Nonresponse Bias and Trip Generation Models

PIYUSHIMITA THAKURIAH, ASHISH SEN, SIIM SÖÖT, AND EDWARD CHRISTOPHER

There is serious concern over the fact that travel surveys often overrepresent smaller households with higher incomes and better education levels and, in general, that nonresponse is nonrandom. However, when the data are used to build linear models, such as trip generation models, and the model is correctly specified, estimates of parameters are unbiased regardless of the nature of the respondents, and the issues of how response rates and nonresponse bias are ameliorated. The more important task then is the complete specification of the model, without leaving out variables that have some effect on the variable to be predicted. The theoretical basis for this reasoning is given along with an example of how bias may be assessed in estimates of trip generation model parameters. Some of the methods used are quite standard, but the manner in which these and other more nonstandard methods have been systematically put together to assess bias in estimates shows that careful model building, not concern over bias in the data, becomes the key issue in developing trip generation and other models.

Nonresponse bias is a well-recognized problem in sample surveys. In the field of transportation, where much of the planning effort rests on estimates of parameters obtained from models, it becomes pertinent to discuss the manner and the extent to which nonresponse bias affects the quality of estimates.

There are two types of nonresponse:

1. *Total nonresponse* occurs if some individuals or households simply do not respond to the survey. Then bias could occur if the preferences, values, or behavior of the nonrespondents are different from those of the respondents on whom estimates are based.
2. *Item nonresponse* occurs if parts of the survey instrument are not completed. A particularly unpleasant form of this in travel surveys occurs when respondents forget to record all trips. The result is that some individual households appear to have taken fewer trips than they actually did.

Trip generation models, whether they are cross-classification models or continuous models, are typically linear models, and estimates can be viewed as least-squares estimates. Since this is not entirely obvious for cross-classification models, a proof is given in a later section. It is well known that if certain conditions are satisfied, least-squares estimates are unbiased.

P. Thakuriah, A. Sen, and S. Sööt, Urban Transportation Center, M/C 357, University of Illinois at Chicago, Reliable Building Annex, Suite 700 South, 1033 West Van Buren Street, Chicago, Ill. 60680. E. Christopher, Chicago Area Transportation Study, 300 West Adams Street, Chicago, Ill. 60606.

On the other hand, if the model does not satisfy these conditions, bias will occur even if there is a 100 percent response rate. These conditions are satisfied by the model if the functional form of the model is correct and all important explanatory variables are included.

Because categorical models do not have any problems with their functional form, and weighting and related issues are taken care of, the authors prefer categorical trip generation models. This preference is discussed in a later section. Therefore, the issue that remains when assessing bias in estimates from categorical trip generation models is whether the model includes all the relevant independent variables or at least all important predictors.

Methods of assessing bias caused by omitting important variables are demonstrated on a trip generation model of trip circuits constructed from data that were collected by a survey in Lake County, Illinois, in the northern part of the Chicago metropolitan area. Two types of approaches are available for this:

1. Try out different variables in addition to those included in the model. Since the variables chosen to include in the model were fairly standard, it should be expected that such analyses have been carried out by model builders in the past.
2. Use various diagnostics for correct model specification. The kind of diagnostic methods used and the results are shown later.

Yet another diagnostic tool is available in the case of categorical models. Since the dependent variable is counted, it is reasonable to expect that its values are approximately Poisson. This assumption is made explicitly or implicitly in nearly all contingency table (discrete multivariate) literature in statistics—and cross-classification models are special cases of such models. An examination of the empirical distribution of the number of circuits revealed an unexpected phenomenon. It showed that in one-member households the distributions were very close to Poisson, whereas in larger households the distributions were different, and the difference appeared to indicate the presence of item nonresponse.

A variable that has been used as a surrogate for the differences between respondents and nonrespondents is also examined. This analysis does not lead to any results that could contradict the conclusions that the estimates obtained from the model are unbiased.

The overall conclusion, with respect to the Lake County model that the authors built and examined, is that the model shows no signs of substantial bias due to variable omission

and that total nonresponse has no noticeable adverse effect. Yet item nonresponse remains a serious problem. This leads to a natural suggestion that individual household members, rather than households, should be asked about trip-making behavior. Clearly this is easier with mail-out/mail-back surveys than with telephone or perhaps even home interview surveys.

Logit as well as gravity models are generalized linear models (*1*). However, it should be noted that in most estimation procedures for the gravity model, the dependent variable is effectively the number of trips going from one zone to another. The effect of total nonresponse at the household level is akin to item nonresponse at the zonal level. Thus the arguments in this paper do not apply to such gravity model parameter estimation procedures. However, a discussion paralleling the one in this paper for trip generation models can be made for logit models.

SOME THEORETICAL CONSIDERATIONS

It is known that when linear least squares is used to estimate regression coefficients (and certain conditions are met), randomness in the independent variable values is not a necessity for unbiasedness. This fact has been exploited to combat bias in sample surveys (*2*). However, because of its importance for this argument and to make the paper more or less self-contained, this fact is demonstrated in following paragraphs.

Cross-classification trip generation analysis is shown to yield least-squares estimates that are the same for a wide range of realistic weights. This leads to the reasons that the authors prefer cross-classification models for trip generation modeling.

Condition for Unbiasedness

A linear regression model can be written as

$$y = X\beta + \varepsilon \tag{1}$$

Equation 1 implies that the linear relationship between the variables in the X-matrix (of independent variables) and y (number of trips) is the same for all households except for minor fluctuations (given by the error term, ε). In the sequel weighted least squares shall be considered for its greater generality and also because trip generation models are concerned with counted data, which typically render ordinary least squares inappropriate.

The weighted least-squares estimate, $\boldsymbol{b}$, of β, is

$$\boldsymbol{b} = (X'WX)^{-1}X'Wy \tag{2}$$

where $w_1, \ldots, w_n$ are positive weights and W is the diagonal matrix whose diagonal elements are $w_1, \ldots, w_n$ [i.e., $W = \text{diag}\ (w_1, \ldots, w_n)$]. If $\boldsymbol{b}$ is to be an unbiased estimate of β, then one of a set of three conditions, collectively called the Gauss-Markov conditions, must be met by Equation 1. This condition is

$$E(\varepsilon) = o \tag{3}$$

where ε equals $(\varepsilon_1, \ldots, \varepsilon_n)'$ and n is the number of households. From Equation 2,

$$\begin{aligned} \boldsymbol{b} &= (X'WX)^{-1}X'W(X\beta + \varepsilon) \\ &= (X'WX)^{-1}X'WX\beta + (X'WX)^{-1}X'W\varepsilon \\ &= \beta + (X'WX)^{-1}X'W\varepsilon \end{aligned}$$

Therefore, when

$$E(\varepsilon) = o \tag{4}$$

it follows that $E(\boldsymbol{b}) = \beta$, showing that $\boldsymbol{b}$ is unbiased. Note that each component ε_i of ε relates only to the ith observation and not to all observations. Thus, if each observation is believed to be valid and the regression model chosen is valid, $E(\varepsilon_i) = 0$. Then, of course Equation 4 would hold.

One way in which Equation 4 will fail to hold is if an explanatory variable is omitted from the analysis. The condition will also be violated if the algebraic form of the model is incorrect or does not apply to all the observations. No other condition is required for the unbiasedness of regression coefficients (*3,4*). In particular, the values of independent variables do not need to be random. The problem then is to develop the model without leaving out any important explanatory variable and by specifying the correct form of the model.

Categorical and Continuous Trip Generation Models

As mentioned earlier, two kinds of trip generation models are in common use. One kind is the so-called categorical or cross-classification model. An example of such a model is a table in which the rows correspond to, say, several levels of household size and the columns correspond to different levels of income. The entries themselves are average number of trips (or trip circuits) made by households in that category (the trip rates). Such models are equivalent to regression models with dummy variables but without an intercept term, as will be shown. In the categorical model presented later, household size and number of workers per household are called factors that have several levels. The other kind of trip generation model is in the form of a single equation containing mainly continuous independent variables. They will be called continuous regression models.

Typical categorical trip generation models are also linear regression models. There are two ways in which this can be seen:

- Approach 1:

Let y_i be the number of trips taken by the ith household, and let

$$x_{ij} = \begin{cases} 1 & \text{if } i\text{th household is in category } j \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Here category means a cell of the cross-classification table. Then a regression using y_i's as dependent values and x_{ij}'s as independent values and using no intercept gives exactly the table entries.

• Approach 2:

Alternatively, let y_i be the number of trips in category j and

$$x_{ij} = \begin{cases} \text{number of households in category } i & \text{if } i = j \\ 0 \quad \text{otherwise} \end{cases} \tag{6}$$

The only difference between Approaches 1 and 2 is that Approach 1 is a disaggregate version of Approach 2. In the first approach, there is one y_i for each i, whereas in the latter approach, y_i is a summation over all i's in category j.

To show that least squares would give exactly the same estimates as cross-classification would, consider Approach 1. Let $\boldsymbol{x}_j$ be the jth column of the matrix X of elements x_{ij}. Each such column contains only 0's and 1's, the 1's in column $\boldsymbol{x}_j$ being only in those rows for which the observation belongs to the jth category. Note that there is no intercept term.

Since the number of trips taken in each household is obtained by counting, it is also called a counted variable. When the dependent variable is counted, it typically has a Poisson distribution and weighted regression is usually called for to avoid a violation of the second of the Gauss-Markov condition given by

$$E(\varepsilon_i^2) = \sigma^2 \tag{7}$$

where $i = 1, 2, \ldots, n$ and σ is a positive constant. For a categorical model, these weights w_i and ω_j would be constant for each category j. Indeed, the condition of constancy of weights within each category is about as general a condition as the authors could require. The authors assume this condition in the sequel.

If $j \neq \ell$, $\boldsymbol{x}_j$ will contain a 1 only in those rows where $\boldsymbol{x}_\ell$ has a 0, since a household belongs to one category. Therefore $\boldsymbol{x}_j' W \boldsymbol{x}_\ell = 0$. The number of 1's in each $\boldsymbol{x}_j$ is the number n_j of households in category j. Therefore, $\boldsymbol{x}_j' W$ is a vector consisting of n_j nonzero terms, each of which is ω_j and $\boldsymbol{x}_j' W \boldsymbol{x}_j = n_j \omega_j$. The matrix $X'WX$ is, therefore, diagonal, diag $(n_1\omega_1, \ldots, n_k\omega_k)$ with diagonal elements $n_j\omega_k$. Consequently,

$$(X'WX)^{-1} = \text{diag}\,(n_1^{-1}\omega_1^{-1}, \ldots, n_k^{-1}\omega_k^{-1}) \tag{8}$$

If y_i is the number of trips taken by the ith household, $\boldsymbol{x}_j W \boldsymbol{y}$ is the total number, t_j, of trips taken by household i in category j, times ω_j. Consequently, $X'W\boldsymbol{y}$ is a vector whose jth element is $t_j\omega_j$. Hence, $\boldsymbol{b} = (X'WX)^{-1}X'W\boldsymbol{y}$ is a vector, the jth element of which is $\omega_j t_j/\omega_j n_j = t_j/n_j$, the trip rate for the jth category. Thus least squares using Model 5 gives exactly the same estimates one gets from categorical analysis—regardless of weighting so long as the weights are constant within a category. Approach 2 can be handled in a similar way.

When adequate data are available, the authors prefer categorical trip generation models over continuous models for the following reasons:

• No explicit weighting is required for categorical models.
• Nonlinearity of the model is usually not an issue for such models.

The only major shortcoming of categorical models is that, since each factor has several levels, the number of independent variables is large and, consequently, large numbers of observations are required. However, this is not a problem in transportation studies. Thus, categorical models would usually be more reliable.

EXAMPLE

A categorical trip generation model was developed. A brief description of the data is given, the model is presented, and the adequacy of this model to give unbiased trip generation estimates is checked.

The data used in the project were obtained by the Chicago Area Transportation Study in October 1989 from a mail-out/mail-back survey of Lake County in northeastern Illinois (Chicago Area Transportation Study, unpublished data, Aug. 1990). Lake County is a rapidly growing region of approximately half a million residents in north suburban Chicago, bordering Wisconsin, Lake Michigan, and Cook County. The county is a low-density suburban area encompassing large estates as well as sizable low-income areas.

The data set includes information on two types of variable: transportation-related variables, including the origin and destination of every trip, its purpose, travel time, mode used, automobile occupancy (for automobile trips) and walking distance if transit modes are involved, and census-related variables, such as number of persons per household, age, vehicle availability, gender, employment status, occupation, and income.

Respondents to the Lake County household travel behavior survey reported their household travel information for one of two days. The first travel date was October 12, 1989. Subsequently, reminder letters were mailed to those households that had not yet returned filled-in questionnaires and whose mail had not been returned by the post office as undeliverable. Two substitutes for the first travel day that had passed (Thursday, October 19 or October 26, 1989) were suggested to these households. A total of 9,143 questionnaires were sent out and 2,480 households returned usable questionnaires (a return rate of 27.1 percent).

TRIP GENERATION MODEL

A categorical household trip generation model was developed on the basis of two commonly used household socioeconomic variables, household size and number of employees per household. These variables are essentially traditional, having withstood the test of numerous studies. However, since such trip generation models have existed for a while and presumably are extensively investigated, it is unlikely that an important variable has not been considered at some stage. The purpose of this paper is not to indicate which socioeconomic variables in the trip generation model are more useful predictors of household travel but to examine such a model in the light of the earlier discussion to see if the estimates are biased.

To illustrate the earlier discussion, the cross-classification trip generation model has been presented both in cross-tabular form (in Table 1) and in the form of a regression model with no intercept and with dummy variables (in Equation 9, as in

Approach 1). In this trip generation model, the base household trip generation rates have been defined in terms of trip circuits. Trip circuits may be defined as the round-trip movements by household members that begin and end at home. Trip circuits, rather than trips, were the focus of this study because the trip generation models of the Chicago Area Transportation Study are in terms of trip circuits, and some consistency with those models was seen to be desirable. Trip generation calculations were obtained after developing cross tabulations of households stratified by occupants and employees. (Trip circuits have also been referred to in the literature as primary trips and trip chains.) The model obtained is

$$\begin{aligned}\text{Trip circuits}_i = {} & 0.81x_{i(0,1)} + 1.90x_{i(0,2)} + 2.08x_{i(0,3)} + 3.00x_{i(0,4)} \\ & + 2.00x_{i(0,5)} + 1.04x_{i(1,1)} + 1.89x_{i(1,2)} + 2.29x_{i(1,3)} \\ & + 2.61x_{i(1,4)} + 3.26x_{i(1,5)} + 2.33x_{i(1,6)} + 2.11x_{i(2,2)} \\ & + 2.45x_{i(2,3)} + 2.85x_{i(2,4)} + 3.84x_{i(2,5)} + 4.19x_{i(2,6)} \\ & + 3.24x_{i(3,3)} + 4.12x_{i(3,4)} + 4.38x_{i(3,5)} + 5.12x_{i(3,6)} \\ & + 4.47x_{i(4,4)} + 5.75x_{i(4,5)} + 6.83x_{i(4,6)} + 6.67x_{i(5,5)} \\ & + \varepsilon_i \qquad R^2 = .669,\ s = 3.12 \end{aligned} \tag{9}$$

where the dependent variable is the trip circuit rates in the ith household and the subscript within parenthesis for each x indicates the number of occupants and the number of workers in the ith household [for instance, $x_{i(1,0)}$ is the dummy variable that takes the value 1 if the ith household has one member who is not employed, and 0 otherwise].

The estimate b_j of β_j for each of the $j = 1, 2, \ldots, J$ categories in the trip generation model (Equation 9), then, is exactly equal to the mean trip circuit rate of that category (as in Table 1). In Table 1, there are 36 categories, whereas in Equation 9, there are 24 independent variables. Therefore, Equation 9 is estimating the trip rate in only some of the categories. The independent variables that were deleted had columns of only 0's in the X matrix (there were no households in these cells), and the columns of the matrix were, consequently, linearly related. Omitting these independent variables to avoid singularity also had the intuitive reasoning that there is no sense in estimating trip rates where there are no household entries at all.

TABLE 1 Categorical Trip Generation Model

Workers	Summary Statistic	Household Size 1	2	3	4	5	6
0	Mean:	0.81	1.90	2.08	3.00	2.00	
	Var:	0.75	2.45	2.63	9.00	12.00	—
	N:	128	189	12	3	3	
1	Mean:	1.04	1.89	2.30	2.61	3.26	2.33
	Var:	0.67	2.29	3.44	3.01	3.76	6.52
	N:	248	228	107	123	61	15
2	Mean:		2.11	2.45	2.85	3.84	4.19
	Var:	—	2.01	3.40	4.52	4.41	12.16
	N:		461	217	240	66	16
3	Mean:			3.24	4.12	4.38	5.12
	Var:	—	—	3.61	8.66	6.17	8.93
	N:			99	56	26	12
4	Mean:				4.47	5.75	6.83
	Var:	—	—	—	9.03	11.11	0.80
	N:				32	12	5
5	Mean:					6.67	
	Var:	—	—	—	—	15.47	—
	N:					6	

The '—' indicates that there are no entries in that category.

Statistical Assessment of Bias

The issue in this section is how one decides if substantial bias is present in the estimates. As mentioned, there are two ways in which bias can occur:

1. If a nonlinear situation is represented with a linear model, or
2. If an important variable is left out of the model.

Because the present model is categorical, where the estimated number of trips in the category is the mean number of trips in each category, no difficulty occurs because of Reason 1.

To detect whether a variable has been left out is always difficult. The statistical literature suggests two ways to check for omitted variables:

1. Examine outliers (which often show the need for additional variables), and
2. Examine if there is a relationship between the predicteds and the residuals of the model (because the orthogonality of the predicteds and the residuals require the unbiasedness of model estimates).

To look for outliers, the authors examined plots of residuals and the Studentized residuals against the predicteds, which indicated a few data points as possible outliers. These points had Studentized residuals of 2.78, 4.06, and 4.58 and DFFITS of 1.96, 1.01, and 1.02. [See works by Sen and Srivastava (*3*) and Belsley et al. (*4*) for a definition of Studentized residuals and DFFITS, the latter being a statistic commonly used to identify influential points.] The cutoff for DFFITS of 0.20, which is a consequence of typically suggested formulae, was not used because it drew attention to 5 percent of the 2,480 data points. Each of these three outliers represented households with large number of household members and no trip circuits. There were no working members in the first two and only one worker in the third. The outliers appear to be due to "natural causes." We found no compelling reason to suspect that the model in Equation 9 did not include all important variables.

As mentioned, it is also useful to see if there is a relationship between the residuals and the predicteds because the deletion of an important variable can cause the residuals of the model to have nonzero expectations. The plots did not reveal any systematic relationship. To corroborate this further, a second regression model was developed in which the dependent variable was the predicteds ($\hat{y}_i$) and the independent variable was the residuals (e_i) from Equation 9. The R^2 of .0003 indicated close to no fit between the predicteds and the residual. But, as pointed out elsewhere (*3*), the lack of a relationship be-

tween $\hat{y}_i$ and e_i does not conclusively lead to the decision that the model is unbiased, because patterns between residuals and the predicteds are not always apparent.

Poisson and Item Nonresponse

Another check for bias that is available for categorical models also did not suggest bias, but it did yield unexpected results. Variables whose values are obtained by counting something are known to have a Poisson distribution if the items that are counted are statistically independent. Although trips are not exactly independent, it is usually conjectured that the Poisson distribution approximately holds. Moreover, this assumption is consistent with the customary assumption made in traffic engineering that flows on links are Poisson—something that has been observed to be approximately true. A Poisson assumption underlies nearly all contingency table (discrete multivariate) analyses in statistics.

From Table 1 it can be seen that the mean-to-variance ratio deviates from 1 without any clearly discernible pattern, although the deviation is more marked for categories with a small sample size. To get a clearer idea of why the deviation from Poisson was occurring, the theoretical Poisson probabilities were computed for each category with means from the trip generation model (Equation 9) as the Poisson parameter.

The comparison of the theoretical Poisson distribution for each category with the actual data points revealed the possibility of item nonresponse (see Table 2; the first set of numbers for each category is the theoretical distribution for that category, the second set is the empirical distribution). In households with one member, no matter whether that member is a worker or a nonworker, the distribution of actual trip circuits conforms approximately to that of the theoretical Poisson distribution of trip circuits and in fact underestimates the theoretical frequency of no trip circuits. As household size increased, the number of households reporting zero trip circuits was far greater than what the theoretical distribution for that category predicted. This finding indicates an important way in which the data collected by sample surveys may bias trip generation information—that when one person in the household fills out the survey questionnaire, he or she might miss recording trips made by other members in the household.

The statistical techniques used reveal that although the data from which the trip generation model was developed showed some item nonresponse, there is no reason to suspect that the estimates from the model are biased (in the sense of total nonresponse). The authors, therefore, do not believe that any

TABLE 2 Comparison of Theoretical Poisson and Empirical Distributions of Trip Circuits by Category

Household Size	Workers	Number of Trip Circuits										
		0	1	2	3	4	5	6	7	8	9	10+
1	0	56.80	46.15	18.75	5.08	1.03	–	–	–	–	–	–
		54	50	20	2	2	–	–	–	–	–	–
1	1	100.86	104.42	54.05	18.66	4.83	1.00	–	–	–	–	–
		77	130	69	7	0	1	–	–	–	–	–
2	0	28.28	53.72	51.03	32.30	15.34	5.83	1.85	–	–	–	–
		53	21	54	27	22	11	–	–	–	–	–
2	1	34.43	65.09	61.52	38.77	18.32	6.93	2.18	0.59	–	–	–
		54	40	63	34	27	5	5	–	–	–	–
2	2	55.73	117.76	124.40	87.61	46.27	19.55	6.89	2.08	0.55	0.13	0.11
		89	33	173	95	52	16	2	0	0	0	1
3	0	1.49	3.11	3.24	2.25	1.17	0.49	–	–	–	–	–
		2	3	3	1	2	1	–	–	–	–	–
3	1	7.20	19.43	26.21	23.59	15.92	8.60	3.87	1.49	0. 50	–	–
		25	12	22	25	10	7	3	2	1	–	–
3	2	18.69	45.84	56.18	45.91	28.14	13.79	5.63	1.97	0.60	0.16	–
		48	14	52	43	34	15	7	1	2	1	–
3	3	3.87	12.54	20.33	21.98	17.81	11.55	6.24	2.89	1.17	0.42	–
		16	2	4	34	21	13	6	1	1	1	–
4	1	9.05	23.61	30.81	26.80	17.49	9.13	1.48	–	–	–	
		20	12	24	32	20	7	6	2	–	–	–
4	2	13.94	39.68	56.45	53.55	38.10	21.68	10.28	4.18	1.49	0.47	–
		51	4	51	48	46	16	13	2	2	2	–
4	3	0.91	3.73	7.70	10.59	10.92	9.01	6.19	3.65	1.89	0.86	0.45
		12	1	3	7	7	9	3	5	5	3	1
4	4	0.34	1.64	3.66	5.46	6.09	5.44	4.06	2.59	1.45	0.72	–
		8	0	0	3	2	4	4	7	3	1	–
5	1	2.34	7.62	12.43	13.52	11.03	7.19	3.91	1.82	0.74	–	–
		9	0	9	16	15	4	4	3	1	–	–
5	2	2.02	7.05	12.29	14.27	12.43	8.67	5.03	2.51	1.09	0.42	–
		10	2	5	13	19	6	7	2	1	1	–
6	1	1.45	3.39	3.95	3.09	1.78	0.84	0.33	0.11	0.03	0.008	–
		8	0	2	3	4	5	0	0	0	9	–
6	2	0.24	1.02	2.13	2.97	3.11	2.61	1.82	1.09	0.57	0.26	0.19
		4	0	1	2	5	4	1	2	0	0	1

important predictor of trip rates is missing from the model. However, work on an additional variable is described in the next section.

Analysis of Late Responses

The variable that was explicitly examined is response itself, if it is assumed that late respondents represent a point on the continuum between respondents and nonrespondents. Under this assumption, those independent variables that would be important in distinguishing between early and late respondents to the survey would also be useful predictors of the difference between respondents and nonrespondents.

The approach presented in this section is based on the premise that individuals who respond late to a survey, and from whom a response is elicited only after a reminder letter had been sent or a follow-up telephone call had been made, are "closer" to nonrespondents than to respondents because of the prodding needed to get the response from them. This is a fairly standard assumption. In a landmark study by Filion using this technique (*6*), nonrespondents were considered as persons who resist the initial waves of the questionnaire. All respondents form a continuum from highly motivated to unmotivated individuals. Each wave probes deeper into the core of the nonrespondents, and the continuum is indicative of the direction and extent of total nonresponse bias. Consequently, Filion claims that extrapolation over successive waves will reflect the characteristics of the hard core of the nonrespondents (*6*) [see also work by Armstrong and Overton (*7*) and Finn et al. (*8*)].

The data gathering design by Chicago Area Transportation Study led most naturally to this part of the analysis. Households that responded to the Lake County household travel behavior survey were divided into two groups, early and late, on the basis of the date on which they reported their travel information. Households that filled the survey form using October 12 or an earlier Thursday (October 5, 1989) as the travel day, were termed as the early respondents. Households that responded using Thursdays after October 12 were termed late respondents. The number of early respondents according to this definition was 1,907 households, and the number of late respondents was 573 households.

The new independent variable introduced to serve this purpose was a dummy variable z_j for each j, which takes the value 1 if the ith household in j responded early to the survey and 0 if the household responded late. A single dummy could have been used for each of the households for which trip circuits are predicted. However, a dummy variable for each j allows greater flexibility. The regression needs to be weighted, but earlier results—that estimates from cross-classification models are weighted least-squares estimates for a wide class of weights—are true only for estimates themselves, not for their standard errors. Thus, for testing purposes specific weights need to be specified.

For simplicity, the authors decided to proceed in a different way. Since the estimates of regression coefficients are trip rates, it was decided to simply compare trip rates using as variance estimates the usual sample variances. That is, a standard t-test was used to compare trip rates for early and late respondents in each category. The results are given in Table 3. A theoretical justification of this approach would follow the lines given by Sen and Srivastava (*3*).

Only one of the differences in Table 3 was significant at a 5 percent level (although in that case the significance was at a substantially lower level). Given 21 separate t-tests, getting one significance at a 5 percent level is about what should be expected if there is no relationship. Although the early re-

TABLE 3 Trip Circuit Rates (TCR) for Early and Late Respondents for Complete Model

Workers	Respondent Type	Household Size 1	2	3	4	5	6
0	Early TCR:	0.83	1.97	1.75			
	Late TCR:	0.73	1.51	2.75			
	Obsns.	[102, 22]	[158, 31]	[8, 4]	[2, 1]	[2, 1]	—
	t value:	0.59	1.59	-0.98			
1	Early TCR:	1.06	1.89	2.40	2.68	3.28	2.22
	Late TCR:	0.95	1.88	1.92	2.40	3.22	2.50
	Obsns.	[228, 56]	[180, 48]	[83, 24]	[91, 32]	[43, 18]	[9, 6]
	t value:	0.29	0.08	1.16	0.78	0.10	-0.21
2	Early TCR:		2.19	2.65	2.93	3.67	4.46
	Late TCR:		1.90	1.89	2.51	3.05	3.00
	Obsns:	—	[331, 130]	[160, 57]	[193, 47]	[46, 20]	[13,3]
	t value:	—	1.86	3.12*	1.06	0.99	0.73
3	Early TCR:			3.25	4.56	4.59	
	Late TCR:			3.22	2.36	4.00	
	Obsns.	—	—	[76, 23]	[45, 11]	[17, 9]	[11, 1]
	t value:			0.06	1.98	0.56	
4	Early TCR:				4.54	6.50	7.00
	Late TCR:				4.17	4.25	8.00
	Obsns:	—	—	—	[26, 6]	[8, 4]	[3, 2]
	t value:				0.25	1.32	-1.73
5	Early TCR:					7.50	
	Late TCR:					6.25	
	Obsns:	—	—	—	—	[2, 4]	
	t value:					0.37	

spondents, in most cells, indicated higher trip rates than late respondents, a reasonable conclusion still is that this variable is not too important.

An examination of the empirical distribution of trip circuits made by early and late respondents in each category also showed that the two groups essentially had the same pattern in the household trip circuits. This corroborates the conclusion that (assuming late respondents are closer to nonrespondents than to early respondents the model is reasonably unbiased and that there is no significant difference between the respondents and the nonrespondents.

CONCLUSION

Estimates from categorical trip generation models are unbiased if certain conditions are met by the model. The focus of developing trip generation models, therefore, shifts from being concerned about overrepresentation or underrepresentation in the data of households possessing certain socioeconomic characteristics to checking the model to see if these conditions are met. A cluster of checks is suggested to verify whether the model gives unbiased estimates of household trip generation parameters.

An empirical analysis was done to illustrate how this bias assessment may be done in practice. A series of statistical diagnostic tools was used, including outlier analysis and examination of the residuals and the predicteds from the model. These tools indicated that the model does not have substantial total nonresponse bias. However, the empirical distribution of only one-member households closely followed the theoretical distribution for such categories of households. This indicated that there is possible item nonresponse in households in which one person records trip information for the other household members. Finally, on the premise that late respondents are closer to nonrespondents than early respondents on a continuum of respondents to nonrespondents, an analysis was done to check if the two groups of respondents are sufficiently different. The results agreed with the earlier analysis that the model is reasonably free of total nonresponse bias.

REFERENCES

1. McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, England, 1989.
2. Särndal, C. E., and T. K. Hui. Estimation for Nonresponse Situations: To What Extent Must We Rely on Models? In *Current Topics in Survey Sampling* (D. Krewski, R. Platek, and J. N. K. Rao, eds.), Academic Press, San Diego, Calif., 1981.
3. Sen, A., and M. Srivastava. *Regression Analysis: Theory, Methods and Applications*. Springer-Verlag, New York, 1990.
4. Belsley, D. A., E. Kuh, and R. E. Welsch. *Regression Diagnostics—Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, Inc. New York, 1980.
5. Rao, C. R. *Linear Statistical Inference and its Applications*. John Wiley & Sons, Inc., New York, 1973.
6. Filion, F. L. Estimating Bias Due to Non-response in Mail Surveys. *Public Opinion Quarterly*, Winter, 1975–1976, pp. 482–492.
7. Armstrong, J. S., and T. S. Overton. Estimating Non-response Bias in Mail Surveys. *Journal of Marketing Research*, Vol. 14, Aug. 1977, pp. 396–402.
8. Finn, D. W., C. Wang, and C. W. Lamb. An Examination of the Effects of Sample Composition Bias in a Mail Survey. *Journal of Marketing Research Society*, Vol. 25, No. 4, Oct. 1983, pp. 331–338.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.