Household size
TELE-RDD	Dummy variable= 1 if household recruited by telephone random digit dialing
SOV-MODE	Dummy variable= 1 if household is an SOV household
BUSDIST	Dummy variable= 1 if nearest bus stop is within 1/4th mile of household
Mode	Binary Choice Dependent Variable= 1 if household is an SOV household
Attrition	Binary Choice Dependent Variable= 1 if household continues to participate in second wave of panel

The results of the model estimation provide clear indications that, in the case of the Puget Sound Transportation Panel, mode choice is not endogenous to attrition. That is, panel participation and the choice process on which endogenous sampling was based can be estimated as independent systems while treating mode choice as exogenous to attrition. This implies that the findings reported in Pendyala, et al. [1992] are based on valid assumptions, i.e., the error covariance across equations is not significant and mode choice is exogenous to attrition.

The above conclusions can be deduced through an examination of the factors, x_1 and x~, which represent the error correlation. From Table *5,* it can be seen that both of these parameters are not statistically significant (i.e., different than zero at the *5%* level). This implies that the representation of unobserved individual attributes in the model system through the incorporation of error components is not necessary. In other words, the two choice processes may be estimated as two independent binary logit models. If, on the other hand,; and x_2 had been significant, then we would have concluded that the initial choice is endogenous to the estimation of choice-based panel attrition weights.

The model system provided consistent and expected indications regarding the signs and magnitudes of coefficient estimates. In the mode choice model, car ownership positively contributes to a household being classified as an SOV household. However, the magnitudes of the coefficients do not appreciably change among car ownership levels, except for no-car ownership which is excluded from the model and whose coefficient is zero. This is likely to be a manifestation of the increased number of licensed drivers in the household owning more cars, making the car availability per driver similar across different levels. One surprising **indication** is that the dummy variable associated with high income households recorded a negative **coefficient.** The distance from the bus stop does not have a significant affect on mode choice to work for household members. The small negative sign, however, does indicate that non-SOV households tended to have bus stops within a quarter mile.

With regard to the attrition model, car ownership, employment, and the term of residence positively influenced households to stay in the panel and respond in the second wave as well. However, low income households, and households with young and middle age adults with no children tended to leave the panel. Larger household sizes also contributed to dropping out of the panel. Households recruited by telephone random digit dialing are significantly more likely to leave the panel. This is very consistent with the fact that households collected by choice-based methods would tend to continue participation. The mode choice was not significantly affecting attrition, presumably because most of its effect was already captured by the TELERDD variable (most households recruited by random digit dialing are SOV or carpool households which had higher attrition rates; see Table 3).

In order to examine the effect of the exclusion of the individual specific error components on model parameter estimates, an independent set of logits was estimated in which mode choice was considered exogenous to attrition. Results of this estimation are summarized in Table 6.

Table 6

Independent Mode Choice and Attrition Model

Variable (Z, X)	**Parameter Estimate (θ, β)**	**t-statistic**
Mode Choice Model (Z. θ)		
Constant	-3.934	-7.58
ONECAR	2.250	5.53
TWOCAR	2.186	5.41
MULTICAR	2.063	5.03
LOTOFCARS	2.346	7.28
YEARHOME	0.128	3.01
LOINCOME	0.279	1.70
HIGHINCOME	-0.294	-2.11
BUSDIST	-0.082	-0.70
Attrition Model (X, β)		
Constant	1.061	2.81
ONECAR	0.883	2.74
TWOCAR	1.210	3.57
MULTICAR	1.282	3.54
NWORKERS	0.219	2.29
YEARHOME	0.131	3.42
LOINCOME	0.361	-2.17
HIGHINCOME	-0.231	-1.40
SNOLADULT	-0.745	-2.46
YNCIADULTS	-0.946	-3.86
MIDADULTS	-0.372	-2.14
HHLDSIZE	-0.285	-4.61
TELE-RDD	-0.618	-2.43
SOV-MODE	0.090	0.61
L(0) = -2331.75	-2[L(0)-L(θ, β)]= 2272.37 with 23 d.f.	
L(C) = -1935.80	-2[L(C)-L(θ, β)]= 1480.47 with 22 d.f.	
L(θ, β) = -1935.80	Number of cases=1682	ρ^2 = 1-L(θ, β)/L(0) = 0.49

A comparison of Tables *5* and 6 corroborates the conclusion presented earlier that, in the case of the PSTP, mode choice is not endogenous to attrition. The model coefficients, t-statistics, and goodness-of-fit measures are found to be nearly identical across the two model estimations. Under these conditions, panel participation probability may be computed for household ***i*** as**:**

$$\boldsymbol{P_1(\gamma|b,j,z,\theta,\beta)} = \frac{e^{\beta' X_i}}{1+e^{\beta' X_i}} \qquad (24)$$

This can be combined with weights developed in Table 1, and Equations 22 and 23 to compute overall choice-based panel weights for each household. For example, a carpool household from Pierce county would be weighted thus:

$$\boldsymbol{\omega(\textit{Pierce County, Carpool}) = 0.949\ X\ 1.108\ X\ [\frac{e^{\beta' X}}{1+e^{\beta' X}}]^{-1}} \qquad (25)$$

After the application of the joint weights similar to that shown in Equation *25,* the weighted sample was found to be as in Table 7.

Table 7

Unweighted and Weighted Sample
(Accounting for Choice-based Sampling and Attrition Biases)

Mode Choice	Unweighted		Weighted	
	N	%	N	%
SOV	890	67	1289	78
Carpool	137	10	200	12
Transit	303	23	161	10
Total	1330	100	1650	100

The weighted stayer sample now has a total sample size of 1650 of which 78% are SOV households. Only 10% are transit households. it is noteworthy that the weighted sample proportions by mode choice are very similar to those in Table 3 where the weighted sample was adjusted for choice-based sampling. The additional weighting applied through the accounting for attrition merely increases the overall sample size without affecting the sample proportions of mode choice. This is presumably because the attrition rates are not very different across household mode strata. The application of the joint weighting procedure produced a total sample size that is close to the original first wave sample of 1682 (with complete information).

6. Discussion on Endogeneity of Weights

Even though the Puget Sound Transportation Panel did not show evidence of endogeneity of mode choice, it is possible for other choice-based panels to have the choice variable endogenous to panel participation. In such cases also, the formulation of weights and the modeling methodology developed in this paper are applicable. The only change would be with regard to the computation of attrition probabilities.

Suppose the estimation of a two-factor error component simultaneous equation system yielded statistically significant values for the factors, $\boldsymbol{\kappa_1}$ and $\boldsymbol{\kappa_2}$ (the error covariance). This means that there are unobserved individual attributes which make mode choice endogenous to attrition. In such a case, the attrition probability would be computed conditionally given the mode choice parameter vector, $\boldsymbol{\theta}$. In turn, this implies that the choice-based panel weights are computed conditional on initial choice.

When the errors are correlated, we need to modify the computation of the panel participation probability. The density function for the error component becomes conditional on the initial mode choice (i.e., being SOV or non-SOV). Then, the bounds for the error component would be dictated by whether ***C**** in Equation 12 takes on a positive or negative value. In turn, this means that the random error component would also influence the bounds for the individual specific error component and needs to be integrated out. This produces a double integral or convolution, which becomes quite complex. The probability of panel participation would be computed for SOV household ***i*** as,

$$\boldsymbol{P_i(\gamma|m=1,z,x,\theta,\beta)} = \int_{-\infty}^{+\infty} \int_{-\theta' Z_i - u_i}^{+\infty} \frac{(e^{\beta' X_i + \kappa_2 \xi_i})^{m_i}}{1+e^{\beta' X_i + \kappa_2 \xi_i}} \frac{e^{-\xi_i^2/2}}{\sqrt{2\pi}} \frac{-e^{u_i}}{(1+e^{u_i})^2} \, d\xi_i \, du_i \qquad (26)$$

and for non-SOV household ***i*** as,

$$\boldsymbol{P_i(\gamma|m=0,z,x,\theta,\beta)} = \int_{-\infty}^{+\infty} \int_{-\infty}^{-\theta' Z_i - u_i} \frac{(e^{\beta' X_i + \kappa_2 \xi_i})^{m_i}}{1+e^{\beta' X_i + \kappa_2 \xi_i}} \frac{e^{-\xi_i^2/2}}{\sqrt{2\pi}} \frac{-e^{u_i}}{(1+e^{u_i})^2} \, d\xi_i \, du_i \qquad (27)$$

These expressions may be substituted in Equation 21 to obtain appropriate weights in the presence of endogeneity.

7. Conclusions

This paper presents a methodology to develop weights for choice-based panel samples where the choice variable may be endogenous to panel participation. The methodology adopted in this paper allows for convenient testing of endogeneity while recognizing the simultaneous nature of the choice processes.

In the case of the Puget Sound Transportation Panel, endogeneity of mode choice was found to be insignificant. As such, the independent logit models of attrition and mode choice could be used to appropriately weight transition tables, A person based mode choice transition table is presented in Table 8 with unweighted and weighted values.

Table 8

Person Mode Choice Transitions

First Wave		Second Wave						Total	
		SOV		Carpool		Transit			
		N	P%	N	P%	N	P%	N	P%
SOV	UW	1004	94.1	48	4.5	15	1.4	1067	73.9
	W	1300	94.5	53	3.9	22	1.6	1375	81.2
Carpool	UW	69	42.1	89	54.3	6	3.7	164	11.4
	W	89	43.3	111	54.1	5	2.6	205	12.1
Transit	UW	31	14.6	11	5.2	170	80.2	212	14.7
	W	16	14.5	7	6.1	90	79.4	113	6.7
Total	UW	1104	76.5	148	10.3	191	13.2	1443	100
	W	1405	83.0	171	10.1	118	7.0	1694	100

N: Number of persons in cell
P%: Transition probability
UW: Unweighted values
W: Weighted values

An examination of Table 8 shows that unweighted and weighted transition probabilities are not very different from each other. In the case of the Puget Sound Transportation Panel, then, the sample transition probabilities very closely reflected the population transitions. However, the necessity to apply weights before drawing inferences is clearly demonstrated. For example, if one examines the unweighted transition from SOV to transit, only 15 persons fall into this cell. The cell corresponding to the transition from transit to SOV has a frequency of
31. If one were to use these unweighted values for deducing population behavior, then the conclusion would be that transit is losing patronage in preference to driving alone. It may be wrongly concluded that twice as many people are switching away from transit as there are people switching to transit. However, the reality as depicted by the weighted frequencies is very different. In fact, transit is gaining ground by drawing people away from SOV. The weighted transition from SOV to transit is 22, while the transition from transit to SOV is only 16. This conclusion, which is totally in contrast to what unweighted transitions indicated, would have far reaching consequences on the implementation of transportation policies and transit service. The importance of applying weights cannot be emphasized more.

The table also indicates the usefulness of adopting a panel approach. In the table, it appears that carpool is losing patronage with 43% switching to driving alone. On the other hand the switch from driving alone to carpool is only at 4%. This may again lead one to believe that plans targeted towards encouraging carpooling need drastic changes. However, this is not necessarily true as the total patronage of carpool remains rather steady across the two waves. The 4% transition from SOV to carpool is almost sufficient to offset the 43% transition away from carpool, so that the total share of carpool is almost steady (12% in the first wave to 10% in the second wave). Such an analysis is possible only through the use of a panel sample.

This paper has successfully developed a method where results of a choice-based panel sample can be appropriately weighted while accounting for attrition, even when initial choice is endogenous to attrition behavior. The empirical examination in this paper indicated the importance of applying weights before drawing inferences regarding population behavior.

8. References

Amemiya, T. (1985) *Advanced Econometrics.* Harvard University Press, Cambridge, MA.

Ben-Akiva, M. and S.R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand.* The MIT Press, Cambridge, MA.

Cosslett, S.R. (1981) Maximum Likelihood Estimator for Choice-Based Samples. *Econometrica*, 49(5), pp. 1289-1316.

Heckman, J. (1976) The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, 5, pp. 475-492.

Kitamura, R. (1992) A Comparison of Approximate Estimators of the Ordered-Response Probit Model with a Lagged Dependent Variable. Unpublished manuscript. University of California, Davis, CA.

Kitamura, R. (1990) Panel Analysis in Transportation Planning: An Overview. *Transportation Research*, 24A, pp. 401-4 15.

Kitamura, R. and P.H.L. Bovy (1987) Analysis of Attrition Biases and Trip Reporting Errors for Panel Data. *Transportation Research*, 21A, pp. 287-302.

Kish, L. (1965) *Survey Sampling*. John Wiley & Sons, New York.

Lancaster, T. and G. Imbens (1990) Choice-Based Sampling of Dynamic Populations. In J. Hartog, G. Ridder, and J. Theeuwes (ed.) *Panel Data and Labor Market Studies*, Elsevier Science Publishers B.V., North Holland, pp. 21-43.

Maddala, G.S. (1983) *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press. Cambridge, MA.

Manski, C.F. and S.R. Lerman (1977) The Estimation of Choice Probabilities from Choice-based Samples. *Econometrica*, 45(8), pp. 1977-1988.

Manski, C.F. and D. McFadden (1981) Alternative Estimators and Sample Designs for Discrete Choice Analysis. In C.F. Manski and D. McFadden (ed.) *Structural Analysis of Discrete Data*, MIT Press**,** Cambridge, pp. 2-50.

Murakami, E. and W.T. Watterson (1990) Developing a Household Travel Panel Survey for the Puget Sound Region**.** *Transportation Research Record* 1285, pp. 40-46.

Nelson, F.D. (1984) Efficiency of the Two-Step Estimator for Models with Endogenous Sample Selection. *Journal of Econometrics*, 24, pp. 181-196.

Pendyala, R.M., K.G. Goulias, R. Kitamura, and E. Murakami (1992) Development of Weights for a Choice-Based Panel Sample with Attrition. *Transportation Research* (forthcoming).

van Wissen, L.J.G. and H.J. Meurs (1989) The Dutch National Mobility Panel: Experiences and Evaluation. *Transportation*, 16(2), pp. 99-119.