Sample Selection Bias with Multiple Selection Rules: Application with Residential Relocation, Attrition, and Activity Participation in Puget Sound Transportation Panel

JIN-HYUK CHUNG AND KONSTADINOS G. GOULIAS

Two sources of sample selection bias emerging simultaneously from panel attrition and residential relocation and their effect on activity participation are examined. The data used were from two time points (Wave 1 in 1989 and Wave 2 in 1990) of the Puget Sound Transportation Panel. Data regarding relocation decisions, taking place between Wave 1 and Wave 2, are available for the households that participated in both waves (participants) and are not available for the households that participated in the first wave only (dropouts). Double selection was associated with the possible simultaneous or sequential decision process underlying participation in the survey and household residential relocation. The method used is based on a bivariate probit model that accounts for selectivity. The method emerges from the unknown relocation status of the dropouts in Wave 2. Subsequent creation of correction terms, needed to account for the lack of data on dropout households' activity participation in Wave 2, uses the probit model. The method, called the Tunali method, is a two-step procedure that follows the usual Heckman method. The models estimated, that is, the bivariate probit model of double selection and activity participation linear regressions corrected and uncorrected for selection, are provided.

Dynamic analysis of travel behavior is greatly facilitated when panel survey data—information from repeated observations of the same individuals over time—are available. A common problem to all panels, however, is the potential selectivity bias emerging from attrition or refusal to participate in a subsequent time point of the survey. Ordinary least-squares (OLS) regression coefficient estimates are inconsistent if attrition occurs in a systematic way, and it is not accounted for in estimation. Analogously, selectivity bias may also emerge from other sources. For example, nonrandom residential relocation (or, more generally, migration) during the panel survey may also produce similar biases. In addition, attrition and residential relocation decision making may also be related. For example, relocating residents may be more likely to refuse participation in the panel in subsequent waves. A method is needed to remove selectivity bias in which attrition and residential relocation are considered simultaneously. This would allow researchers to test hypotheses about the relationship between attrition and relocation, derive sample weights that can be used for subsequent waves of a panel, and provide for a complete correction method for regression models that suffer from selectivity biases.

The most common selectivity bias correction method, used in transportation modeling, takes the form of an equation that represents the selection process with a discrete dependent variable (e.g., participation or nonparticipation in a survey). Another equation represents the outcome of some decision-making process (e.g., number of household trips or number of cars owned by a household). This is the equation for which consistent estimates are needed. The usual technique to account for selectivity bias has been to create "correction" terms used to augment the target regression equation and "eliminate" the selectivity bias as if it were a specification error. This method treats selectivity as a specification error and is named the Heckman correction method (*1,2*). The method has been used by Mannering (*3*), Kitamura and Bovy (*4*), Hensher et al. (*5*), and Monzon et al. (*6*). In this paper this method is called the single-selection model because it includes only one source of selectivity. When the sources of selectivity are several, similar methods can be devised and multiple correction terms can be used to eliminate the bias. These methods, however, are more complex than the single-selection method. Their complexity increases exponentially when relationships exist between the selectivity sources and when portions of the "selected" sample are unobserved (*7*).

In this paper two sources of selectivity are considered: panel attrition and residential relocation. Their effect on activity participation is also examined. The data used are from the first two time points (Wave 1 in 1989 and Wave 2 in 1990) of the Puget Sound Transportation Panel [PSTP, described by Murakami and Watterson, (*82*)]. Data about relocation decisions, taking place between Wave 1 and Wave 2, are available for the households that participated in both waves (participants) and are not available for the households that participated in the first wave only (dropouts). This precludes the use of the methods devised by Kitamura et al. (*9*) and may be the source of "double selection," as a result of the possible simultaneous or sequential decision process underlying participation in the survey and residential relocation. The method, based on a bivariate probit model, accounts for selectivity caused by the unknown relocation status of the dropouts. The lack of data on activity participation for the panel dropouts is another source of selectivity. The method creates two correction terms to be used in the Wave 2 activity participation equations.

First the paper presents a more general model of double selection. Then, the selectivity model is described with a few estimation issues. It then provides a short description of the data analyzed. Then, estimation results for the bivariate probit model of attrition and residential relocation and the augmented regressions (with the two correction terms) of activity participation are provided. A summary and conclusion are offered last.

Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, Pa. 16802.

MODEL

The general model of double selectivity, of which the model used in this paper is a particular case, is formulated as follows. Each household in the sample is characterized by two discrete-outcome decisions, to participate in the Wave 2 of the panel and to change the residential location between Wave 1 and Wave 2. A third decision is characterized by a "continuous" outcome, that is, frequency of activity participation in Wave 2. Using the dichotomous variables, Y_1 and Y_2, to represent the two discrete outcome decisions and the continuous variable Y_3 to represent the continuous outcome, it is possible to write the two selection "rules" in terms of explanatory variables such as

$$Y^*_{1i} = \beta'_1 X_{1i} + \epsilon_{1i}$$

$$Y_{1i} = 1 \qquad \text{if } Y^*_{1i} > 0$$

$$Y_{1i} = 0 \qquad \text{if } Y^*_{1i} \leq 0 \tag{1}$$

and

$$Y^*_{2i} = \beta'_2 X_{2i} + \epsilon_{2i}$$

$$Y_{2i} = 1 \qquad \text{if } Y^*_{1i} > 0$$

$$Y^*_{1i} = 0 \qquad \text{if } Y^*_{1i} \leq 0 \tag{2}$$

The third equation describing the continuous dependent variable is as follows:

$$Y_{3i} = \beta'_3 X_{3i} + \sigma_3 \epsilon_{3i} \tag{3a}$$

where

X_{ki} = vectors of explanatory variables (k = 1,2,3),
σ_3 = unknown scale parameter, and
β_k = unknown regression coefficient vectors to be estimated with the elements of the variance-covariance matrix of (ϵ_{1i}, ϵ_{2i}, ϵ_{3i}) reported in Equation 4:

$$\sum = \begin{bmatrix} 1 & \rho & \rho_{13} \\ \rho & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} \tag{4}$$

Equations 1 through 4 describe the structure of the model under consideration. The household observations contain information on Y_{1i}, Y_{2i}, Y_{3i}, and X_{1i}, X_{2i}, X_{3i}. Y^*_{1i} and Y^*_{2i} can be interpreted as the propensity of the household to relocate and to participate in the second wave of the panel survey, respectively. Considering the two discrete outcome variables, described by Equations 1 and 2, there are four possible joint outcomes. In Figure 1 this can be indicated by a four-cell table containing the frequency of the number of households in each combination of outcomes. Assuming that the assumptions ϵ_{1i}, ϵ_{2i}, ϵ_{3i} are trivariate normally distributed with 0 mean and covariance given by Equation 4, and error terms independent across households and the explanatory variables, then it is possible to write the joint cell probabilities reported in the second part of Figure 1.

The probability density, associated with each cell, of Y_{3i}, can be written as a function of the cell probability and the trivariate normal density of the ϵ's. These components in turn can be used to derive a likelihood function for the entire system of equations and then use it for estimation via maximum likelihood. A problem arises, however, when some cells in Figure 1 are not observed.

In Figure 1 the data present four possible distinct regimes defined by the combination in outcomes depicted by the variables Y_1 and Y_2. [There will be four pairs of possible joint outcomes for Y_1 and Y_2. (0,0), (0,1),(1,0), and (1,1).] Letting [$Y_1 \times Y_2$] be the joint outcome of the two variables in Figure 1, the expectation of Equation 3a can be written as

$$E(Y_{3i}, Y_1 \times Y_2) = \beta'_3 X_{3i} + \sigma_3 E(\epsilon_{3i} \mid X_{3i}, Y_1 \times Y_2) \tag{5}$$

In Figure 1 there are four distinct subsamples. One equation of the type described in Equation 5 applies to each. However, panel attrition and residential relocation are characterized by the lack of information on residential relocation of households that dropped out of the panel. In terms of Figure 1, there are only three distinct cells:

Frequencies

	Y_2	
Y_1	0	1
0	N_1	N_2
1	N_3	N_4

Probabilities

	Y_2	
Y_1	0	1
0	$BN(\beta_1 X_1, \beta_2 X_2, \rho) = \int_{-\infty}^{-\beta_1 X_1} \int_{-\infty}^{-\beta_2 X_2} f(\epsilon_1 \epsilon_2)\, d\epsilon_1 d\epsilon_2$	$BN(-\beta_1 X_1, \beta_2 X_2, -\rho) = \int_{-\infty}^{-\beta_1 X_1} \int_{-\beta_2 X_2}^{\infty} f(\epsilon_1 \epsilon_2)\, d\epsilon_1 d\epsilon_2$
1	$BN(\beta_1 X_1, -\beta_2 X_2, -\rho) = \int_{-\beta_1 X_1}^{\infty} \int_{-\infty}^{\beta_2 X_2} f(\epsilon_1 \epsilon_2)\, d\epsilon_1 d\epsilon_2$	$BN(-\beta_1 X_1, -\beta_2 X_2, \rho) = \int_{-\beta_1 X_1}^{\infty} \int_{-\beta_2 X_2}^{\infty} f(\epsilon_1 \epsilon_2)\, d\epsilon_1 d\epsilon_2$

FIGURE 1 Four discrete outcomes and associated probabilities.

1. Participants who change residential location (movers) and took part in both panel waves (participants),
2. Participants who did not change their residential location (stayers) and took part in both panel waves, and
3. Participants in Wave 1 only (dropouts) of unknown residential relocation choice.

It is clear then, that observation of residential status is conditional on panel attrition (herein called incomplete information). In terms of Figure 1 this is equivalent to "collapsing" two cells into one. For these cells instead of a bivariate normal cell probability one obtains a univariate normal probability (e.g., corresponding to the probability of panel attrition). Estimation of Equations 1 and 2 also can be performed using a log likelihood function that is analogous to the usual bivariate probit likelihood function.

Consider Y_1 representing residential relocation status (taking the value of 1 if the household did not move and 0 otherwise) and Y_2 representing panel participation (taking the value of 0 if the household is a dropout and 1 otherwise). The cells with incomplete information are ($Y_1 = 0$, $Y_2 = 0$) and ($Y_1 = 1$, $Y_2 = 0$). The sample size of each distinct cell is ($N_1 + N_3$) for the dropouts, N_2 for participant-movers, and N_4 for participant-stayers. The log likelihood function associated with Equations 1 and 2 is as follows:

$$L^* = \sum_{i=1}^{N_4} \ln BN\,[\beta_1' X_{1i}, \beta_2' X_{2i}, \rho] + \sum_{i=1}^{N_2} \ln BN\,[\beta_1' X_{1i}, -\beta_2' X_{2i}, -\rho] + \sum_{i=1}^{N_1+N_3} \ln \Phi[-\beta_2' X_{2i}]$$

where BN is the bivariate normal standard distribution and Φ is the univariate normal standard distribution (this is the effect of "collapsing" two cells because of a lack of residential relocation data on the dropouts). This function can be used to estimate the regression coefficients in Equations 1 and 2 and the correlation coefficient between their two error terms (ρ). One can use either maximum likelihood or any other method as in work by Amemiya (*10*). A pseudo t-test associated with ρ can be used to verify that a bivariate probit model is a more appropriate formulation than two univariate probit models for Equations 1 and 2. Alternatively, a nested likelihood ratio chi-square test can also be applied.

The second objective of estimation in this paper is to obtain consistent estimates of β_3 and to examine the sign and magnitude of the parameters in Equation 5. The selectivity "problem" arises when $E(\epsilon_{3i} \mid X_{3i}, Y_1 \times Y_2) \neq 0$ and OLS is used to estimate Equation 3a. For the cells in which Y_{3i} is observed a trivariate normal density applies and the related likelihood function is analogous to the complete cell membership discussed before. In this paper, instead of employing a method that involves trivariate normal densities, an alternative procedure that produces equally consistent estimates is used.

The method was devised by Tunali (*11*) and is the double-selection analog of the Heckman single-selection correction method (called the Tunali method here). It is a two-step procedure, which at the first step employs maximum likelihood estimation for Equations 1 and 2 to obtain consistent estimates of the two correction terms (λ_1 and λ_2). At the second step, the estimates of the λ's are used to correct for specification error (emerging from selection bias) in the regression of Y_{3i}. The system of the equations to consider is given by Equation 1, Equation 2, and the following, augmented continuous dependent variable regression:

$$Y_{3i} = \beta_3' X_{3i} + \gamma_1 \lambda_1 + \gamma_2 \lambda_2 + \sigma_3 \epsilon_{3i}^* \qquad (3b)$$

where γ_1 and γ_2 are functions of σ_3 and the correlations in Equation 4 and can be estimated by least-squares regression. λ_1 and λ_2 are the double-selection analogs of the Mill's ratios in single selection. The λ's are functions that involve data from the selection rules in Equations 1 and 2. ϵ_{3i}^* is a heteroskedastic error term. When OLS is applied to Equation 3b the usual standard errors of the coefficient estimates are biased. This is allowed for by "correcting" the OLS standard error estimates used for hypothesis testing. Estimation of the correction terms (λ_1 and λ_2), their associated coefficients (γ_1 and γ_2), and the associated standard error follows LIMDEP (*12*), which follows the Heckman two-step method.

DATA

PSTP is the first general-purpose urban transportation survey in the United States. The major goals of the panel are to (*a*) track changes in employment, work characteristics, household composition, and vehicle availability; (*b*) monitor changes in travel behavior and response to changes in the transportation environment; and (*c*) examine changes in attitudes and values of transit and nontransit users. PSTP includes household, person, trip, and attitude information of four waves, with each pair of waves a year apart. The first-wave data collection took place from September to early December 1989. The second-wave survey was conducted in the fall of 1990. An extensive description of the panel is provided by Murakami and Watterson (*8*).

In this paper, the analysis uses selected travel diary information from the first two waves. The travel diary includes continuous 48-hr activities (excluding the in-home activities) for each wave. It includes every trip a person made in 2 days. Each trip was characterized by trip purpose, type, mode, start/end time, travel duration, origin/destination, and distance. From this data set out-of-home activity engagement information can be derived using the trip purposes. The raw data were "cleaned" from any inconsistencies and the records with complete information are used here.

In the original data set, trip purposes are classified into eight different types (work, school, college, shopping, personal business, appointments, visiting, and free time). Models for all the activities considered together (sum of activities) and by grouping activities in a few categories were estimated. Assuming that a household, within a given 24-hr period, prioritizes its activity participation according to the relative importance of each activity, a natural grouping would be the following hierarchy (with a decreasing degree of constraint and importance): subsistence (work, school, college), maintenance (shopping, personal, appointments), and leisure (visiting, free-time) activities. The models treated for selectivity are models of subsistence frequency, maintenance frequency, and leisure frequency, each considered separately. A fourth model representing the sum of all activities is also estimated to identify possible "loss" of information when usual trip generation models are formulated.

Information on residential relocation was also collected within the panel. The data analyzed in this paper are from 1,662 households, of which 1,313 (79 percent) participated in both panel waves and 349 (21 percent) participated in Wave 1 only. From among the 1,313 participants, 111 (8 percent) changed residential location between the two waves, whereas 1,202 (92 percent) did not.

EMPIRICAL EXAMPLE

An application of the double-selection model mentioned earlier is provided here to address two related issues. The first is with respect to potential sample biases in a Wave 2 sample emerging from selec-

tive attrition and possible selective residential relocation. Restoring representativeness in the PSTP can be performed using weights derived from the bivariate probit model (joint attrition and relocation) mentioned earlier. Sample weights for subsequent waves aim at recreating population representativeness in the panel. One can use the results up to this point as seen elsewhere (*9*) to create sample weights. The second is with respect to consistent parameter estimation for the regression equations representing activity participation in Wave 2. The sample in Wave 2 contains only partial information on the population because of the double selection with part of the observations containing incomplete classification (i.e., the dropouts cannot be classified into movers and stayers). This affects the expectation of the error term in Equation 3b. The Tunali double-correction terms can be used to gain coefficient consistency. The definition of variables, cell frequencies, and average characteristics per group are presented in Table 1. The average value for each variable used in the models is presented separately for each of the three groups considered in this paper.

The first model of interest is the bivariate probit model with selection. Table 2 contains the single equation results, that is, estimates of two independent univariate probit equations ($\rho = 0$) and the bivariate probit estimates ($\rho \neq 0$). Model specification was defined mainly on the basis of past results using a similar data set on attrition and relocation and indications from past literature.

The regression parameter estimates are consistent (in terms of signs and relative magnitude) in the two models. With respect to the attrition model, as expected, the results confirm previous research using a similar data set. Households with a higher car ownership level, higher employment, and longer duration of residence in Wave 1 are more likely to participate in both waves of the panel. Confirming the usual tendency reported in other surveys, low-income households, single-adult households, and childless households with relatively young household composition tend to drop out after the first panel wave. People recruited via random digit dialing (in the sample analyzed here 92 percent are recruited via random digit dialing and 8 percent by special choice-based methods) tend to stay in the panel. The relocation equation exhibits agreement between the single-equation estimation and bivariate probit estimates. The household life-cycle stage is an important determinant of relocation (that is, households at their earlier stages are more likely to move than at their later stages). This is reflected by the coefficients of the two variables representing the number of children in the household. An interesting result is that the residence tenure (the dummy variable associated with 5 years or more in the current residence) has a negative coefficient. This may be an indication that, as residence tenure increases, the household is less likely to move. All three indi-

TABLE 1 Definition of Variables and Sample Characteristics

Variable	Description
FEMALESx	Number of females in the household in wave x
DRIVERSx	Number of drivers in the household in wave x
WORKERSx	Number of workers in the household in wave x
KID(0-5)x	Number of children whose age is less than five years in wave x
MIDINCOMEx	Dummy variable =1 if annual household income is between $15,000 and $50,000 in wave x ; 0 otherwise
HIGHINCOMEx	Dummy variable =1 if annual household income is more than $ 50,000 in wave x ; 0 otherwise
SGLADULTx	Dummy variable =1 if household has only one adult less than 35 years and no children in wave x ; 0 otherwise
YNGADULTSx	Dummy variable =1 if household has two or more adult less than 35 years and no children in wave x ; 0 otherwise
MIDADULTSx	Dummy variable =1 if household has two or more adult aged 35-64 years and no children in wave x ; 0 otherwise
YRHOME(0-1)x	Dummy variable =1 if number of years in current residence is less than one year in wave x ;0 otherwise
YRHOME(1-5)x	Dummy variable =1 if number of years in current residence is between one and five years in wave x ;0 otherwise
YRHOME(5-10)x	Dummy variable =1 if number of years in current residence is between five and ten years in wave x ;0 otherwise
ONECARx	Dummy variable =1 if household owns one car in wave x ; 0 otherwise
TWOCARSx	Dummy variable =1 if household owns two cars in wave x ; 0 otherwise
MULTICARSx	Dummy variable =1 if household owns more than two cars in wave x; 0 otherwise
TELE-RDD	Dummy variable =1 if household recruited by telephone random digit dialing
HHLDSIZEx	Household size in wave x
KINGx	Dummy variable =1 if residence locate in King County in wave x ; 0 otherwise
PIERCEx	Dummy variable =1 if residence locate in Pierce County in wave x ; 0 otherwise
SNOHOMISx	Dummy variable =1 if residence locate in Snohomish County in wave x ; 0 otherwise
Relocation	Binary Choice Dependent Variable = 1 if household has moved in second wave in panel
Attrition	Binary Choice Dependent Variable = 1 if household continues to participate in second wave of panel

Note : x=1 and 2 in variables indicate wave 1 and wave 2.

(*continued on next page*)

TABLE 1 (*continued*)

	Sample Mean of Variables		
	Participants and stayers	*Participants and movers*	*Non-participants in Wave2*
FEMALES1	.975	.883	.966
DRIVERS1	1.735	1.523	1.653
WORKERS1	1.256	1.243	1.206
KID(0-5)1	.218	.297	.310
KID(0-5)2	.204	.288	
MIDINCOME1	.651	.685	.590
HIGHINCOME1	.194	.153	.198
MIDINCOME2	.523	.478	
HIGHINCOME2	.333	.396	
SGLADULTS1	.029	.117	.063
YNGADULTS1	.050	.153	.109
MIDADULTS1	.292	.207	.238
YRHOME(0-1)1	.122	.297	.241
YRHOME(1-5)1	.333	.469	.384
YRHOME(5-10)1	.546	.234	.375
ONECAR1	.229	.324	.264
TWOCARS1	.449	.414	.415
MULTICARS1	.289	.225	.255
ONECAR2	.216	.270	
TWOCARS2	.426	.297	
MULTICARS2	.297	.162	
TELE-RDD	.950	.793	.560
HHLDSIZE1	2.575	1.820	2.752
HHLDSIZE2	2.513	1.182	
KING1	.400	.541	.410
PIERCE1	.207	.109	.261
SNOHOMIS1	.262	.198	.249
Frequency	1202	111	349

cators of county of residence (King, Pierce, and Snohomish) show that the movers are more likely to be from the fourth county (Kitsap). The most important result here is the lack of significance (and relatively small magnitude) of the error correlation coefficient between relocation and attrition (ρ). (The use of this method provides for clearer indications about the relationship between relocation and attrition. The usual caveat on the estimated standard error of ρ applies as well.) Similar to previous results on attrition and mode choice (*9*) and based on this paper, attrition is not correlated with other choices households make.

The results here provide some guidance on sample weight creation procedures. The results also reinforce past approaches to "sequential" and independent weight creation, that is, deriving weights that transform the Wave 2 panel sample into a representative sample by sequentially applying single-source derived weights to account for each source-specific sample bias.

The estimated bivariate probit model is used to create consistent estimates for the λ's for two out of the four cells in Figure 1. The first, corresponding to ($Y_1 = 0$, $Y_2 = 1$), represents the panel participants in both waves who did not relocate (participant stayers) and the second, corresponding to ($Y_1 = 1$, $Y_2 = 1$), represents the panel participants in both waves who relocated (participant movers). Four models are presented here for Y_3. The first three, in Table 3, depict 2-day household activity participation frequencies for subsistence, maintenance, and leisure. The fourth model depicts the sum of subsistence, maintenance, and leisure (called the total frequency of household activity participation resembling a trip generation model).

Table 3 provides a comparison between OLS and the Tunali method. The specification of all the models is the same in an attempt to provide a common basis for comparison. Alternative specifications provided similar results and are not presented here. Some of these models are underspecified, and this has an effect on the significance of the correction terms (*11*).

The standard errors of the coefficient estimates reported here (denominators in the "t-stats") are also corrected for selection on the basis of the method reported in LIMDEP (*7*). This is the same method used by Tunali for the two groups analyzed here (*11*). A consistent estimator is used for the standard error of the regression equation (Equation 3b) and is based on the usual OLS residuals with a correction (*12*). Estimates for the error correlation coefficients (ρ_{13} and ρ_{23}) are obtained with algebraic manipulations that involve the coefficients of the correction terms, the correlation in the bivariate probit model, and the standard error of the regression in Equation 3b. Unfortunately, in practice, this may produce correlation coefficients that are not within the unit circle, posing great difficulties in interpreting the coefficients.

With respect to the subsistence equation, one can observe a general agreement in the signs and relative magnitudes of the coefficients between the OLS and the Tunali models for both groups,

TABLE 2 Residential Relocation and Panel Attrition Models

	Univariate Probit		Bivariate Probit	
	Coef.	"t-stat"	Coef.	"t-stat"
Relocation				
Constant	-.831	-3.680	-.959	-2.916
FEMALES1	.015	.121	.007	.067
DRIVERS1	-.212	-2.104	-.200	-2.028
KID(0-5)1	.118	1.398	.100	.965
MIDINCOME1	-.010	-.067	.009	.051
HIGHINCOME1	-.075	-.386	-.067	-.315
SGLADULT1	.565	2.572	.513	1.845
YNGADULTS1	.597	3.131	.533	2.027
MIDADULTS1	.087	.624	.083	.559
YRHOME(5-10)1	-.511	-4.289	-.478	-3.328
TELE-RDD	-.050	-1.469	-.046	-1.226
KING1	.021	.135	.031	.187
PIERCE1	-.383	-1.926	-.367	-1.708
SNOHOMIS1	-.121	-.682	-.113	-0.611
Attrition				
Constant	.869	4.431	.880	4.376
ONECAR1	.405	2.201	.396	2.095
TWOCARS1	.582	3.098	.569	2.963
MULTICARS1	.574	2.896	.558	2.785
WORKERS1	.135	2.534	.132	2.448
YRHOME(0-1)1	-.425	-4.065	-.435	-4.158
YRHOME(1-5)1	-.209	-2.489	-.204	-2.403
LOWINCOME1	-.188	-1.522	-.198	-1.580
HIGHINCOME1	-.116	-1.229	-.114	-1.211
SGLADULT1	-.457	-2.533	-.455	-2.519
YNGADULTS1	-.532	-3.614	-.526	-3.575
MIDADULTS1	-.211	-2.151	-.210	-2.207
HHLDSIZE1	-.174	-4.800	-.172	-4.811
TELE-RDD	.037	1.618	.037	1.497
ρ (1,2)			.310	.415
Goodness-of-fit Statistics				
Relocation				
Log-likelihood	-343.37		Log-Likelihood	-1161.03
Restricted Log-likelihood	-380.40		Restricted Log-likelihood	-1234.57
Chi-Squared (df=13)	74.06		Chi-squared (df=27)	147.08
Attrition				
Log-likelihood	-817.82			
Restricted Log-likelihood	-854.17			
Chi-Squared (df=13)	72.70			

that is, participant stayers and participant movers. For the stayers, as car ownership increases, the households are more likely to participate more frequently in these activities. The movers provide the exact opposite relationship between car ownership and activity frequency (but with loss of significance). Higher-income households tend to have higher frequencies, and the presence of young children inhibits participation in these activities (presumably to school and college). As expected, as household size increases, subsistence frequency also increases. Household size may also capture the effect of employed people in the household. In the OLS model its associated coefficient is unity; this was increased by 25 percent when the regression was corrected for selectivity. In Equation 3b a variable X influences Y in two ways: directly via its associated β and indirectly through the correction terms (λ's), and this explains the difference between the two models. The significance of the γ's indicates substantial selectivity bias for the participant stayers, whereas this is not true for the participant movers.

The maintenance activity frequency provides similar indications to the subsistence model. An exception to this is the effect of income. It appears that lower-income households are more likely to engage in this type of activity than higher-income households. One correction is significant for the participant-stayer model, and none is significant for the participant-mover model. Evidence of selectivity is present or absent depending on the type of frequency examined. This is even clearer when one examines the results in the leisure frequency models. None of the correction terms is signifi-

TABLE 3 Activity Regression Models

	SUBSISTENCE ACTIVITY FREQUENCIES				MAINTENANCE ACTIVITY FREQUENCIES				LEISURE ACTIVITY FREQUENCIES			
	OLS (without correction)		Tunali method (with correction)		OLS (without correction)		Tunali method (with correction)		OLS (without correction)		Tunali method (with correction)	
	Coef.	"t-stat"	Coef.	"t-stat"	Coef.	"t-stat"	Coef.	"t-stat"	Coef.	"t-stat"	Coef.	"t-stat"
Participants and stayers												
Constant	1.547	3.157	2.342	2.497	2.836	4.675	3.782	4.159	1.791	3.650	2.093	3.239
ONECAR2	-1.308	-2.590	-1.835	-2.791	-.724	-1.158	-.662	-.979	-.925	-1.830	-.908	-1.716
TWOCARS2	-1.145	-2.297	-1.742	-2.641	-.190	-.308	-.297	-.433	-1.009	-2.021	-1.045	-1.944
MULTICARS2	-.220	-.422	-.758	-1.118	-.194	-.301	-.442	-.621	-.462	-.886	-.543	-.969
HHLDSIZE2	1.005	9.255	1.252	8.751	1.372	10.204	1.251	7.983	1.322	12.161	1.285	10.282
KID(0-5)2	-1.151	-5.379	-1.341	-4.869	-.657	-2.481	-.317	-1.049	-1.396	-6.517	-1.289	-5.750
MIDINCOME2	.949	2.991	.773	2.342	-.287	-.730	-.103	-.261	.243	.764	.300	.940
HIGHINCOME2	2.208	6.438	2.106	5.710	-.091	-.215	.046	.108	.728	2.120	.771	2.245
R^2		*.159*		*.173*		*.122*		*.141*		*.153*		*.156*
λ_1			-6.415	-2.296			6.961	2.490			2.175	1.330
λ_2			-4.779	-2.148			.946	.465			.281	.204
ρ_{13}			-1.693				2.135				.599	
ρ_{23}			- .958				-.388				-.122	
σ_3			2.914				3.122				3.488	
Participants and movers												
Constant	-.306	-.189	-6.603	-.721	3.729	3.453	-7.806	-.585	4.090	3.617	-8.393	-.609
ONECAR2	-.592	-.390	-.015	-.008	- .843	- .833	.345	.138	- .858	- .810	.525	.205
TWOCARS2	- 906	-.562	1.756	.759	-2.725	-2.534	-.939	-.338	-1.120	-.994	.986	.346
MULTICARS2	-2.293	-1.263	-1.226	-.439	-1.222	-1.010	1.024	.291	-.060	-.047	2.708	.748
HHLDSIZE2	1.315	2.642	.906	1.180	.989	2.981	.226	.241	.684	1.969	-.151	-.155
KID(0-5)2	-.839	-1.055	-.450	-.371	.654	1.233	1.318	.816	-.305	-.548	.376	.227
MIDINCOME2	2.056	1.668	2.493	1.539	-1.964	-2.389	-1.019	.543	-2.051	2.383	-.919	-.482
HIGHINCOME2	4.507	3.003	4.724	2.341	-.531	-.531	.099	.040	-1.025	-.978	-.168	-.066
R^2		*.199*		*.212*		*.249*		*.330*		*.138*		*.239*
λ_1			2.782	.689			4.720	.772			4.826	.763
λ_2			6.242	.585			13.105	.847			15.438	.961
ρ_{13}			0.152				.080				.005	
ρ_{23}			0.964				1.410				1.484	
σ_3			5.579				8.260				9.393	

cant in these models. The higher standard error of the regression equation for all the models of the participant movers indicates higher variation in activity participation when compared with the stayers. The possibility of this functioning as an indicator of misspecification is discarded mainly because of the lack of significance of the correction terms. Table 4 presents a model with dependent variable the sum of the three activities in Table 3. The results parallel the indications of the subsistence models (signs of coefficients and relative magnitude). In general, the coefficients are higher because of the higher values of the dependent variable. Unlike the subsistence model, the correction terms are not significant. This leads to the conclusion that the effects of selectivity can be better captured by considering frequency of activity types separately. A refinement of the method here is under way using more complete specifications for the regression models, for example, incorporating transportation system attributes and better descriptors of household composition. In addition, a sensitivity analysis of the method to the specification of the bivariate probit model is also needed. The residential relocation model needs to consider additional determinants of relocation. This is also left as a future task.

SUMMARY AND CONCLUSIONS

A method to account for the possible simultaneity of multiple selection in panel surveys is presented in this paper. Two sources of selectivity are considered together—residential relocation and panel attrition—using a bivariate probit model that considers the lack of observed residential relocation for the sample of the dropouts. The method can be applied to derive sample weights for subsequent panel waves and to create correction terms that can be used to obtain consistent estimates of activity participation equations.

In the first two waves of PSTP, residential relocation and attrition are not correlated. This supports the use of sequential weighting for the Wave 2 sample. The application of correction terms to regression models of activity participation provided many insights. The effect of selectivity on activity participation may depend strongly on the type of activity analyzed. When all the activity types are aggregated to form a single model of frequencies (e.g., a trip generation model) selectivity bias may appear to be absent. When activities are considered separately, selectivity bias is present in some equations.

Many extensions and improvements are needed in the method presented here. The models need to be specified in radically different ways and the results need to be compared with those from this paper. This will provide some guidance on the effects of misspecification on selectivity equations. The Tunali method provides consistent estimates but is not fully efficient. Efficiency loss is associated with the two steps involved. A full information maximum likelihood method would be a suitable alternative. The three activity equations in Table 3 have been considered separately. It is well known that participation in one type of activity influences participation in another. This can be easily modeled by creating a system of equations and applying the Tunali method to the system. During an earlier review it was suggested that potential improvements in the method here may emerge from alternate forms of the activity frequency equations. One could extend the method using a system of "TOBIT" models with

TABLE 4 Sum of Activity Frequencies (Trip Generation)

	OLS (without correction)		Tunali method (with correction)	
	Coef.	"t-stat"	Coef.	"t-stat"
Participants and stayers				
Constant	6.174	5.653	8.217	5.754
ONECAR2	-2.957	-2.627	-3.404	-2.898
TWOCARS2	-2.344	-2.110	-3.085	-2.578
MULTICARS2	-.876	-.754	-1.744	-1.397
HHLDSIZE2	3.699	15.283	3.788	13.517
KID(0-5)2	-3.203	-6.719	-2.947	-6.088
MIDINCOME2	2.905	1.280	.970	1.361
HIGHINCOME2	2.845	3.722	2.923	3.797
R^2		*.243*		*.249*
λ_1			2.722	.788
λ_2			-3.553	-1.160
ρ_{13}			.454	
ρ_{23}			-.522	
σ_3			8.426	
Participants and movers				
Constant	7.512	2.837	-22.802	-.656
ONECAR2	-2.294	-.925	.855	.132
TWOCARS2	-2.939	-1.114	1.802	.251
MULTICARS2	-3.456	-1.164	2.505	.275
HHLDSIZE2	2.989	3.673	.981	.403
KID(0-5)2	-.490	-.377	1.244	.297
MIDINCOME2	-1.958	-.972	.555	.115
HIGHINCOME2	2.951	1.203	4.655	.726
R^2		*.223*		*.319*
λ_1			12.328	.773
λ_2			34.785	.864
ρ_{13}			.072	
ρ_{23}			1.428	
σ_3			21.693	

double "Probit" selectivity. With respect to model specification, the method here can be improved by the inclusion of level-of-service variables that are currently created for PSTP. In addition, for the relocation model a more in-depth specification analysis is needed.

ACKNOWLEDGMENTS

The authors thank the Puget Sound Regional Council for providing the data. Funding was provided by the U.S. Department of Transportation Region 3 Mid-Atlantic Universities Transportation Center. Comments by anonymous reviewers are also acknowledged.

REFERENCES

1. Heckman, J. J. 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,* Vol. 5, No. 4, 1976, pp. 475–492.
2. Heckman, J. J. Sample Selection Bias as a Specification Error. *Econometrica,* Vol. 47, 1979, pp. 153–161.
3. Mannering, F. L. Selectivity Bias in Models with Discrete and Continuous Choice: An Empirical Analysis. In *Transportation Research Record 1085,* TRB, National Research Council, Washington, D.C. 1985, pp. 58–62.
4. Kitamura, R., and P. H. L. Bovy. Analysis of Attrition Biases and Trip Reporting Errors for Panel Data. *Transportation Research A,* Vol. 21, 1987, pp. 287–302.
5. Hensher, D. A., P. O. Bernard, N. C. Smith, and F. W. Milthorpe. *Modeling the Dynamics of Car Ownership and Use: A Methodological and Empirical Synthesis.* Working Paper 32. Transport Research Group, Macquarie University, Sydney, Australia, 1987.
6. Monzon, J., K. G. Goulias, and R. Kitamura. Trip Generation Models for Infrequent Trips. In *Transportation Research Record 1220,* TRB, National Research Council, Washington, D.C. 1989, pp. 40–46.
7. Maddala, G. S. *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, England, 1983.
8. Murakami, E., and W. T. Watterson. Developing a Household Travel Panel Survey for the Puget Sound Region. In *Transportation Research Record 1285,* TRB National Research Council, Washington, D.C., 1990, pp. 40–46.
9. Kitamura, R., R. M. Pendyala, and K. G. Goulias. Weighting Methods for Choice-Based Panels with Correlated Attrition and Initial Choice. In *Transportation and Traffic Theory* (C. F. Daganzo, ed.), Elsevier Science Publishers, Amsterdam, The Netherlands, 1993, pp. 275–294.
10. Amemiya, T. *Advanced Econometrics.* Harvard University Press, Cambridge, Mass, 1985.
11. Tunali, I. A General Structure for Models of Double-Selection and an Application to a Joint Migration/Earnings Process with Remigration. *Research in Labor Economics,* Vol. 8, Part B, 1986, pp. 235–282.
12. *LIMDEP User's Manual and Reference Guide,* Version 6., Econometric Software, Inc., New York, 1992.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.