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

Table 3. Estimation statistics for sectors.

Item	Sector 1	Sector 2	Sector 3	Region
No. of cases	744	746	746	2,236
No. of observations	2,078	1,997	2,156	6,240
Log-likelihood at zero	-755	-722	-790	-2,266
Log-likelihood at convergence	-580	-636	-688	-1,928
Likelihood ratio statistic	350	171	203	678
Likelihood ratio index	0.232	0.118	0.129	0.150

Thus the variance of this standardized measure, given a particular model and population, is a function of sample size only.

Scattergrams of estimates for the standardized measure of the seven slope parameters defined in Table 1 have been plotted against estimation sample size for three different sectors, and they were found to be similar. The scattergram for one parameter in all three sectors is shown in Figure 2, along with the 95 percent confidence limits. As expected, the estimated parameters are distributed around the true parameters, with the range of the distribution decreasing as the number of cases in the estimation sample increases. It appears that the mean of Q is approximately zero (as expected), and its variance is described by Equation 6a. Further, approximately 95 percent of the reported deviations are within the expected range. Finally, as expected, the distribution appears to be independent of the estimation sector.

Parameter Precision and Required Sample Size

The deviations of sample parameter estimates from population parameters for each sector and variable are related to the standardized deviation (Q) by the population parameter standard deviation (s_p), as shown in Equation 5. Thus the variance of observed parameter deviations (z_s) is

$$V(z) = s_p^2 \times V(Q)$$
$$= s_p^2 \times [(N_p - N_s)/N_s] \quad (7)$$

which is a function of the estimation precision of the parameter in the population and the sample size. Thus it is possible to determine a priori the sample size necessary to obtain a predetermined level of precision in parameter estimates if the population estimation precision and the population size are known.

The interpretability of this relationship can be improved by formulating an index of estimation precision that is independent of both population size and sample size. Thus,

$$s_*^2 = \underset{N\to\infty}{E} (s_N^2 \times N) \quad (8)$$

This index, which can be estimated by

$$s_*^2 = s_p^2 \times N_p \quad (9a)$$

or

$$s_*^2 = s_s^2 \times N_s \quad (9b)$$

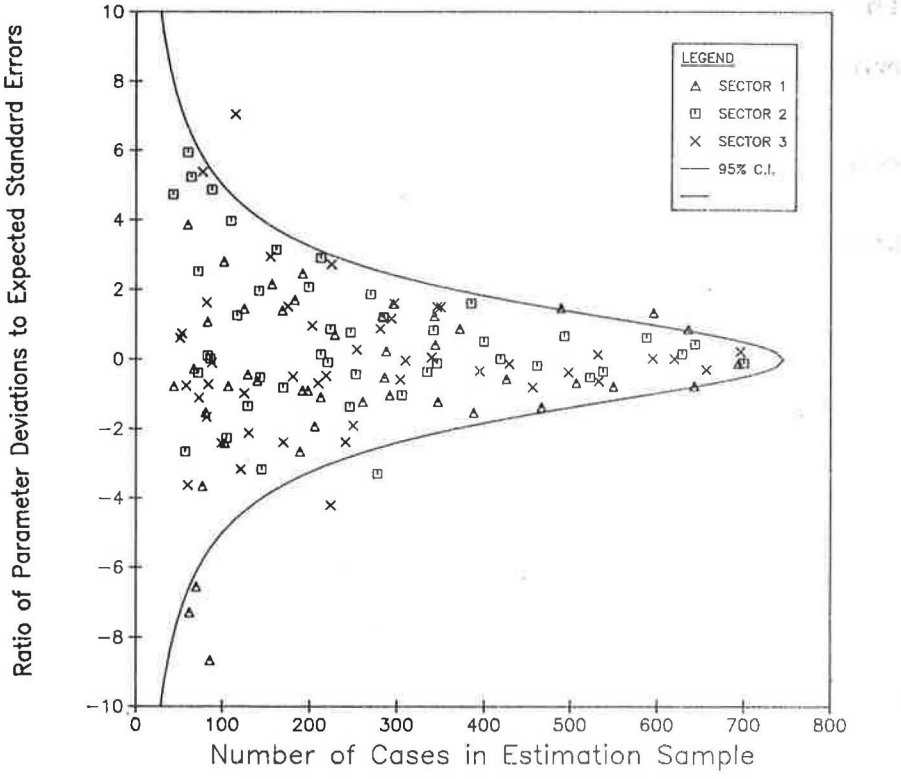
characterizes the underlying precision of a parameter independent of population or sample size. This index is used in Equation 7 to obtain

$$V(z) = s_*^2 [(N_p - N_s)/(N_p \times N_s)] \quad (10)$$

By using this formulation, the sample size required to obtain a desired level of precision in parameter estimation can be obtained as a function of population size and as the precision index for the parameter of interest. Specifically, an N_S is sought that satisfies

$$t_{\alpha/2}\, s_* [(N_p - N_s)/(N_p \times N_s)]^{1/2} = z^* \quad (11)$$

Figure 2. Scattergram of parameter precision with estimation sample size.



where $t_{\alpha/2}$ is the t value associated with the desired (α) confidence interval for z, and z* is the desired level of precision for parameter deviations. Thus the required sample size is

$$N_s^* = [N_p + N_p (z^*/s_* \, t_{\alpha/2})^2] / [1 + N_p (z^*/s_* \, t_{\alpha/2})^2] \quad (12)$$

which, when N_p is large, simplifies to

$$N_s^* \simeq (s_* \, t_{\alpha/2}/z^*)^2 \quad (13)$$

This relationship (Equation 12) is plotted in Figure 3 for the case where the parameter deviation (z) is to be within a prespecified fraction of the parameter precision index (s_*) with 95 percent confidence.

Equation 12 (or Equation 13 for large populations) can be used to predetermine the sample size required to obtain a desired level of parameter estimation precision. This determination is based only on prior knowledge of population parameter precision (s_*) and population size. Estimates of population parameter precision may be obtained by reviewing estimation results of similarly specified models in other contexts or by using a small data sample. The use of small data samples to obtain information to optimally design the sample collection procedure for a given sample size has been treated extensively by Daganzo (12).

The use of Equation 12 is demonstrated by calculating the sample size required to have 80 percent confidence so that the absolute value of z is less than 25 percent of the true parameter value. (More generally, this analysis can be undertaken by setting limits to the deviations of each parameter based on required or desired precisions in model sensitivity and the differences in the corresponding variable across plan alternatives. However, use of an arbitrary proportional range provides useful insight in an abstract context.) The calculation process and results are given in Table 4. These results illustrate again that as population increases, the number of sample observations needed to obtain parameter estimates in a prespecified range increases at a decreasing rate. When the population is large (i.e., more than 100,000), the required estimation sample size approximates that for an infinite population.

More important, the sample sizes required to obtain what would appear to be a modest level of parameter precision are substantially greater than those commonly used in the estimation of disaggre-

Figure 3. Required sample size with 95 percent confidence.

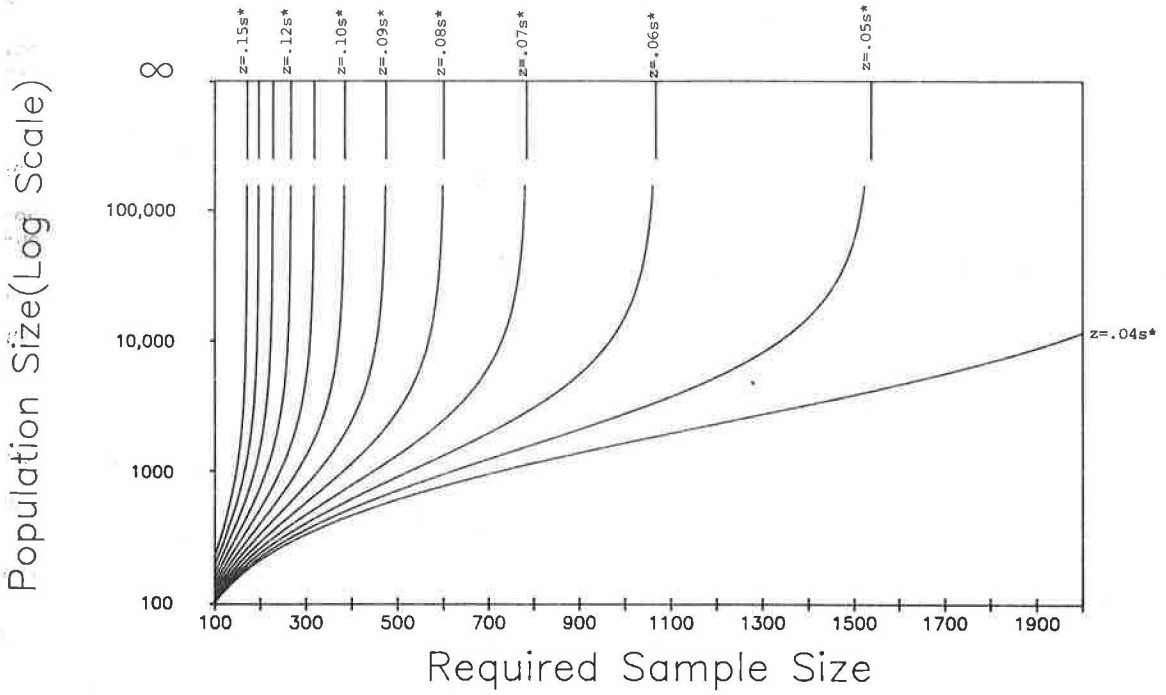


Table 4. Computation of required sample size to obtain parameter estimate with 80 percent confidence within 25 percent of true values.

Variable	β^* (see Tables 2 and 3)	Z ($\pm 0.25\,\beta^*$)	S_s (see Tables 2 and 3)	S_* (From Equation 8b)	Required Estimation of Sample Size for Different Population Sizes: N_p = 100,000	N_p = 1,000,000
DAD	-2.366	±0.5915	0.2261	10.69	533	536
SRD	-2.349	±0.5873	0.1747	8.26	324	325
CPDDA	3.047	±0.8518	0.2268	10.72	260	261
CPDSR	1.767	±0.4418	0.1593	7.53	475	477
GWSR	0.6477	±0.1619	0.0962	4.55	1,276	1,291
NWORKSR	0.3084	±0.0771	0.0674	3.19	2,721	2,789
OPTCINC	-0.0297	±0.0074	0.0084	0.395	4,422	4,605
TVTT	-0.0233	±0.0058	0.0031	0.147	1,032	1,042
OVTTD	-0.0588	±0.0147	0.0393	1,086.0	20,756	25,523

the predictions obtained by model transfer describe behavior in the prediction context." Transferability is a function of the quality of the model being transferred and similarity of behavior between the estimation and application contexts.

If choice behavior in the estimation and application contexts is based on the same behavioral process, the transfer predictive accuracy will be increased with increasing estimation sample size. In this case a model that is able to provide an accurate description of choice behavior in the estimation context will be able to provide an accurate description in the transfer or prediction context. However, if the behaviors are different between contexts, increasing sample size will not overcome these differences.

The objective of this paper is to examine the effect of sample size on parameter stability, parent population replication, and transferability of disaggregate discrete choice models of multinomial logit structure. In each case the expected relationship is formulated, an empirical analysis to scale the relationship is executed, the implications of the results obtained are identified, and the conclusions are stated. Also described in the paper are the data used and the structure of the empirical analysis undertaken.

DATA DESCRIPTION AND EXPERIMENTAL DESIGN

Data

The data used in this study are drawn from the Washington Council of Governments travel to work modal-choice data collected in Washington, D.C., in 1968. The data used describe the central business district (CBD) work trips of 2,236 persons. A total of 1,768 persons have drive-alone, shared-ride, and transit alternatives available, and 468 persons have only the shared-ride and transit alternatives because of a lack of driver's license or cars available in the household.

The data set is partitioned into three geographic sectors of the region according to worker residential location. Each sector includes approximately one-third of the sample observations. The partition allows for the examination of the first two relationships (parameter precision and parent population replication) within each sector and the investigation of the transferability prediction relationship for six possible transfers between sectors.

Experimental Design

The experiment is constructed by defining the full sample in each sector as the population of interest, and then subsamples of varying size are selected. These subsamples are used to estimate multinomial logit model parameters, predict choice behavior for the population from which each sample is drawn, and predict choice behavior in each of the other populations (different sectors). The flowchart of this experimental design is shown in Figure 1, which describes the sampling and estimation process and also the data used in each step.

The first task of the experiment is to obtain subsamples of each data set with varying sample sizes. Forty-five sets of random subsamples are independently generated within each of the three sectors. Within each sector the number of individuals in samples varies from approximately 50 to approximately 700.

The second task is to estimate travel modal-choice models for each data sample. A nine-variable model previously used in a related study of model transferability (11) is used in this study. These variables are described in Table 1. By using a single-model specification, it is possible to examine the effect of sample size without any confounding effects caused by differences in model specification. The estimation results for these models that use the full set of cases (the population) in each sector, as well as additional data, are reported in Tables 2 and 3. These estimation results serve as a reference point for the models estimated with each data subsample. The subsample estimation results are discussed later in this paper.

The third step in this study is to use the 45 models estimated in each sector to predict travel choices for the full population in each of the three

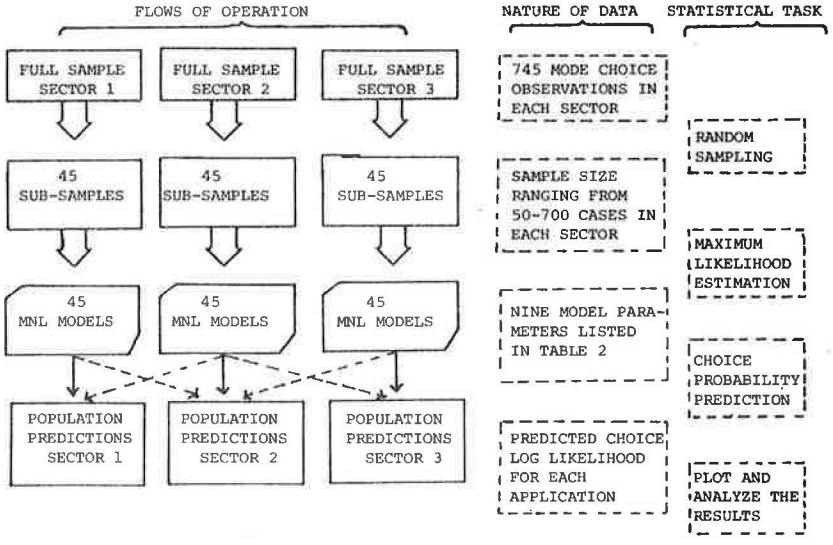


Figure 1. Flowchart for experimental design.

sectors. Thus each estimated model is used for three predictions (one local and two transfer predictions). Population replication performance and transferability measures are developed for each of these predictions and used to interpret the model accuracy relationships.

EFFECT OF SAMPLE SIZE ON PARAMETER PRECISION

Parameter precision is the inverse of the variance of parameter estimates obtained in repeated samples. In this section the effect of sample size on parameter precision is evaluated by comparing estimated parameter values for each sample with the population parameters reported in Tables 2 and 3.

Relation Between Parameter Precision and Sample Size

The total available data sample is treated as the population of interest, and the difference between models estimated on subsamples and models obtained from the population (full sample) is examined. As all the data included in each subsample are also included in the full sample, the parameter estimates obtained from samples are not independent of parameter estimates obtained from the full data. The variance-covariance matrix of estimates of the difference between sets of parameter estimates is

$$\Sigma_z = \Sigma_s + \Sigma_p - 2\Sigma_{sp} \quad (1)$$

where

Σ_z = error variance-covariance matrix for difference between subsample and full sample parameters (i.e., $z = \beta_s - \beta_p$);
Σ_s, Σ_p = error variance for subsample and full sample parameter estimates, respectively; and
Σ_{sp} = covariance matrix of error between subsample and full sample parameter estimates.

When the subsample is a subset of the full sample $\Sigma_{sp} = \Sigma_s$ (see Appendix),

$$\Sigma_z = \Sigma_s - \Sigma_p \quad (2)$$

which is a positive semidefinite covariance matrix of the differences between parameter estimates obtained from the full and partial samples. The expected relationship between the full and partial sample error variances is

$$\Sigma_p = \Sigma_s (N_s/N_p) \quad (3)$$

Thus, from Equations 2 and 3,

$$\Sigma_z = [(N_p - N_s)/N_p]\ \Sigma_s \quad (4a)$$

and

$$\Sigma_z = [(N_p - N_s)/N_s]\ \Sigma_p \quad (4b)$$

A standardized variable of differences is formulated in parameter estimates (Q) by dividing observed differences (z) by the standard error in population estimates (s_p); i.e., square root of diagonal elements in Σ_p,

$$Q = z/s_p \quad (5)$$

where Q is the difference between sample parameter and population parameter values in units of standard error of estimate for population parameters. Then the variance and 95 percent confidence interval of Q are

$$V(Q) = (N_p - N_s)/N_s \quad (6a)$$

and

$$-1.96\ [(N_p - N_s)/N_s]^{1/2} \leq Q^* \leq 1.96\ [(N_p - N_s)/N_s]^{1/2} \quad (6b)$$

Table 1. Model specifications.

Variable Name	Variable Description
DAD, SRD	Dummy variable specific to drive-alone and shared-ride alternative; measures average bias between pairs of alternatives other than that represented by the included variables
CPDDA, CPDSR	Cars per driver included separately as alternative specific variables from the drive-alone and shared-ride modes; measures the change in bias among modes caused by changes in automobile availability within the household
OPTCINC	Round trip out-of-pocket travel cost divided by income (cents/$1,000 per year); measures the effect of travel cost on mode utility with cost effect modified by household income level
TVTT	Round trip total travel time in minutes; measures the linear effect of combined in- and out-of-vehicle travel time in mode utility
OVTTD	Round trip out-of-vehicle travel time divided by trip distance (minutes/mile); measures the additional effect of out-of-vehicle travel time in utility in addition to the effect represented in TVTT; this added effect is structured to decline with increasing trip distance
GWSR	Dummy variable that indicates if the breadwinner is a government worker specific to the shared-ride alternative; measures the effect on shared-ride utility of shared-ride incentives for government workers
NWORKSR	Number of workers in the household specific to the shared-ride alternative; measures the change in utility of shared ride when there is an opportunity to share ride with a household member

Table 2. Parameter estimates and standard errors.

	Sector 1		Sector 2		Sector 3		Region	
Variable	Estimated Parameter	Standard Error	Estimated Parameter	Standard Error	Estimated Parameter	Standard Error	Estimated Parameter	Standard Error
DAD	-3.30	0.425	-1.44	0.388	-2.73	0.402	-2.67	0.226
SRD	-2.62	0.321	-1.92	0.277	-2.52	0.345	-2.35	0.175
CPDDA	4.06	0.426	2.70	0.382	3.58	0.396	3.41	0.227
CPDSR	2.06	0.319	1.67	0.235	1.59	0.315	1.77	0.159
OPTCINC	-0.0138	0.0155	-0.0282	0.0139	-0.0280	0.0163	-0.0297	0.0084
TVTT	-0.0459	0.0070	-0.0110	0.0050	-0.0223	0.0049	-0.0233	0.0031
OVTTD	-0.0019	0.0668	-0.1068	0.0666	-0.0421	0.0781	-0.0588	0.0393
GWSR	0.775	0.179	0.481	0.166	0.680	0.163	0.648	0.096
NWORKSR	0.133	0.128	0.275	0.110	0.502	0.123	0.308	0.067

gate choice models. The sample size required is substantially greater than the 300 to 500 observations that are commonly believed to be adequate for estimation of disaggregate choice models (13,14) for more than half of the model parameters. Use of the smaller samples can be expected to produce parameter estimates that have a high probability of being different from the true parameters. This problem is most serious for level-of-service parameters in this data set.

Conclusions

Two important observations are drawn from these results. First, as expected from sampling theory, the variability of parameter estimates is inversely related to sample size in a nonlinear fashion. This relationship is described in Equation 6a and is shown in Figure 3. Second, the sample size needed to obtain a reasonable degree of precision for managerial policy analysis may be substantially larger than is commonly suggested for the estimation of disaggregate choice models. The commonly held belief that 300 to 500 observations are satisfactory seriously underestimates the sample size suggested in this analysis to be needed to obtain estimators with a reasonable level of precision, especially for service variables. The importance of these results, if verified in other studies, is heightened by noting that many studies use samples of 1,000 or less observations (15-20), whereas this study suggests a need for at least 1,000 observations to estimate the influence of travel time--a most important variable--within an error of 25 percent with 80 percent confidence.

EFFECT OF SAMPLE SIZE ON REPLICATION OF PARENT POPULATION BEHAVIOR

In this study an examination was made of the accuracy with which a model, based on a data sample, will replicate the choice behavior in the parent population.

Relation Between Replication Precision and Sample Size

A prediction test statistic was formulated to test the hypothesis that the subsample model β_s is equivalent to the population model β_p,

$$PTS_p(\beta_s) = -2[LL_p(\beta_s) - LL_p(\beta_p)] \quad (14a)$$

This statistic, which is approximately chi-squared, can be expressed as a quadratic function of the difference in parameter vectors (5):

$$PTS_p(\beta_s) \simeq (\beta_p - \beta_s)' \Sigma_p^{-1} (\beta_p - \beta_s) \quad (14b)$$

Entering the relationships of $z = \beta_s - \beta_p^*$ and $\Sigma_z = [(N_p - N_s)/N_s]\Sigma_p$ into Equation 14b:

$$PTS_p(\beta_s) \simeq [(N_p - N_s)/N_s]\, z' \Sigma_z^{-1} z \quad (15)$$

where the quadratic term has a chi-square distribution. Thus the mean, variance, and $1 - \alpha$ confidence limit of PTS are

$$E(PTS) = [(N_p - N_s)/N_s] \times DF \quad (16)$$

$$V(PTS) = 2\,[(N_p - N_s)/N_s]^2 \times DF \quad (17)$$

and

$$PTS_\alpha \leq [(N_p - N_s)/N_s]\, \chi^2_{DF,\alpha} \quad (18)$$

Thus both the average and the variance of PTS decrease at decreasing rates as estimation sample size increases and are asymptotic to zero as sample size approaches population size.

Empirical Population Replication Analysis

To empirically demonstrate the results derived in the previous subsection, the predicted population log-likelihood by subsample models was compared with the maximum population log-likelihood by the full sample model in each sector by using Equation 14a. To examine the distribution and the 95 percent confidence limit of the prediction test statistic, scattergrams were plotted of the prediction test statistic against the size of estimation subsamples in Figure 4, and different symbols were used to represent observations in three different sectors. The results were as follows. First, as expected, PTS is subject to large variance when estimation sample size is small. The variance decreases quickly as estimation sample size increases for observations in all three sectors. Second, the curve that represents the expected value of PTS appears to fit the data well in all three sectors. Third, it appears that approximately 95 percent of the observations are within the 95 percent confidence limit shown in the figure. Thus these observations are consistent with the analytic results in the previous subsection.

Next, a prediction index was formulated that defines the degree to which the model estimated from the sample describes the population choice behavior relative to a model based on the full population. First, the common sample-based rho-square measure was considered:

$$\rho^2_{ss} = [LL_s(\beta_s) - LL_s(NM)]/[LL_s^* - LL_s(NM)] = 1 - [LL_s(\beta_s)/LL_s(NM)] \quad (19)$$

and then the corresponding population-based rho-square measure based on sample estimates was considered:

$$\rho^2_{ps} = [LL_p(\beta_s) - LL_p(NM)]/[LL_p^* - LL_p(NM)] = 1 - [LL_p(\beta_s)/LL_p(NM)] \quad (20)$$

Based on population estimates,

$$\rho^2_{pp} = [LL_p(\beta_p) - LL_p(NM)]/[LL_p^* - LL_p(NM)] = 1 - [LL_p(\beta_p)/LL_p(NM)] \quad (21)$$

Next, the prediction index as the ratio of Equations 20 and 21 were formulated to obtain

$$PI = [LL_p(\beta_s) - LL_p(NM)]/[LL_p(\beta_p) - LL_p(NM)] \quad (22)$$

The degree to which the sample-based model provides information about population behavior relative to that provided by the population-based model (when both referred to a common base or null model) is described by this ratio. To interpret this index, it was reformulated in terms of the population test statistic defined in Equation 14,

$$PI = 1 - \{PTS_p(\beta_s)/2[LL_p(\beta_p) - LL_p(NM)]\} \quad (23)$$

Note that the denominator in the second term is fixed for any population and model specification. Further, this term is the population model likelihood ratio statistic reported in Table 3 for each of the population models. These results can be used to obtain the expected value of the prediction index for fixed population size as

$$E(PI) = 1 - \{[(N_p - N_s)/N_s] \cdot DF/LRS\} \quad (24)$$

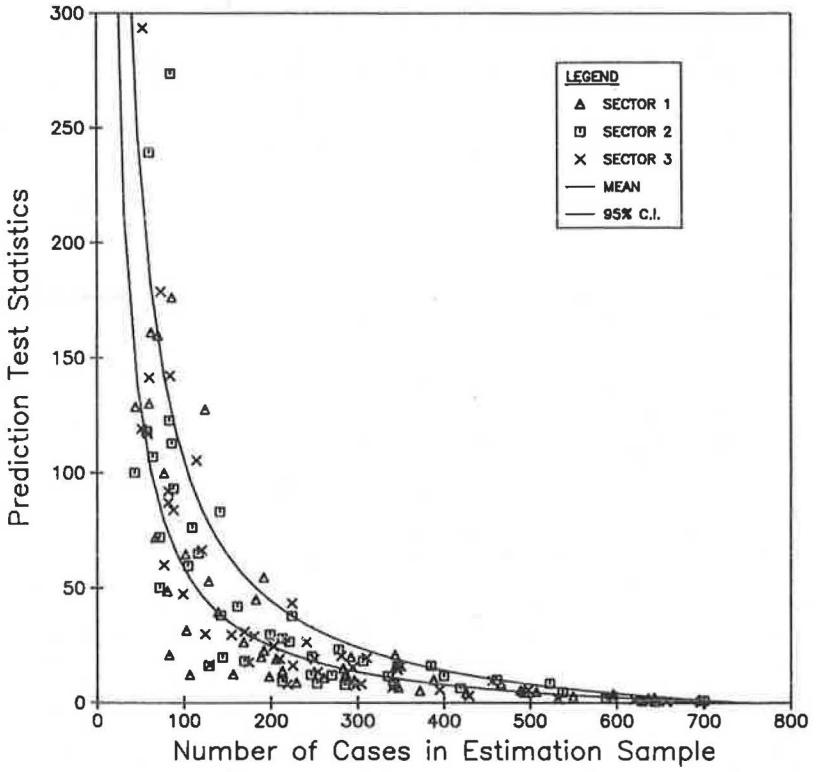


Figure 4. Scattergram of prediction test statistics with estimation sample size.

Finally, these results are modified for populations of varying size but otherwise identical characteristics by defining the likelihood ratio statistic per individual in the population (obviously, population data from which to compute the population model likelihood ratio statistic are not generally available; however, LRS* can be estimated by dividing sample likelihood ratio statistics by sample size):

$$LRS_* = LRS/N_p \quad (25)$$

to obtain

$$E(PI) = 1 - \left\{[(N_p - N_s)/N_s] \cdot DF/(N_p \times LRS_*)\right\} \quad (26)$$

which, when population size is much greater than sample size, is

$$E(PI) = 1 - (1/N_s)(DF/LRS_*) \quad (27)$$

The expected values of the prediction index for the three Washington sectors for different sample sizes are given in Table 5. The proportion of information provided by models estimated on samples of different sizes depends on the ability of the model to provide information about the behavior under study, as represented by the value of the likelihood ratio statistic per person. Sectors in which estimated models provide a higher level of information require smaller samples to achieve a specified level of relative accuracy. The results reported in Table 5 indicate that samples of 500 observations will provide 90 percent of the potential model information in each of the three Washington sectors.

Table 5. Expected value of prediction index (large population cases).

Sample Size	Expected Value of Prediction Index		
	Sector 1	Sector 2	Sector 3
50	0.62	0.21	0.34
100	0.81	0.61	0.67
200	0.90	0.80	0.83
300	0.94	0.87	0.89
500	0.96	0.92	0.93
1,000	0.98	0.96	0.97

Conclusions

The theoretical relationship between sample size and population description accuracy in the form of the prediction test statistic is developed in Equations 14-18. The empirical results reported in Table 4 are consistent with those relationships. The prediction index provides a somewhat more intuitive description of the relationship between sample size and descriptive accuracy. This relationship suggests that, in terms of descriptive accuracy alone, disaggregate samples of approximately 500 observations may be adequate. It is important to recognize the distinction between the ability to describe parent population choice behavior and prediction of behavior under different travel service conditions, which is most closely related to the precision of estimated parameters discussed previously.

EFFECT OF SAMPLE SIZE ON TRANSFERABILITY

Statistical Measure and General Expectation

Model transferability at the disaggregate level can be measured by indices formulated as a function of the difference in log-likelihood for the application sample of a transferred model [$LL_i(\beta_j)$] and the corresponding log-likelihood of a model estimated on that sample [$(LL_i(\beta_i)$]. The transfer test statistic formulated by Koppelman and Wilmot (11) is used to evaluate the transferability of disaggregate models. The transfer test statistic

$$TTS_i(\beta_j) = -2\,[LL_i(\beta_j) - LL_i(\beta_i)] \quad (28)$$

is chi-squared distributed with degrees of freedom equal to the number of model parameters under the assumption of fixed values of parameters for the transferred model. The smaller this statistic is, the more applicable is the transferred model to the application population.

This transfer test statistic is used to evaluate each of the sample-based models for transfer prediction of the population in each of the other sectors. Based on the results given previously, it is expected that the sample size of estimation subsamples will affect both the prediction accuracy and variability of a transferred model in the application context, according to a function that has a term of $(N_p - N_s)/N_s$ to reflect the sampling effect in the estimation context. It is also expected that there is a constant term in the transfer test statistic that reflects the real difference between the population of estimation and the population of prediction. These relationships are developed in the following subsection.

Relation Between Transfer Test Statistics and Estimation Sample Size

The transfer test statistic of a subsample-based model ($_j$), predicted on an alternative population, is defined as

$$TTS_{ij} = -2\,[LL_i(\beta_j) - LL_i(\beta_i^*)] \quad (29a)$$

which is approximately (5),

$$TTS_{ij} = (\beta^* - \beta_j)' \Sigma_{p_i}^{-1} (\beta_i^* - \beta_j) \quad (29b)$$

Let TTS_{ij}^* represent the transfer test statistic of the population-based model,

$$TTS_{ij}^* = (\beta_i^* - \beta_j^*)' \Sigma_{p_i}^{-1} (\beta_i^* - \beta_j^*) \quad (30)$$

which is nonstochastic, and assume that $N_{p_i} \times \Sigma_{p_i} \simeq N_{p_j} \times \Sigma_{p_j}$ (i.e., the underlying model parameter covariance matrices are equivalent) for the two populations; thus (21),

$$TTS_{ij} = TTS_{ij}^* + (N_{p_i}/N_{p_j}) \cdot [(N_{p_j} - N_{s_j})/N_{s_j}]\,[2\,(\beta_i^* - \beta_j^*)' \Sigma_{z_j}^{-1} z_j + z_j' \Sigma_{z_j}^{-1} z_j] \quad (31)$$

That is, the transfer test statistic for a model estimated on a sample from population j and used to predict population i is composed of a deterministic term that describes the difference between the two populations and a random variate composed of two terms. The first term, which is random because of the inclusion of z_j, is normally distributed with mean zero and variance-covariance matrix Σ_{z_j}. The second term, which is random because of the inclusion of $z_j' \Sigma_{z_j}^{-1} z_j$, is a chi-square variate with DF degrees of freedom. Thus TTS_{ij} is the sum of a fixed term, a normal variate and a chi-square variate. (Note that this breakdown of TTS_{ij} ignores the interaction between terms and the constraint required to ensure that TTS_{ij} is nonnegative.) The expected value and variance of TTS_{ij} are

$$E(TTS) = TTS_{ij}^* + (N_{p_i}/N_{p_j}) \times [(N_{p_j} - N_{s_j})/N_{s_j}] \cdot DF \quad (32a)$$

and

$$V(TTS) = 4\,(N_{p_i}/N_{p_j}) \times \{[(N_{p_j} - N_{s_j})/N_{s_j}]\,TTS_{ij}^*\} + 2\,(N_{p_i}/N_{p_j}) \times [(N_{p_j} - N_{s_j})/N_{s_j}] \times DF \quad (32b)$$

Thus both the mean and variance of the transfer test statistic increase with the difference between the two populations involved in the transfer process and decrease with the sample size of the estimation data set so that increased estimation sample size improves model transferability.

Empirical Analysis

The relationship between the transfer test statistic and sample size is examined empirically. The values of the population transfer test statistic (TTS*) are given in the following table:

Estimation Sector	Transfer Test Statistics by Prediction Sectors 1	2	3
1	--	67.2	72.2
2	48.6	--	29.0
3	52.6	27.2	--

A scattergram of the transfer test statistic is plotted with varying estimation sample size for transfers from sectors 1 and 3 to sector 2. This scattergram (Figure 5) can be used to examine the expected values and variances that were derived. In this figure, the expected value of the transfer test statistic, as defined by Equation 32a, is included. As expected, these lines fit the data in the respective transfer conditions satisfactorily. It was also observed that the variance of the transfer test statistic decreases as the estimation sample size increases, as suggested by Equation 32b. Further, it was noted that the sample values of TTS for transfers from sector 3 with the smaller value of TTS* have both lower mean and variance than the transfer from sector 1.

Conclusions

The expected relationship between sample size and transfer prediction accuracy is confirmed by the analytic decomposition of the transfer test statistic into a deterministic component that is independent of sample size and a stochastic component, the distribution of which is related to sample size for any given pair of populations. Empirical transferability tests are consistent with these analytically formulated relationships.

Increases in sample size cannot be used to offset real differences in the behavior of two populations reflected in TTS*. However, they can reduce the stochastic component. Additional analysis may be useful to clarify these relationships, but the empirical results suggest that samples in excess of 500 observations may be necessary to obtain transfer predictive accuracy that is close to that which might be obtained by a population-based model.

SUMMARY OF CONCLUSIONS

The conclusions reported in the preceding sections are summarized as follows.

1. Increased size of estimation samples leads to (a) parameter estimates that are likely to be closer to the true population parameters, (b) smaller standard errors of such parameter estimates, and (c) more accurate prediction of population choice behavior.
2. The sample size required to obtain choice model parameter estimates that are reasonably close

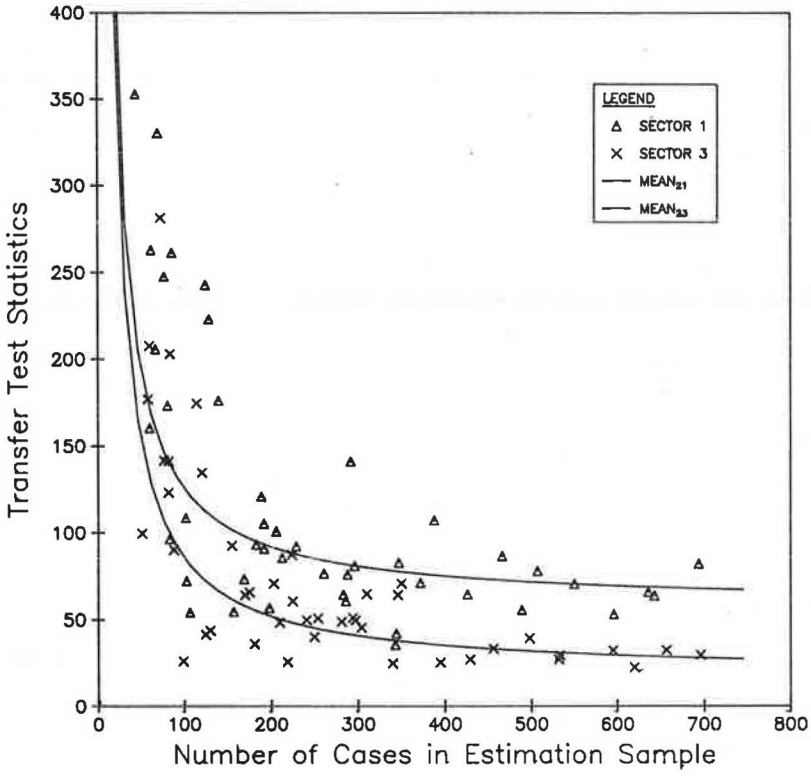


Figure 5. Scattergram of transfer test statistics predicted on sector 2.

to the true population parameters appears to be substantially larger than the sample sizes commonly prescribed for the estimation of disaggregate choice models.

3. The sample size required to obtain a model that accurately replicates parent population choice behavior appears to be somewhat smaller than that required to obtain accurate parameter estimates and accurate prediction under changed transportation service coordination.

4. Model transferability is a function of both the estimation sample size and the difference between the populations involved in the model transfer. Increasing estimation sample size has a positive effect on transferability at a decreasing rate. When the difference between two populations is large, it is expected that there will be large and highly variable transfer errors.

5. The required sample size needed to obtain a desired level of parameter estimation or prediction accuracy can be determined from pilot sample model estimation.

Overall, these results suggest the need to use data samples on the order of 1,000 to 2,000 observations rather than 500 observations as formerly believed. Although some reduction in sample size may be feasible when optimal sample stratifications are used (12, and paper by Sheffi and Tarem elsewhere in this Record), it is unlikely that samples as small as 500 observations can be adequate for model estimations.

Obviously, the importance of this issue suggests that additional research be undertaken to obtain further analysis of sample size requirements for models of different travel choices in different contexts. Further, transportation planners must formulate judgments about the desired precision of estimated model parameters and model prediction.

Appendix: Derivation of Sample Population Covariance Matrix

The population (full sample) estimation covariance matrix is the negative inverse of the Hessian (4) or

$$V_p = \left[\sum_{t \subset p} \sum_i (X_{it} - \bar{X}_t)' P_{it} (1 - P_{it}) (X_{it} - \bar{X}_t)\right]^{-1} \quad \text{(A1)}$$

where

- V = covariance matrix,
- Σ = summation,
- x_{it} = variable vector of alternative i for individual t,
- $\bar{X}_t$ = probability weighted average of x_{it},
- P = choice probability of alternative i for individual t,
- p = population,
- t = individual, and
- i = alternative.

Similarly, the sample estimation covariance matrix is

$$V_s = \left[\sum_{t \subset s} \sum_i (X_{it} - \bar{X}_t)' P_{it} (1 - P_{it}) (X_{it} - \bar{X}_t)\right]^{-1} \quad \text{(A2)}$$

where s is the sample indicator.

Finally, the covariance matrix between the population and sample estimates is given by

$$V_{sp} = \left[\sum_{t \subset s,p} \sum_i (X_{it} - \bar{X}_t)' P_{it} (1 - P_{it}) (X_{it} - \bar{X}_{it})\right]^{-1} \quad \text{(A3)}$$

where sp indicates the covariance matrix between population and sample estimations, and t C s,p implies summation over observations included in both the sample and the full population.

In this case, where the population includes all sample elements, the summation over s,p is equiva-

lent to the summation over s and $V_{sp} = V_s$ or, by using the notation in the body of the paper, $\Sigma_{sp} = \Sigma_s$.

ACKNOWLEDGMENT

We acknowledge the support of the Office of University Research, U.S. Department of Transportation, for support of this work under a contract to investigate "Conditions for Effective Transferability of Disaggregate Choice Models", and James Ryan of UMTA, who as contract monitor has provided substantial stimulation to this research effort.

We also acknowledge earlier analysis undertaken by Joseph N. Prashker of Technion University in collaboration with Koppelman, and valuable discussions with Eric I. Pas of Duke University. Finally, we thank our colleagues at the Northwestern University Transportation Center for intellectual and analytic support in the research. These colleagues include Chester G. Wilmot, Peter Kuah, and Albert Lunde.

REFERENCES

1. C.F. Manski. The Analysis of Qualitative Choice. Massachusetts Institute of Technology, Cambridge, Ph.D. dissertation, 1973.
2. C. Chu. Structural Issues and Sources of Bias in Residential Location and Travel Mode Choice Models. Department of Civil Engineering, Northwestern Univ., Evanston, Ill., Ph.D. dissertation, 1982.
3. H. Theil. Principles of Econometrics. Wiley, New York, 1971.
4. D. McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. In Frontiers in Econometrics (P. Zarembka, ed.), Academic Press, New York, 1973.
5. C. Daganzo. Multinomial Probit. Academic Press, New York, 1979.
6. G. Maddala. Econometrics. McGraw-Hill, New York, 1977.
7. C. Rao. Linear Statistical Inference and Its Application. Wiley, New York, 1965.
8. F.S. Koppelman. Methodology for Analyzing Errors in Prediction with Disaggregate Choice Models. TRB, Transportation Research Record 592, 1976, pp. 17-23.
9. T.J. Atherton and M.E. Ben-Akiva. Transferability and Updating of Disaggregate Travel Demand Models. TRB, Transportation Research Record 610, 1976, pp. 12-18.
10. D. McFadden. Properties of Multinomial Logit Model. Univ. of California, Berkeley, Working Paper 7617, 1976.
11. F.S. Koppelman and C.G. Wilmot. Transferability Analysis of Disaggregate Choice Models. TRB, Transportation Research Record 895, 1982, pp. 18-24.
12. C. Daganzo. Optimal Sampling Strategies for Statistical Models with Discrete Dependent Variables. Transportation Science, Vol. 14, No. 4, 1980.
13. M.F. Richards. Application of Disaggregate Techniques to Calibrate a Trip Distribution and Modal-Split Model. TRB, Transportation Research Record 592, 1976, pp. 29-34.
14. M. Richards and M. Ben-Akiva. A Disaggregate Travel Demand Model. Saxon House and Lexington Books, Lexington, Mass., 1975.
15. A.P. Talvitie and D. Kirschner. Specification, Transferability, and the Effect of Data Outliers in Modeling the Choice of Mode in Urban Travel. Transportation, Vol. 7, No. 3, 1978.
16. K.L. Sobel. Travel Demand Forecasting by Using the Nested Multinomial Logit Model. TRB, Transportation Research Record 775, 1980, pp. 48-55.
17. D. McFadden. The Theory and Practice of Disaggregate Demand Forecasting for Various Modes of Urban Transportation. In Emergency Transportation Planning Methods (W.F. Brown et al., eds.), U.S. Department of Transportation, Rept. DOT-RSPA-DPB-50-78-2, 1978.
18. K. Train. A Validation Test of a Disaggregate Mode Choice Model. Transportation Research, Vol. 12, 1978, pp. 167-174.
19. C.F. Manski and L. Sherman. An Empirical Analysis of Household Choice Among Motor Vehicles. Transportation Research, Vol. 14A, No. 5-6, Oct.-Dec. 1980.
20. T.A. Domencich and D. McFadden. Urban Travel Demand: A Behavioral Analysis. North-Holland, Amsterdam, 1975.
21. F. Koppelman and C. Chu. Sample Size Effects on Parameter Stability, Prediction Precision, and Transferability of Multinomial Logit Models. In Identification of Conditions for the Effective Transferability of Disaggregate Choice Models, Transportation Center, Northwestern Univ., Evanston, Ill., 1982.

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