Transportation and Traffic Theory / C.F. Daganzo (Editor)

Weighting Methods for Choice-Based Panels with Correlated Attrition and Initial Choice

R. Kitamura[a], R. M. Pendyala[b] and K. G. Goulias[c]

[a]Department of Civil and Environmental Engineering, University of California at Davis, Davis, CA 95616, U.S.A.

[b]Department of Civil Engineering, Box 42291, University of Southwestern Louisiana, Lafayette, LA 70504, U.S.A.

[c]Department of Civil Engineering, 223 Sackett Building, The Pennsylvania State University, University Park, PA 16802, U.S.A.

Abstract

This paper develops a weighting method that can be applied to choice-based panel samples. The need to study infrequent travel choices has motivated the use of choice-based sampling procedures where sample entities are chosen based on endogenous choice variables. As a choice-based sample is not representative of the population, unbiased inferences can be drawn only after applying weights to the sample. The issue is further complicated if a choice-based sampling technique is employed in a panel survey, where the same behavioral units are observed over time. A choice-based panel sample would need additional treatment for selective attrition, the non-random leaving of panel survey participants. 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 region to obtain unbiased population estimates of transitions in mode choice. This is accomplished by estimating a bivariate probit simultaneous equation system of mode choice and attrition.

1. INTRODUCTION

A choice-based sample is derived by selecting sample entities based on the endogenous choices. 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. A choice-based sample is not representative of the population as the sample share of the infrequent choices exceeds the population share. Drawing unbiased inferences regarding population behavior would require the treatment of choice-based samples using appropriate weighting methods. Several methods have been developed previously [Cosslett, 1981; Imbens, 1992; 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 [Lancaster and Imbens, 1990]. Panel surveys, which offer longitudinal information on the

same behavioral units, facilitate policy analyses and travel demand forecasting based on measured changes in behavior while controlling for unobserved, individual-specific factors that do not change over time [Kitamura, 1990]. 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 may 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].

A choice-based panel sample could offer valuable information regarding dynamics of infrequent choices. This motivates the examination of how choice-based panels, which would involve both endogenous sampling biases and attrition biases, can be treated for drawing unbiased population inferences. This paper aims at developing a weighting method that would jointly account for endogenous sampling and attrition biases.

The Puget Sound Transportation Panel (hereafter, PSTP) [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 to ensure that a sizable number of households using mass transit was included in the overall sample.

In a previous study [Pendyala, et al., 1993], a weighting method was developed and applied to the PSTP sample to generate population estimates of mode choice transitions. However, it was assumed that the choice behavior (on which sampling was based) was exogenous to attrition. Specifically, it was assumed that the error terms for the mode choice and attrition equations were independent, and mode choice was a pre-determined variable in the estimation of attrition probabilities.

This paper tests the veracity of these assumptions by treating mode choice as endogenous to attrition. This is accomplished by estimating a simultaneous equation system with correlated error terms. A bivariate probit formulation is adopted to estimate choice and attrition probabilities. A model system treating attrition and choice behavior independently and a model system incorporating endogeneity are estimated and compared using the PSTP data. The methodology developed in this paper is applicable to any choice-based panel which may need the recognition of endogeneity.

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. Finally, Section 6 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. If the endogenous variable represents a discrete choice, the resulting sample is a choice-based sample.

This section reviews the mathematical formulations of weights for choice-based samples. The extension to incorporate attrition in the case of panels is the focus of the next section. The discussion here closely follows that of Cosslett [1981], Manski and McFadden [1981], Lancaster and Imbens [1990] and Thill and Horowitz [1991].

Let C represent a finite choice set consisting of M mutually exclusive discrete alternatives. Let Z represent the space of explanatory attributes characterizing the population. The population is contained in the product space $C \, X \, Z$. 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

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

$P(j|z,\theta)$ = 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 pure 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 subsets C_b, $b=1,....,B$, where C_b refers to the b-th 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 b-th sampling stratum may be represented as $A_b = C_b \, X \, Z$. It is 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,

$$\sum_{j \in C_b} \int_{z \in Z} f(j,z|\theta) \, dz = \sum_{j \in C_b} \int_{z \in Z} P(j|z,\theta) p(z) \, dz$$

$$= \sum_{j \in C_b} Q(j|\theta)$$

$$= Q(b|\theta) \tag{3}$$

where $Q(j|\theta)$ is the marginal probability of choice j, and $Q(b|\theta)$ is the marginal probability that $j \in C_b$. In other words, $Q(j|\theta)$ is the aggregate market share of alternative j in the population, while $Q(b|\theta)$ is the combined aggregate market share of alternatives contained in stratum A_b. Then, sampling a (j,z) pair involves the joint 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:

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

$$= \frac{H(b)\, f(j,z|\theta)}{\sum_{j \in C_b} \int_{z \in Z} f(j,z|\theta)\, dz}$$

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

$$L_c = \frac{H(b)\, f(j,z|\theta)}{Q(b|\theta)} \tag{4}$$

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,

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

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

An explanation on the choice of weights is provided by Lancaster and Imbens [1990]. The true market share of alternative j is $Q(j|\theta)$. For a sample to be representative of the population, the proportion of the sample choosing alternative j should be $Q(j|\theta)$ also. However, in a choice-based sample, the fraction of the sample choosing alternative j in the bth stratum would be $H(b)Q(j|\theta)/Q(b|\theta)$. When alternative j belongs to several strata, the sample proportion of alternative j would be the summation of such terms over all strata containing it. Therefore, the application of the weight expressed in Equation 5 would ensure

that each alternative j occurs according to the true market share, $Q(j|\theta)$, similar to simple random sampling. The weight $\omega(j)$ compensates for over- or under-sampling alternative j. These weights can be applied to choice-based samples to draw unbiased inferences regarding the population.

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.

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

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

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

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. 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, the adoption of stock or flow sampling procedures merely changes the definition of strata. As the mathematical formulation of choice-based weights is not affected by the definition of strata, the weights derived in the previous section are equally applicable to stock and flow samples for the treatment of endogenous sampling biases.

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 the modeling framework and methodology for deriving joint choice-based attrition weights.

For convenience, let us consider a binary choice variable, i.e., $m=\{0,1\}$. 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):

$$C^* = \theta' z + \psi$$

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

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

where
- C^* = latent variable underlying initial choice behavior
- m = observed indicator of initial choice
- A^* = latent variable underlying attrition behavior
- w = observed indicator of attrition; 1 if continued to participate in panel and 0 otherwise
- θ, β = coefficient vectors
- γ = scalar coefficient
- z = explanatory variables influencing choice behavior
- X = explanatory variables influencing attrition behavior
- ψ, ϵ = random error terms

Vectors Z and X may contain common explanatory variables. The subscript b for the coefficient vector β 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) C^* is independent of X and $\gamma = 0$; and

(b) Error terms are uncorrelated, i.e., $E[\psi \epsilon] = 0$.

A discussion regarding the estimation of the simultaneous equation system under conditions (a) or (b) can be found in Pendyala, et al. [1993]. When either one of the conditions is assumed to be true, the system of equations can easily be estimated using single-equation estimation procedures. In Pendyala, et al. [1993], 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.,

$$P(w=1 \mid X, m) = \Phi(\beta' X + \gamma m) \quad (11)$$

where the left hand side represents the probability of continuing to participate in the panel given the vector X and initial choice m and the right hand side is the standard normal cumulative distribution function evaluated at $\beta' X + \gamma m$.

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

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 reliable.

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. This approach is thus adopted in this study.

Distributional assumptions need to be made on the random error terms ψ and ϵ in order to express response probabilities. The probit offers a theoretically sound formulation for discrete responses. Adoption of the probit formulation in a situation involving two binary choice endogenous variables would imply that the joint distribution of ψ and ϵ is given by the bivariate standard normal. The bivariate probit formulation was first considered by Ashford and Sowden [1970] and Amemiya [1974], but did not see application until recently due to the computational requirements in evaluating two-dimensional bivariate normal integrals.

For the system of equations represented in Equations 9 and 10, the full-information likelihood function for the bivariate probit formulation is now developed. Define sample strata as:

S_1: $m=1$ and $w=1$
S_2: $m=1$ and $w=0$
S_3: $m=0$ and $w=1$
S_4: $m=0$ and $w=0$

Let the joint density of the error terms, ψ and ϵ, be

$$f(\psi,\epsilon) = \left[\frac{2\pi}{\sqrt{(1-\rho^2)}}\right] \exp\left[-\frac{(\psi^2 - 2\rho\psi\epsilon + \epsilon^2)}{2(1-\rho^2)}\right]$$

where ρ represents the correlation between the error terms, ψ and ϵ [see Johnson and Kotz, 1972].

The likelihood function for the first set of observations, S_1, is derived by considering the joint probability of the events, $m=1$ and $w=1$. That is,

$$\begin{aligned}\Pr[m=1,\ w=1] &= \Pr[m=1]\ \Pr[w=1 \mid m=1] \\ &= \Pr[C^* \geq 0]\ \Pr[A^* \geq 0 \mid C^* \geq 0] \\ &= \Pr[\psi \geq -\theta' z]\ \Pr[\epsilon \geq -(\beta' X + \gamma m) \mid \psi \geq -\theta' z] \\ &= \Pr[\psi \geq -\theta' z,\ \epsilon \geq -(\beta' X + \gamma)]\end{aligned}$$

$$= \int_{-\theta' z}^{\infty} \int_{-(\beta' X+\gamma)}^{\infty} f(\psi,\epsilon)\, d\psi\, d\epsilon$$

The likelihood function for this set of observations is,

$$L_1 = \prod_{S_1} \int_{-\theta' z}^{\infty} \int_{-(\beta' X+\gamma)}^{\infty} f(\psi,\epsilon)\, d\psi\, d\epsilon$$

Likelihood functions for sets of observations, S_2, S_3, and S_4 can be derived in a similar manner. The likelihood function for the entire sample will be obtained as,

$$L = \prod_{S_1} \int_{-\theta' z}^{\infty} \int_{-(\beta' X+\gamma)}^{\infty} f(\psi,\epsilon)\, d\psi\, d\epsilon \prod_{S_2} \int_{-\theta' z}^{\infty} \int_{-\infty}^{-(\beta' X+\gamma)} f(\psi,\epsilon)\, d\psi d\epsilon \prod_{S_3} \int_{-\infty}^{-\theta' z} \int_{-\beta' X}^{\infty} f(\psi,\epsilon)\, d\psi d\epsilon$$

$$\prod_{S_4} \int_{-\infty}^{-\theta' z} \int_{-\infty}^{-\beta' X} f(\psi,\epsilon)\, d\psi d\epsilon \qquad (12)$$

Parameter vectors, θ, β, and γ are estimated so as to maximize L. The evaluation of double integrals of the bivariate normal density function is computationally intensive, but tractable. Bivariate probits with observed endogenous indicators as explanatory variables can be estimated using LIMDEP [Greene, 1990].

Once parameter estimates are obtained, the next step is to derive weights for choice-based panel samples. Consider the joint probability of three events; the first that the b-th stratum is chosen, the second that a (j,z) pair is chosen from this stratum, and the third that the unit (j,z) continues to participate in the panel (event represented by Υ) . Then the likelihood that a particular sample unit continues to participate in the choice-based panel is given by

$$\begin{aligned}L_{cp} &= P(b,j,z,\Upsilon \mid \theta,\beta) \\ &= P(b)\ P(j,z \mid b,\theta,\beta)\ P(\Upsilon \mid b,j,z,\theta,\beta) \\ &= H(b)\ P(j,z \in A_b \mid \theta,\beta)\ P(\Upsilon \mid b,j,z,\theta,\beta) \qquad (13)\end{aligned}$$

where $P(\Upsilon|b,j,z,\theta,\beta)$ 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 $L_c = H(b)P(j,z \in A_b|\theta,\beta)$. Substituting the expression for L_c into Equation 13, we obtain,

$$L_{cp} = \frac{H(b)\ f(j,z|\theta)\ P(\Upsilon|b,j,z,\theta,\beta)}{Q(b|\theta)} \tag{14}$$

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

$$\omega(j) = \left[\sum_{j \in C_b, b \in B} \frac{H(b)\ \hat{P}(\Upsilon|b,j,z,\theta,\beta)}{Q(b|\theta)} \right]^{-1} \tag{15}$$

This weight provides logically consistent indications. For example, household types that are over-represented, i.e., households for whom $H(b)/Q(b|\theta)$ is greater than 1, would have low weights applied to them. Similarly, households that tended to leave the panel, i.e., households for which $P(\Upsilon|b,j,z,\theta,\beta)$ is small, would be weighted more heavily.

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 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 transit households was collected through special

recruiting methods. These households were recruited through on-board solicitations of randomly selected bus routes, and by re-contacting 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 ($\geq$ 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.

If a household met multiple criteria, it was assigned to the transit category. The special choice-based recruitment of transit households made the enriched sample have a larger proportion of transit households than in the population.

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.

Table 1
Composition of First Wave and Stayer Samples by Mode Choice

Recruitment Method		Mode			Total Sample
		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	58	58
Metro-Seg.	First Wave	5	1	44	50
	Stayer	4	1	39	44
Metro R/NR	First Wave	1	0	41	42
	Stayer	0	0	34	34
Total	First Wave	1138	193	382	1713
	Stayer	890	137	304	1331

Tele-RDD: Telephone random digit dialing
On-Bus: On-bus solicitation of volunteer participants
Metro Seg.: Volunteers from the Metro Market Segmentation Study
Metro R/NR: Volunteers from the Metro Rider/Non Rider Survey

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, 1331 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.

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 [1965]. 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 $[41.4/57.9]^{-1} = 1.399$. As residences from King county were under-sampled, these households are applied with a weight greater than unity.

Table 2
Households by County of Residence

	Survey Sample		Population(1989)		
County	N	%	N	%	Weight
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

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)

	Unweighted		Weighted	
Mode Choice	N	%	N	%
SOV	890	67	1,238	79
Carpool	137	10	193	12
Transit	304	23	143	9
Total	1,331	100	1,574	100

Mode choice and attrition behavior are modeled as per Equations 9 and 10 using the full-information maximum likelihood approach outlined in Equation 12. The model system was estimated on 1682 first-wave households for which complete data were available. The percent attrition in the sample by initial mode choice is shown in Table 4. There are 1331 stayers and 351 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	1,116	20.3
Carpool	137	54	191	28.3
Transit	304	71	375	19.2
Total	1,331	352	1,682	20.9

Four different specifications were used to estimate the model system. The first represents a model system in which initial choice and attrition are assumed to be mutually independent. The error correlation, ρ, is specified to be zero, and the endogenous explanatory variable, m, is eliminated from the attrition equation (i.e., $\gamma=0$). This

specification reduces the estimation effort to one of two independent binary probits. In the second model specification, error correlation, ρ, is not constrained to be zero, but the endogenous explanatory variable, m, is absent in the attrition equation, leading to a bivariate probit system with no endogenous explanatory variables. The third specification is one in which the error correlation, ρ, is set to zero, but the endogenous explanatory variable, m, is included in the attrition equation. This again represents two independent binary probits, but includes an endogenous explanatory variable. Finally, the fourth specification is the most general case, including both error correlation and the endogenous explanatory variable.

Results of the simultaneous equation estimation effort are shown in Table 5. The first portion corresponds to the mode choice model of Equation 9, while the latter portion corresponds to the attrition model.

Table 5
Mode Choice and Attrition Models

	Model 1 ($\rho=0$, $\gamma=0$)		Model 2 ($\rho\neq0$, $\gamma=0$)		Model 3 ($\rho=0$, $\gamma\neq0$)		Model 4 ($\rho\neq0$, $\gamma\neq0$)	
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Mode Choice								
Constant	-2.306	-8.57	-2.306	-10.58	-2.306	-8.57	-2.306	-10.56
ONECAR	1.273	5.60	1.273	6.03	1.273	5.60	1.273	6.02
TWOCARS	1.231	5.42	1.231	5.75	1.231	5.42	1.231	5.74
MULTICARS	1.152	4.97	1.152	5.24	1.152	4.97	1.152	5.24
#CAR=#DRVR	1.362	7.65	1.363	7.66	1.362	7.65	1.362	7.61
YEARHOME	.076	2.97	.076	2.86	.076	2.97	.076	2.87
LOINCOME	.155	1.61	.155	1.58	.155	1.61	.155	1.58
HIGHINCOME	-.188	-2.21	-.188	-2.22	-.188	-2.21	-.188	-2.22
BUSDIST	.018	1.44	.018	1.39	.018	1.44	.018	1.41
Attrition								
Constant	.657	2.94	.663	2.90	.658	2.94	.658	2.85
ONECAR	.520	2.70	.525	2.62	.501	2.56	.501	1.74
TWOCARS	.714	3.59	.717	3.50	.689	3.39	.689	2.22
MULTICARS	.751	3.57	.751	3.54	.723	3.35	.723	2.28
NWORKERS	.123	2.31	.127	2.33	.129	2.38	.129	2.25
YEARHOME	.100	3.38	.101	3.50	.100	3.38	.100	3.34
LOINCOME	-.209	-2.15	-.208	-2.12	-.211	-2.17	-.211	-2.09
HIGHINCOME	-.136	-1.43	-.137	-1.45	-.135	-1.43	-.135	-1.42
SGLADULT	-.451	-2.51	-.446	-2.47	-.447	-2.49	-.447	-2.47
YNGADULTS	-.560	-3.86	-.557	-3.77	-.556	-3.82	-.556	-3.74
MIDADULTS	-.212	-2.19	-.210	-2.23	-.209	-2.16	-.209	-2.22
HHLDSIZE	-.169	-4.74	-.168	-4.66	-.167	-4.65	-.167	-4.64
TELE-RDD	-.312	-2.34	-.337	-2.35	-.340	-2.39	-.340	-2.36
SOV(γ)					.049	.57	.050	.13
$\rho_{\varepsilon\varepsilon}$			.028	0.53			-.001	-.01

Table 5 (Continued)
Mode Choice and Attrition Models

Goodness-of-fit Statistics				
L(0)	-2331.7	-2331.7	-2331.7	-2331.7
L(C)	-1935.8	-1935.8	-1935.8	-1935.8
$L(\theta,\beta)$	-1815.9	-1815.8	-1815.7	-1815.7
$-2[L(0)-L(\theta,\beta)]$	1031.7(df=22)	1032.0 (df=23)	1032.0 (df=23)	1032.0 (df=24)
$-2[L(C)-L(\theta,\beta)]$	239.8 (df=20)	240.1 (df=21)	240.1 (df=21)	240.1 (df=22)

Description of Variables

Variable	Description
ONECAR	Dummy variable=1 if household owns one car; 0 otherwise
TWOCARS	Dummy variable=1 if household owns two cars; 0 otherwise
MULTICARS	Dummy variable=1 if household owns more than two cars; 0 otherwise
#CAR=#DRVR	Dummy variable=1 if number of cars ≥ number of drivers in household
YEARHOME	Number of years in current residence
NWORKERS	Number of employed persons in household
LOINCOME	Dummy variable=1 if annual household income ≤ $15,000
HIGHINCOME	Dummy variable=1 if annual household income > $50,000
BUSDIST	Dummy variable=1 if nearest bus stop is within 1/4th mile of household
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	Dummy variable=1 if household is an SOV household
Mode Choice	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

For purposes of model estimation, the mode choice variable was dichotomized into SOV and non-SOV households. Non-SOV households included both carpool and transit households. Model estimation was performed using the econometric software package LIMDEP [Greene, 1990]. It conveniently allows the user to specify parametric restrictions in model specification. Bivariate probit models were estimated using LIMDEP and later confirmed with routines written in GAUSS [1992]. Results obtained through LIMDEP and GAUSS were found to be very similar. As such, the estimates provided by LIMDEP are used in this paper.

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. This can be deduced through an examination of the estimates of ρ which are not statistically significant (i.e., not different than zero at the 5% level). Panel participation and the choice process on which endogenous sampling was based are mutually independent. Mode choice may therefore be treated exogenous to attrition. This implies that the findings reported in Pendyala, et al. [1993], based on the assumption of an uncorrelated error structure, are valid.

All four model specifications provided consistent and expected 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. Households with higher levels of car availability per driver show a greater propensity to be SOV households. 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 (at a 5% level) on mode choice to work for household members.

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, single adult households, and households with young and middle age adults with no children tended to leave the panel. Larger household sizes also contributed to households leaving the panel. Households recruited by telephone random digit dialing are significantly more likely to leave the panel and those collected by special choice-based methods tended to continue participation. The variable SOV, representing the endogenous explanatory variable, m, does not exhibit a significant effect on attrition.

A comparison across model specifications 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 all four model estimations. Panel participation in subsequent waves is not dependent on the initial choice variable based on which the households were sampled. Under these conditions, panel participation probability may be computed for a household using an independent univariate binary probit model as,

$$P(w=1 \mid X,m) = \Phi(\beta' X + \gamma m) \tag{18}$$

which is the same as Equation 11.

This can be combined with weights developed in Table 1, and Equations 16 and 17 to compute overall choice-based panel weights for each household. For example, a carpool household from Pierce county would be weighted thus:

$$\omega(\text{Pierce County, Carpool}) = 0.949 \text{ X } 1.108 \text{ X } \Phi(\beta' X + \gamma m) \tag{19}$$

After the application of the joint weights similar to that shown in Equation 19, the weighted sample was found to be as in Table 6. 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 attrition was found to be independent of the sampling procedure. The application of the joint weighting procedure produced a total sample size that is close to the original first wave sample of 1682.

Table 6
Unweighted and Weighted Sample

	Unweighted		Weighted	
Mode Choice	N	%	N	%
SOV	890	67	1,289	78
Carpool	137	10	200	12
Transit	304	23	161	10
Total	1,331	100	1,650	100

7. CONCLUSIONS

In the case of the Puget Sound Transportation Panel, initial mode choice was found to be not endogenous to attrition. As such, the independent probit 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 7 with unweighted and weighted values.

An examination of Table 7 shows that unweighted and weighted transition probabilities are quite similar to one another. 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 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 23, while the transition from transit to SOV is only 14. This conclusion, which is totally in contrast to what unweighted transitions indicated, could have far reaching policy implications.

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; the switch from driving alone to carpool is only at 4%. This may again lead one to believe that the market share of carpool is diminishing. However, this is not necessarily true as the total

patronage of carpool remains rather steady between 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.

Table 7
Person Mode Choice Transitions

		Second Wave							
		SOV		Carpool		Transit		Total	
First Wave		N	P%	N	P%	N	P%	N	%
SOV	UW	1,004	94.1	48	4.5	15	1.4	1,067	73.9
	W	1,308	94.4	54	3.9	23	1.7	1,385	81.6
Carpool	UW	69	42.1	89	54.3	6	3.7	164	11.4
	W	90	42.5	116	54.7	6	2.8	212	12.5
Transit	UW	31	14.6	11	5.2	170	80.2	212	14.7
	W	14	13.9	6	5.9	81	80.2	101	5.9
Total	UW	1,104	76.5	148	10.3	191	13.2	1,443	100.0
	W	1,412	83.1	176	10.4	110	6.5	1,698	100.0

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

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 methodology adopted in this paper allows for convenient testing of endogeneity while recognizing the simultaneous nature of the choice processes. The empirical examination in this paper indicated the importance of applying weights before drawing inferences regarding population behavior.

8. REFERENCES

Amemiya, T. (1974) Bivariate Probit Analysis: Minimum Chi-Square Methods. *Journal of the American Statistical Association*, 69, 940-944.

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

Ashford, J.R. and R.R. Sowden (1970) Multi-variate Probit Analysis. *Biometrics*, 26, 535 - 546.

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), 1289-1316.

_____ (1992) *GAUSS386 version 3.0*. Aptech Systems, Inc. Maple Valley, WA.

Greene, W.H. (1990) *LIMDEP version 5.1*, Econometric Software Inc., New York.

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**, 475-492.

Imbens, G. (1992) An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling. *Econometrica*, **60**(5), 1187-1214.

Johnson, N. L. and S. Kotz (1972) *Distributions in Statistics: Continuous Multivariate Distributions*. John Wiley & Sons, New York.

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**, 401-415.

Kitamura, R. and P.H.L. Bovy (1987) Analysis of Attrition Biases and Trip Reporting Errors for Panel Data. *Transportation Research*, **21A**, 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, 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), 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, 2-50.

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

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

Pendyala, R. M., K. G. Goulias, R. Kitamura, and E. Murakami (1993) Development of Weights for a Choice-Based Panel Sample with Attrition. *Transportation Research* (forthcoming).

Thill, J. C. and J. L. Horowitz (1991) Estimating a Destination-Choice Model from a Choice-based Sample with Limited Information. *Geographical Analysis,* 23 (4), 298-315

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