Bus, Taxi, and Walk Frequency Models That Account for Sample Selectivity and Simultaneous Equation Bias

JESSE JACOBSON

A 2-year user-side subsidy experiment that provided the handicapped and the elderly with discounted coupons to be used on buses and taxis was conducted in a small northeastern metropolitan city. The effect of the user-side subsidy experiment on bus and taxi travel by the elderly population is described. As expected, the subsidy experiment increased the number of trips taken by bus and by taxi. Furthermore, able-bodied elderly persons who do not own automobiles and handicapped elderly persons who are either employed or students are more likely to purchase discounted bus coupons than the population of elderly persons as a whole. Also, the number of walk trips was not affected by the number of bus and taxi trips taken. Therefore, people who have participated in the subsidy program have enjoyed a net increase in mobility (in the form of additional bus and taxi trips) because bus and taxi trips have not simply replaced walk trips.

Starting in July 1978 and for 24 consecutive months thereafter, the U.S. Department of Transportation (DOT) conducted an experiment of user-side subsidies for public transportation in Lawrence, Massachusetts, a small metropolitan city north of Boston. A select group of individuals--the elderly (65 years and older) and the handicapped of all ages--was eligible to receive financial assistance in the form of a reduced bus fare (the regular bus fare for elderly and handicapped persons was $0.15, but only $0.01 if project coupons were used) and a 50 percent discount on taxi rides (the discount was limited to $1.25 per ride and $20 per month). To establish eligibility individuals were to register at a downtown office, which was also the only location where discount coupons for bus and taxi rides could be purchased.

In conjunction with the experiment, a sample of individuals who were eligible to receive the assistance was contacted and asked to report sociodemographic information and to record a diary of travel for May 1978 and May 1979 (before the experiment and during the tenth month of the experiment). Although the total sample included both elderly and transportation-handicapped persons, only the subsample of the elderly (handicapped and able-bodied persons) was selected for this study. From this group, 130 completed returns were available; 48 percent of these returns were from transportation-handicapped persons, and 40 percent of the returns were from individuals who chose to become project users.

The purposes of this paper are to measure the travel impact of the experiment on the elderly population and to understand the reasons that attracted some of the eligible population to purchase discounted coupons and to use bus and taxi for their travel.

There is a problem in measuring the impact of the project because the purchase of the discounted coupons is prompted by expected benefits and other exogenous factors that are not fully measurable. If the incidence of these factors was known, the variables that identify them could be used in the analysis. Unfortunately, these variables are often not known or measured; thus in this paper a method to represent their effect is presented.

In the following sections two models that measure bus and taxi trip frequency, and a model that measures the number of walk trips, are presented. The latter model is used to determine whether walk trips are being replaced by bus or taxi trips.

PROBLEM OF SELF-SELECTION TO TREATMENT

Although the goal of this research is to measure the effectiveness of the project in increasing travel mobility, it is recognized that the inevitable limitations of the data generate issues that the model has to deal with explicitly. This is so, in particular, because the choice of becoming a project user (i.e., registering in the project and purchasing the discount coupons) rests entirely on the individuals who participate in the survey. Therefore, a definition has to be found for the following dichotomous variable for individual t,

$$d_t = \begin{cases} 1 \text{ if individual purchases discounted coupons} \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

and for the following model of travel demand,

$$y_t = \beta' X_t + \delta d_t + \epsilon_t \quad (2)$$

where

y_t = number of trips taken by individual t;
β = column vector of coefficients;
X_t = column vector of independent variables;
δ = a scalar, which is the coefficient of the dichotomous variable d_t; and
ϵ_t = stochastic component of the model.

At first glance it would appear that δ would represent the effect of the project. However, those who became project users did so because, as a general rule, they expected their travel demand to be higher than otherwise, and those who chose not to become users did so because they did not expect their travel to increase by becoming users. In other words, the benefits that users derive from purchasing the discounted coupons are larger than the benefits foregone by nonusers. This implies that d_t and ϵ_t are correlated; thus the model of trip generation that was proposed could not be estimated either by ordinary regression or by conventional cross-classification, a method that assumes, much like ordinary regression, independently distributed stochastic components.

As mentioned previously, if it was possible to measure all the variables that determine project participation, the variables could be incorporated in the analysis explicitly. However, because some of these variables are unmeasured, it is necessary to consider d_t as being an endogenous variable. Thus the estimation of a model that recognizes this endogenicity, which is also called selectivity bias, is presented. The theoretical justification for such a model is straightforward, and the reader is referred to the extensive literature on the subject (1,2) for more detail.

BUS FREQUENCY MODEL

Purchase of discounted coupons for bus travel is clearly a major factor in the frequency with which individuals take bus trips. However, as discussed earlier, the use of the variable that represents the

observed purchase decision in the model could yield inconsistent estimates of the project effect because of the likely presence of sample selectivity. Accordingly, the bus frequency model is estimated by a two-stage procedure first proposed by Maddala and Lee (3). The procedure requires estimation of a probit model of the decision to purchase discounted bus coupons, and estimation of a model (which incorporates as an independent variable the expected value of the dependent variable of the probit) of bus trip frequency.

The probit model of purchase of discounted bus fares is estimated from data on the actual purchase of these fares in May 1979. The observed dependent variable of the model is equal to one if bus coupons were purchased (in May 1979) and zero otherwise. The probability of purchasing discounted bus coupons (i.e., the expected value of the dependent variable) is equal to $\Phi(\gamma' Z_t)$, where Z_t is a column vector of independent variables, γ is a column vector of coefficients, and $\Phi(.)$ is the cumulative of the standard normal distribution. The estimated coefficients (γ), together with some goodness-of-fit measures, are given in Table 1. Although the probit was formulated as a single-equation model, different coefficients were estimated for able-bodied and transportation-handicapped persons.

Table 1. Probit estimates of use of bus coupons.

User	Coefficient	Asymptotic t-Statistic
Able-bodied person		
Constant	-0.945	3.2
Zero automobiles in household	0.711	1.7
Bus trips in May 1978	0.0678	2.5
Transportation-handicapped person		
Constant	-0.856	3.1
Employed or student	2.09	2.9
Bus trips in May 1978	0.281	3.6

Note: Log-likelihood with estimated coefficients = -50.16, log-likelihood with constants alone = -74.86, log-likelihood ratio statistic (4 df) = 49.4, number of observations = 130, 85.4 percent of sample was correctly classified, 10.8 percent of sample was erroneously classified as nonuser, and 3.8 percent of sample was erroneously classified as user.

For able-bodied elderly persons, automobile ownership (a zero-one variable) was found to affect the purchase of bus coupons significantly, whereas for transportation-handicapped persons, the most important variable was that of employment and student status, again a zero-one variable. The log-likelihood ratio statistic is equal to 49.4, a value that allows rejection, with a large level of confidence, of the hypothesis of no effect of the independent variables.

The second-stage model--a limited dependent variable model of the number of bus trips--is estimated from bus trips reported in the May 1979 diary survey. As discussed earlier, instead of including a zero-one variable for actual coupon purchase (or nonpurchase), the probability of being a project user is included in this model, i.e., the expected value of the dependent variable from the probit model. This ensures that the coefficient for the bus coupon purchase variable is consistent because sample selectivity is accounted for. A single-equation specification is again used for the groups of able-bodied and transportation-handicapped persons. The estimated coefficients are given in Table 2.

To test the effectiveness of the program, further statistical tests are performed on the subsample of actual project users. Specifically, the expected number of bus trips of project users, had they been nonusers, is compared with the actual number of bus trips taken. Because the distribution of the number of trips is truncated normal, the probability that the expected number of bus trips (conditional on nonpurchase of the project coupons is lower than the actual number of bus trips) is written as $(X'\beta - \mu)/\sigma$, where β is a column vector of coefficients, X is a vector of independent variables, μ is the actual number of bus trips taken in May 1979, and σ is the standard deviation of the underlying nontruncated distribution of the stochastic component of the model. For the subsample of program users, this probability averages 80 percent, and the Pearson's P_λ is 252.90 with 70 df, a value that clearly permits rejection of the null hypothesis of no-project effect on bus travel. Note also that the mean number of bus trips for the individuals who purchased discounted bus coupons in May 1979 is 16.51, whereas the mean expected number of bus trips for the same individuals, had they been nonusers, is 4.70, a difference of approximately 12 monthly trips.

Table 2. Estimates of May 1979 bus trips (limited dependent variable model).

User	Coefficient	Asymptotic t-Statistic
Able-bodied person		
Constant	-6.07	2.5
Probability of being a user for individuals who are neither students nor employed	28.7	2.2
No. of bus trips in May 1978	0.791	1.9
Transportation-handicapped person		
Constant	-10.1	3.9
Probability of being a user	23.4	4.3
No. of bus trips in May 1978	0.630	4.2
σ	10.4	10.6

Note: $y^* = X'\beta + e$

$$y = \begin{cases} 0 & \text{if } y^* < 0.5 \\ y^* & \text{otherwise} \end{cases}$$

and log-likelihood with estimated coefficients = -270.36, log-likelihood with constants alone = -318.01, log-likelihood ratio statistic (4 df) = 95.3, and number of observations = 130.

TAXI FREQUENCY MODEL

The estimation of a probit model of taxi coupon purchases did not yield acceptable results. Specifically, standard statistical tests pointed to the low explanatory power of the model. Several different specifications of the probit model were tested, but those also met with little success. Although it would have been possible to investigate the failure of the probit formulation to yield a satisfactory model, doing so would have been beyond the scope of this research. As a consequence, the two-stage procedure adopted for the bus frequency model was replaced by a simpler model. This model, which measures the monthly taxi trips taken, includes as an independent variable the actual purchase (or nonpurchase) of taxi coupons in May 1979 (a zero-one variable) and not the expected value from a probit model.

It is recognized that the coefficient estimate of the coupon purchase variable will be biased because of its endogenicity. However, it should be mentioned that this endogenicity is expected to be much less severe in the taxi model than in the bus model, particularly because the subsidy is only 50 percent (versus 93 percent for bus trips) and it is more limited in availability (the maximum taxi subsidy is $1.25 per trip and $20.00 per month per person). Accordingly, although the model presented in the following paragraphs has some evident limitations, it was decided to include it in this paper for completeness.

The taxi frequency model, like the model for bus travel, is a limited dependent variable model. As for the previous model, the taxi trip rate cannot be negative, and 79 of the 130 persons in the sample (61 percent) did not take any taxi trips in May 1979. In addition to the zero-one variable for individuals who purchased taxi coupons, the number of household automobiles has, as expected, a significant effect on taxi trip frequency (see Table 3).

Table 3. Estimates of May 1979 taxi trips (limited dependent variable model).

Item	Coefficient	Asymptotic t-Statistic
Constant	-7.54	4.6
No. of household automobiles	3.48	2.1
No. of taxi trips in May 1978	0.812	8.5
Purchased taxi coupons (1 if yes, 0 otherwise)	9.10	5.3
σ	6.91	9.6

Note: $y^* = X'\beta + e$

$y = \begin{cases} 0 & \text{if } y^* < 0.5 \\ y^* & \text{otherwise} \end{cases}$

and log-likelihood with estimated coefficients = -199.56, log-likelihood with constant alone = -247.35, log-likelihood ratio statistic (3 df) = 95.6, and number of observations = 130.

To test the effectiveness of the program in increasing taxi travel, statistical tests identical to the ones used for the bus travel model are applied here. Specifically, Pearson's P_λ (which has a value of 167.92 for the subsample of the 30 individuals who are taxi coupon purchasers) allows rejection of the null hypothesis of no increase in taxi travel because of project participation. The analysis also indicates that the mean number of taxi trips taken in May 1979 by taxi coupon purchasers is 8.6, whereas the expected value conditional on nonpurchase is 3.45 taxi trips for the same group of individuals, a difference of approximately 5 trips per month.

WALK TRIPS FREQUENCY MODEL

Although vehicular trips in general, and bus and taxi trips in particular, increased as a result of the user-side subsidy, it was hypothesized that some of the new vehicular trips might have replaced what were formerly walk trips. To test this hypothesis a walk frequency model that includes bus and taxi trip frequency as explanatory variables is estimated.

Because bus and taxi trips are endogenous to the walk trips model (i.e., the models for each travel mode are part of a system of structural equations), it was decided to use the expected trip rates from the models presented in the previous two sections as instruments instead of using the observed trip rate for bus and taxi trips.

The specification chosen for the estimation is again a limited dependent variable model. As for the previous models, the walk trip rate cannot be negative, and 17 of the 130 persons in the sample (13 percent) did not take any walk trips in May 1979. The coefficient estimates for the model are given in Table 4. If bus and taxi trips were actually replacing potential walk trips, the coefficients of the frequency of bus and taxi trips would be negative (and statistically significant). The results, however, reveal these coefficients to be positive and not statistically different from zero, which indicates that the hypothesis of modal substitution is unlikely to be valid.

Table 4. Estimates of May 1979 walk trips (limited dependent variable model).

Item	Coefficient	Asymptotic t-Statistic
Constant	0.40	0.13
Expected no. of bus trips in May 1979	0.15	0.85
Expected no. of taxi trips in May 1979	0.43	1.3
No. of walk trips in May 1978	0.79	18.0
σ	20.7	15.0

Note: $y^* = X'\beta + e$

$y = \begin{cases} 0 & \text{if } y^* < 0.5 \\ y^* & \text{otherwise} \end{cases}$

and log-likelihood with estimated coefficients = -515.862, log-likelihood with constant alone = -602.458, log-likelihood ratio statistic (3 df) = 173.19, and number of observations = 130.

CONCLUSIONS

The models presented in this paper have confirmed quite strongly the a priori hypothesis regarding travel by bus, taxi, and walk. The large increases in bus and taxi travel observed in May 1979 by those individuals who purchased discounted coupons can be directly attributed to the project. Also, it was shown that the increase in bus and taxi trips was not achieved at the expense of walk trips. Rather, the additional bus and taxi trips were trips that would have not been taken in the absence of the subsidy project.

The data in Table 5 further confirm the findings of the models. Note in particular the increase (between 1978 and 1979) in bus trips for bus subsidy users (i.e., for those individuals who purchased bus coupons), and the increase in taxi trips for taxi subsidy users. These increases are much larger than the increases for the sample as a whole and for the subsample of nonusers of the program.

Table 5. Trip rates by mode and project participation status.

Mode	Month and Year	Project Participation Status: All Sample (n = 130)	Project Users, Taxi and Bus (n = 49)	Project Bus Users (n = 35)	Project Taxi Users (n = 30)
Bus	May 1978	3.52	7.18	9.97	8.27
	May 1979	6.22	12.82	16.51	11.77
Taxi	May 1978	2.43	3.69	3.83	5.33
	May 1979	3.22	5.57	4.37	8.6
Walk	May 1978	40.27	49.27	54.54	42.9
	May 1979	37.12	44.98	49.69	39.93
All modes	May 1978	109.05	100.00	103.63	100.67
	May 1979	106.69	99.61	104.14	99.37

Walk trips are mostly unaffected by program use, which confirms the findings of the model of walk trips. Note that only bus subsidy users take a larger number of walk trips than other groups.

ACKNOWLEDGMENT

This paper was prepared as part of an exploratory study of the travel behavior of the elderly and the handicapped; it was sponsored by the Service and Methods Demonstration program of UMTA. I wish to acknowledge various comments by Lawrence B. Doxsey of the Transportation Systems Center.

REFERENCES

1. B.S. Barnow, G.G. Cain, and A.S. Goldberg. Issues in the Analysis of Selectivity Bias. In Evaluation Studies Review Annual, Volume 5 (E.W. Stromsdorfer and G. Farkas, eds.), Sage Publications, Beverly Hills, Calif., 1980.
2. J. Heckman. Simultaneous Equation Models with Both Continuous and Discrete Endogenous Variables with or Without Structural Shift in the Equations. In Studies in Nonlinear Estimation (S.M. Goldfeld and R.E. Quandt, eds.), Ballinger Publishing Company, Cambridge, Mass., 1976.
3. G.S. Maddala and L.-F. Lee. Recursive Models with Qualitative Endogenous Variables. Annals of Economic and Social Measurement, Vol. 5, 1976, pp. 525-545.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Notice: This paper represents the views of the author and not necessarily those of U.S. Department of Transportation.

Effect of Sample Size on Disaggregate Choice Model Estimation and Prediction

FRANK S. KOPPELMAN AND CHAUSHIE CHU

Sampling error is one of several types of error in econometric modeling. The relationship between sampling error and sample size is well known for both estimation and prediction. The objective of this paper is to provide an empirical foundation for using these relationships to guide researchers and planners in the determination of sample size for model development. Analytic relationships are formulated for sample size, precision of parameter estimates, replication of parent population, and replication of an alternative (transfer) population. Application of these relationships to an empirical case indicates that the sample sizes required to obtain reasonably precise parameter estimates are substantially larger than the sample sizes generally considered to be needed for disaggregate model estimation. Nevertheless, these sample sizes appear to be adequate for obtaining reasonably accurate replication of observed choice behavior in the parent population. The corresponding results for prediction to a different population are complicated by the issue of intrapopulation transferability. Although the results reported in this paper should be validated in other contexts, it appears that accurate estimation requires the use of samples that are substantially larger than formerly believed. Samples on the order of 1,000 to 2,000 observations may be needed for estimation of relatively simple disaggregate choice models. Although some reduction in this requirement may be obtained by improved sample design, it is unlikely that the final sample requirements can be reduced to less than 1,000 observations.

Econometric model development is subject to errors in sampling, model specification, and measurement (1,2). In this paper the effect of sampling error is examined for model parameter estimates, prediction to the parent population, and transfer prediction to alternative populations. Sampling error can be avoided only by observation and analysis of the entire population. In practice, the resources needed to collect data for an entire population and to analyze such extensive data are not available. Thus there is concern with the magnitude of the errors that are introduced by use of samples of the population.

EXPECTED EFFECTS OF SAMPLE SIZE

The precision of parameter estimates for a given model structure depends on the estimation method used, the multidimensional distribution of the explanatory variables of the model, the range of observed behavior, the quality of model specification, and the sample size of the estimation data set. Maximum likelihood estimation obtains consistent estimators of the parameters of disaggregate choice models and provides estimates of the precision with which model parameters are estimated (3-5).

The relationship between parameter precision and sample size is well known. The variance-covariance matrix of estimated parameters in linear models is inversely proportional to sample size (3,6). The variance-covariance matrix of maximum likelihood estimated parameters for quantal choice models is asymptotically equal to the negative inverse of the Hessian of the log-likelihood function (3,7). The asymptotic expectation of this matrix is inversely proportional to sample size. Thus the error variance-covariance matrix for maximum likelihood estimations for quantal choice models is also inversely proportional to sample size.

Prediction accuracy describes how well the choice model replicates observed population behavior. Prediction performance of discrete choice models is a function of the validity of model theory, the validity of the derived model structure, the quality of model specification, the quality of variable measurement and prediction, and the accuracy of estimated parameters (8). As noted earlier, precision of model parameter estimates is proportional to sample size. It follows that the portion of prediction error attributable to errors in parameter estimation is inversely proportional to sample size. Specifically, the expected squared prediction error caused by errors in parameter estimates is inversely proportional to sample size (5, p. 189). Models estimated from large samples are more likely to accurately describe the behavioral process in the general population, and consequently such models will have satisfactory prediction performance. Thus it is expected that increased sample size in model estimation will yield improved prediction precision. When excessively small samples are used, both parameter estimates and parent population predictions will be highly variable.

Transferability of disaggregate discrete choice models is based on the argument that choice models describe the underlying behavioral response mechanisms or decision rules of decision makers in the selection among available alternatives (9,10). If the behavioral response or decision rules of decision makers is constant across contexts, models that describe this behavior will be transferable. Koppelman and Wilmot (11) define transferability of choice models as "the degree of success with which