Selectivity Bias in Models of Discrete and Continuous Choice: An Empirical Analysis

FRED L. MANNERING

In this paper is discussed an application of a recently developed econometric technique for correcting selectivity bias in discrete and continuous modeling systems with multiple discrete choices. The case studied is the household's choice of type of vehicle to own and the extent to which it is utilized. An appropriate model structure is formulated, and vehicle utilization equations that do and do not account for selectivity bias are estimated. The empirical results strongly underscore the importance of proper econometric treatment of discrete and continuous modeling systems.

In recent years there has been considerable research examining behavioral choices that involve jointly determined discrete and continuous components. One of the primary factors motivating such research is the relatively frequent occurrence of discrete and continuous interrelationships in actual choice situations. Typical examples include the choice of durable goods, such as an appliance or automobile (i.e., discrete), and the extent to which it is used (i.e., continuous), occupation choice and resulting income, union participation and earnings, and the choice of freight mode and quantity shipped.

From an econometric standpoint, the modeling of discrete and continuous choices presents a number of interesting implications. Perhaps the most important is the sample selection or selectivity bias that will be present in the continuous equations. To address this problem, Heckman *(1-3)*, Schmidt and Strauss *(4)*, Westin and Gillen *(5)*, Duncan *(6)*, and McFadden and Winston *(7)* have all developed or applied, or both, corrective econometric techniques. Essentially, such corrective methods involve the joint estimation of a discrete model, traditionally derived from random utility theory (e.g., probit or logit), and a continuous model, normally estimated by regression procedures. Unfortunately, most existing studies on this subject are based on econometric methods that make extensions beyond the consideration of simple binary discrete choices exceedingly difficult, if not impossible.

In this paper is demonstrated, by empirical example, a recently derived method *(8, 9)* of correcting selectivity bias in discrete and continuous modeling systems with multiple discrete choices. The objective here is not to develop new econometric theory or methods but to make recent econometric contributions in the area of selectivity bias more widely known.

SELECTIVITY BIAS PROBLEM

Virtually any modeling system that has interrelated discrete and continuous choices will exhibit selectivity bias in its estimation of continuous equations. To illustrate this, the household decision of type of vehicle to own and the extent to which it is utilized will be assessed. This particular decision process has recently been recognized as a classic example of interrelated discrete and continuous choice *(10–12)*.

For the purposes of this paper, consider households owning only one vehicle and choosing among three types of vehicles that are defined as *(a)* vehicles with less than 25,000 accumulated mileage, *(b)* vehicles with accumulated mileage between 25,000 and 60,000 mi, and *(c)* vehicles with accumulated mileage in excess of 60,000 mi. These mileage classifications are loosely based on observed utilization behavior and annual maintenance and repair costs, although a rigorous statistical classification along these lines would be a preferred approach. Also, by restricting the analysis to households owning only one vehicle, the problems associated with assigning utilization among household vehicles is avoided. For models that explicitly address this multivehicle utilization assignment problem see the work of Mannering *(13)* and Greene and Hu *(14)*.

Intuitively, a basic understanding of the selectivity bias problem, as it relates to vehicle usage, can be gained by noting that usages are observed only for the vehicle type actually selected by the household, and no information is available on the extent to which other vehicle types would have been used by the household had they been selected. As an example of the selectivity bias that can result from such a problem, consider Figure 1. Let Line a represent a least squares estimation of an equation that is defined for the low accumulated mileage vehicle type, assuming that usage data on the low-mileage vehicle type is available for all households. In reality, the observed sample of low-mileage vehicle type users is composed primarily of high-usage households that have selected the low accumulated mileage vehicle type for reasons such as the pleasure of driving a newer automobile and the need for vehicle reliability. Therefore, using the realistic observed sample, a least squares equation estimation will produce bias parameter estimates as reflected by Line b of Figure 1.

To formalize this selectivity bias from an econometric standpoint, it is necessary to consider the problem as a correlation of residuals. To illustrate this, consider a utility-maximizing modeling system defined for each household (j) such that, for discrete choice of vehicle type,

$$U_{ij} = \theta_i Z_{ij} + \varepsilon_{ij} \quad (1)$$

where

U_{ij} = total indirect utility provided by vehicle type i,
Z_{ij} = a vector of household and vehicle type attributes,
ε_{ij} = a disturbance term accounting for unobserved effects, and
θ_i = a vector of estimable parameters.

For continuous choice of household vehicle utilization,

$$y_{ij} = \beta_i X_j + \nu_j \quad (2)$$

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

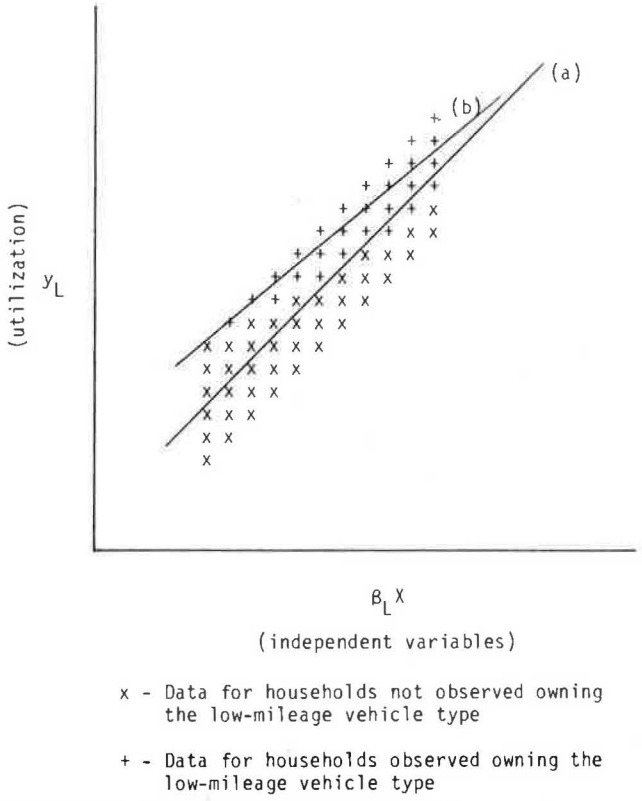

FIGURE 1 Illustration of the selectivity bias problem.

where

y_{ij} = household utilization of vehicle type i (e.g., miles per year);
X_j = vector of household socioeconomic conditions;
ν_j = unobserved characteristics of the household; and
β_i = a vector of estimable parameters that vary across vehicle types (i's).

Econometrically, the problem of selectivity bias arises because correlation is likely to be present between the unobservables ε_{ij} and μ_j [i.e., $E(\varepsilon_{ij}\, \nu_j) \neq 0$]. For example, the unobserved effects that tend to increase usage (e.g., household's value of driving pleasure) will adversely affect the probability of owning a vehicle that has high accumulated mileage because such a vehicle is likely to be decrepit and therefore provide little driving pleasure. Similarly, the unobserved effects that increase the utility of selecting a low-mileage vehicle (e.g., need for reliability in travel) are likely to be associated with a higher degree of vehicle usage.

If correlation of error terms does exist, estimation of Equation 2 by ordinary least squares (OLS) will produce biased and inconsistent parameter estimates. This follows because such correlation implies that households observed to own specific vehicle types (i.e., low accumulated mileage) may indeed be a nonrandom sample that is censored by usage, a dependent variable.

Alternatively, estimating a single vehicle utilization equation with dummy variables that indicate the choice of vehicle type may be considered. However, this too would result in biased and inconsistent parameters because the dummy variables would most assuredly be correlated with the disturbance term (e.g., a low-mileage car is likely to have a high utilization disturbance term). To effectively correct for selectivity bias, it is necessary to estimate vehicle utilization equations as part of a joint model that includes the discrete choice of vehicle type.

Econometric Methods

To resolve the selectivity bias problem and arrive at consistent estimates of the parameters that comprise the utilization equations, Equation 2 is written as

$$E(y_{ij}|\, i) = \beta_i X_j + E(\nu_j|\, i) \tag{3}$$

where $E(y_{ij}|\, i)$ is the utilization conditional on the choice of vehicle type i; $E(\nu_j|\, i)$ is the conditional unobserved household characteristics; and other terms are as previously defined.

The estimation of Equation 3 will provide bias-corrected and consistent estimates of the parameter vector (β) because the selectivity bias induced by the nonrandom observed utilization samples is explicitly accounted for by the conditional expectation of ν_j [i.e., $E(\nu_j|\, i)$]. The problem in estimating Equation 3 becomes one of obtaining a closed-form representation of $E(\nu_j|\, i)$ that can be used in equation estimation. Such a closed-form representation has been derived by both Hay *(8)* and Dubin and McFadden *(9)*, on the assumption that the discrete choice can be represented by a multinomial logit model. The general form of the derivation is as follows (suppressing the household subscripting for notational convenience).

Let γ denote the vector of discrete choice disturbance terms $(\varepsilon_1, \varepsilon_2, \ldots \varepsilon_K)$ where K is the total number of alternatives. It follows that the expectation of ν conditional on the choice of i can be written as

$$E(\nu|\, i) = (1/P_i) \int_{\gamma|i} E(\nu|\, \gamma) \prod_{k=1}^{K} f(\varepsilon_k)\, d_\gamma \tag{4}$$

where P_i is the probability of selecting Alternative i. If it is assumed that γ is generalized extreme value distributed with σ^2 denoting the unconditional variance of ν and ρ_i denoting the correlation of ν and the resulting discrete choice logistic error terms (i.e., $\varepsilon_i - \varepsilon_k$'s), Hay *(8)* has demonstrated [see also Dubin and McFadden *(9)* and Dubin *(15)*] that Equation 4 can be written as

$$E(\nu|\, i) = (-1)^{K+1} (\sigma\, 6\, \rho_i/\pi^2) \left((1/K) \sum_{k \neq i}^{K} \{[P_k \ln P_k/(1 - P_k)] + \ln P_i\}\right) \tag{5}$$

Thus the selectivity bias correction becomes a simple ratio of the multinomial logit discrete choice probabilities.

Therefore, to obtain utilization equations free from selectivity bias, the following estimation procedure is used:

1. Estimate a multinomial logit model of vehicle type choice.
2. Use the predicted type choice probabilities (P_{kj}'s) to arrive at consistent estimates of the selectivity bias correction represented by the portion of Equation 5 in large parentheses.
3. With the values from Step 2, estimate Equation 3 by OLS; note here that the $\sigma\rho_i/\pi^2$ term in Equation 5 becomes an estimable OLS parameter.

An empirical application of the procedure is presented later.

Model Structure

Before proceeding with the actual estimation, it is necessary to provide a more detailed representation of the model structure than that given in Equations 1 and 2. To begin, let the household's vehicle choice indirect utility function be

$$U_{ij} = \theta_i Z_{ij} + \phi\, y_{ij} + \varepsilon_{ij} \quad (6)$$

where y_{ij} is the household utilization of vehicle type i; ϕ is an estimable parameter; and other terms are as defined for Equation 1. The inclusion of y_{ij} in the indirect utility function reflects, logically, the significance of vehicle utilization in determining the choice of vehicle type. However, because y_{ij} is endogenous to the type choice process and observed for only the chosen vehicle type, the reduced form of the indirect utility function must be used such that (by substitution of Equation 2),

$$U_{ij} = \theta_i Z_{ij} + \phi\, \beta_i X_j + \phi\, \nu_j + \varepsilon_{ij} \quad (7)$$

If the ε_{ij}'s are generalized extreme value distributed, then the type choice probabilities are given by the standard multinomial logit model

$$P_{i'j} = e^{V_{i'j}} / \sum_i e^{V_{ij}} \quad (8)$$

where $P_{i'j}$ is the probability of household j selecting vehicle type i' and V_{ij} is the deterministic component of the indirect utility U_{ij} (i.e., all terms except ε_{ij}). Note here that because the error term $\phi\nu_j$ (see Equation 7) does not vary across vehicle alternatives (i's), it will not affect the assumption of a logit model structure.

With this model structure, it follows that the selectivity bias corrected utilization equations are of the estimable form,

$$y_{ij} = \beta_i X_j + \alpha_j \lambda_{ij} + \eta_i \quad (9)$$

where

$\alpha_i = (-1)^{K+1}\,(\sigma\, 6\rho_i/\pi^2)$, a parameter estimable by OLS,

and

$$\lambda_{ij} = (1/K) \sum_{k \neq i}^{K} \{[P_{kj} \ln P_{kj}/(1 - P_{kj})] + \ln P_{ij}\}$$

It is important to note here that, to improve exposition of forthcoming empirical results, ρ_i is not included in the summation over K. This exclusion implies an equality restriction, across choice alternatives, of the correlation of ν and $\varepsilon_i - \varepsilon_k$'s. Relaxation of this restriction complicates the model structure by making additional parameter estimation necessary (i.e., a total of $K - 1$ α's must be estimated for each i). However, it has been empirically demonstrated that such a restriction of ρ_i is not unreasonable (*8*).

As described earlier, OLS estimates of Equation 9 will result in consistent parameter estimates. An empirical demonstration is presented in the following section.

Estimation Results

As described earlier, the empirical analysis considers single-vehicle households that own low, medium, or high accumulated mileage vehicles. The data used for model estimation consisted of a 364-household sample collected by the U.S. Department of Energy in the spring of 1980. Utilization of the vehicles owned by these households is defined as the mileage accumulated over the 6-month period from January 1980 to June 1980. Of these 364 households, 81, 185, and 98 owned low-, medium-, and high-mileage vehicles, respectively. A complete list of variables used in subsequent estimations, along with the corresponding means and standard deviations associated with the household sample, is given in Table 1.

Turning first to the estimation of the multinomial logit vehicle type choice model, it was found that vehicle capital cost, age of the head of the household, household income, residential location, number of household members, and number of drivers in the household all entered the reduced form indirect utility function as specified by Equation 7. The resulting coefficient estimates are given in Table 2.

TABLE 1 VARIABLES USED IN MODEL ESTIMATION

Variable	Mnemonic	Mean		
		Low Mileage	Medium Mileage	High Mileage
No. of miles driven in 6-month period (dependent variable in utilization equations)	y	4908.1 (3220.8)[a]	4681.7 (4706.6)	2836.3 (2670.9)
Age of household head (yr)	AGE	45.0 (17.9)	47.5 (18.8)	52.6 (20.5)
Location indicator (1 if SMSA, 0 otherwise)	SMSA	0.77 (0.43)	0.68 (0.47)	0.66 (0.48)
No. of drivers in household	NDRIV	1.52 (0.62)	1.43 (0.66)	1.34 (0.61)
Household income ($/yr)	INC	18,723 (11,800)	15,520 (12,517)	11,396 (8,112)
Vehicle capital cost ($)/income ($/yr)	COST	0.809 (0.826)	0.404 (0.413)	0.116 (0.118)
No. of household members	NMEM	2.40 (1.21)	2.33 (1.34)	2.35 (1.47)
Selectivity bias correction[b] (λ)	SBC	−1.31 (0.30)	−0.75 (0.077)	−1.15 (0.354)

[a]Standard deviations in parentheses.
[b]As defined in Equation 9.

TABLE 2 MULTINOMIAL LOGIT VEHICLE TYPE CHOICE ESTIMATES

Variable	Estimated Coefficient	t-Ratio
Cost (Alt1, Alt2, Alt3)	−0.835	−2.16
AGE (Alt1)	−0.019	−2.13
AGE (Alt2)	−0.04	−1.89
INC (Alt1)	3.23 E-05	1.56
INC (Alt2)	2.47 E-05	1.40
SMSA (Alt1)	0.296	0.83
SMSA (Alt2)	−0.108	−0.39
NMEM (Alt1)	−0.209	−1.49
NMEM (Alt2)	−0.159	−1.40
NDRIV (Alt1)	0.253	0.87
NDRIV (Alt2)	0.111	0.46
Constant (Alt1)	0.725	0.82
Constant (Alt2)	1.551	2.26

Note: Alt = alternative, Alt1 = low mileage (less than 25,000 accumulated miles), Alt2 = medium mileage (25,000 to 60,000 accumulated miles), and Alt3 = high mileage (more than 60,000 accumulated miles). Variable coefficient value is defined only for those alternatives listed in parentheses and is zero for nonlisted alternatives. Number of observations = 364; log-likelihood at zero = −339.9 and at convergence = −327.4.

The estimates are of plausible sign and reasonably significant statistically. It is interesting to note the importance of the age and income variables in determining vehicle type choice. These results reflect a propensity, as expected, of young, high-income households to own low-mileage vehicles. Also, the vehicle capital cost variable produced an anticipated strong negative effect on vehicle type choice probabilities.

It is important to note that the model presented in Table 2 represents the best specification obtained from a large number of estimation trials. Other model specifications tested resulted in substantially lower *t*-statistics or lower log-likelihoods at convergence, or both. On the basis of this assessment, the model presented herein can be presumed to be fairly well specified. The issue of proper specification is potentially important because a seriously misspecified model can lead to erroneous interpretations of the magnitude of selectivity bias (*8*).

Given the specification of the type choice model presented in Table 2, the selectivity bias correction can be readily determined (see Equation 9).The means and standard deviations of the resulting selectivity bias corrections are given in Table 1. It is interesting to point out that the selectivity bias term in Equation 9 (see also Equation 4) will be higher (in absolute value) for a specific alternative when the probability of a household selecting that alternative is lower. This is intuitively reasonable because it is expected that the estimated coefficients (the β_i vector) will require greater correction when the household's vehicle selection probability is low.

With the calculated selectivity bias terms in hand, utilization equations for low, medium, and high accumulated mileage vehicles can be estimated. The estimations include the natural log of utilization as the dependent variable with head of household's age, residential location, number of household drivers, and income as independent variables. The natural logarithm of utilization was used to improve statistical fit and to avoid the possibility of negative predicted values. Note that, by using $\ln y_{ij}$ in place of y_{ij} in all equations, the modeling system represented by Equations 7–9 is still valid. Also, it should be mentioned that the independent variables incorporated in this model are similar to those used by Mannering (*12, 13*) and Greene and Hu (*14*). Previously, however, only Greene and Hu estimated different coefficients for different vehicle types, and their work ignores possible selectivity bias. Finally, note that the utilization equations do not contain any vehicle-specific attributes. As demonstrated by Heckman (*2*, pp.935–936), this is a necessary condition for the existence of the model as defined by Equations 7–9.

Table 3 gives the estimation results for equations not corrected for selectivity bias and for those that are corrected for selectivity bias. The results indicate that the age variable has a strong negative effect on vehicle utilization in all equations. In addition, residential location and the number of household drivers were found to provide generally reasonable but less significant coefficient estimates. The income variable was generally not significant, and this was particularly true for those equations that were corrected for selectivity bias. Overall, the differences between corrected and uncorrected coefficients, though noticeable, are not large when the corresponding standard errors are considered. It is speculated that larger sample sizes would produce more statistically significant differences between corrected and uncorrected model coefficients.

Turning specifically to the selectivity bias terms, it is notable that the selectivity bias coefficients were highly significant in the medium- and high-mileage utilization equations. This is a strong

TABLE 3 UTILIZATION EQUATION ESTIMATES, UNCORRECTED AND CORRECTED FOR SELECTIVITY BIAS

	Low Mileage		Medium Mileage		High Mileage	
Variable	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
AGE	−0.014	−0.012	−0.017	−0.012	−0.016	−0.011
	(−3.27)	(−2.13)	(−4.95)	(−3.26)	(−3.18)	(1.87)
SMSA	−0.186	−0.287	−0.267	−0.035	−0.255	−0.234
	(−1.01)	(−1.36)	(−1.90)	(−0.230)	(−1.25)	(−1.15)
NDRIV	0.187	0.151	0.102	0.125	0.066	−0.004
	(1.42)	(1.11)	(1.02)	(1.28)	(0.03)	(0.02)
INC	−6.75 E-06	−1.35 E-06	1.07 E-05	3.11 E-06	1.87 E-05	−5.56 E-06
	(−0.98)	(−0.37)	(1.99)	(0.55)	(1.44)	(−0.26)
Constant	8.91	9.64	8.74	11.15	8.31	7.42
	(25.54)	(11.48)	(32.61)	(14.92)	(18.94)	(9.97)
SBC		0.467		3.66		−0.784
		(1.06)		(3.44)		(−1.65)
R^2	.175	.185	.168	.220	.159	.178
Corrected R^2	.131	.131	.150	.198	.123	.134
No. of observations	81		185		98	

Note: Dependent variable is the natural log of miles driven in 6 months (*t*-statistics in parentheses).

TABLE 4 PERCENTAGE DIFFERENCE BETWEEN CORRECTED AND UNCORRECTED HOUSEHOLD VEHICLE UTILIZATION EQUATION PREDICTIONS

	Low Mileage	Medium Mileage	High Mileage
Average of all households	5.9	13.4	9.9
Standard deviation of all households	9.1	14.0	7.9
Highest difference	53.2	82.1	47.2

indication that sample selectivity bias is present in such equations because the null hypothesis of no selectivity bias can be tested directly using the t-statistics of the estimated selectivity bias coefficient (i.e., under the null hypothesis of no selectivity bias, the selectivity bias coefficient would be zero). The selectivity bias coefficient in the low-mileage utilization equation, although of reasonable magnitude, was not highly significant. This may have been caused by the relatively low number of observations or the diversity of households owning low-mileage vehicles, or both. As a final point, it is important to note that the standard errors of the coefficient estimates presented in Table 3 will be biased downward, to some extent, because estimates of the discrete choice probabilities are used in the selectivity bias correction term as opposed to the true probabilities. This obviously results in an upward bias in reported t-statistics.

To demonstrate the potential importance of selectivity bias in household vehicle utilization equations, utilization was predicted for low-, medium-, and high-mileage vehicles for all 364 households, using both corrected and uncorrected equations. A summary of the results is given in Table 4. The data in this table indicate that, by ignoring selectivity bias, substantial errors can be introduced into predictions of household vehicle utilization. In the case presented here, it was found that as much as an 82 percent difference between predicted values can result.

CONCLUSION

As demonstrated by the empirical analysis undertaken in this paper, the problem of selectivity bias in the continuous equations of discrete and continuous modeling systems can have significant consequences. Specifically, if selectivity bias is ignored, substantial errors in equation estimation may result. It is therefore essential that proper econometric treatment be given to discrete and continuous modeling systems to ensure the credibility of resulting parameter estimates.

REFERENCES

1. J. Heckman. 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.
2. J. Heckman. Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, Vol. 46, No. 6, 1978.
3. J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, Vol. 47, No. 1, 1979.
4. P. Schmidt and R. Strauss. The Effect of Unions on Earnings and Earnings on Unions: A Mixed Logit Approach. *International Economic Review*, Vol. 17, No. 1, 1976.
5. R. Westin and D. Gillen. Parking Location and Transit Demand: A Case Study of Endogenous Attributes in Disaggregate Model Choice Models. *Journal of Econometrics*, Vol. 8, 1978.
6. G. Duncan. Formulation and Statistical Analysis of Mixed, Continuous/Discrete Dependent Variable Model in Classical Production Theory. *Econometrica*, Vol. 48, No. 4, 1980.
7. D. McFadden, C. Winston, and A. Boersch-Supan. "Joint Estimation of Freight Transportation Decisions Under Non-Random Sampling." In *Analytic Studies in Transport Economics* (A. Daugherty, ed.), Cambridge University Press, New York, 1986.
8. J. Hay. *Occupational Choice and Occupational Earnings*. Ph.D. dissertation. Yale University, New Haven, Conn., 1980.
9. J. Dubin and D. McFadden. An Econometric Analysis of Residential Electric Appliance Holdings and Consumption. *Econometrica*, Vol. 52, 1984.
10. K. Train and M. Lohrer. "Vehicle Ownership and Usage: An Integrated System of Disaggregate Demand Models." Presented at 62nd Annual Meeting of the Transportation Research Board, Washington, D.C., 1983.
11. F. Mannering and C. Winston. A Dynamic Empirical Analysis of Household Vehicle Ownership and Utilization. *The Rand Journal of Economics*, Vol. 16, No. 1, 1986.
12. F. Mannering. A Note on Endogenous Variables in Household Vehicle Utilization Equations. *Transportation Research*, Vol. 20B, No. 1, 1986.
13. F. Mannering. An Econometric Analysis of Vehicle Use in Multivehicle Households. *Transportation Research*, Vol. 17A, No. 3, 1983.
14. D. L. Greene and P. S. Hu. "The Influence of the Price of Gasoline on Vehicle Use in Multivehicle Households." In *Transportation Research Record 988*, TRB, National Research Council, Washington, D.C., 1984, pp. 19–24.
15. J. Dubin. *Economic Theory and Estimation of the Demand for Consumer Durable Goods and their Utilization: Appliance Choice and the Demand for Electricity*. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, 1982.

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