The Usefulness of Prediction Success Tables for Discriminating Among Random Utility Travel Demand Models

JOEL L. HOROWITZ

ABSTRACT

The development of an empirical random utility travel demand model, like the development of most other statistically based models, typically includes testing and comparing several different functional specifications of the model to determine which specification best explains the available data. This paper is concerned with comparisons based on prediction success tables and indices. It is shown by example that prediction success tables and indices can lead to selection of the incorrect model when a correctly specified model is compared with an incorrectly specified one. This can happen even with data sets sufficiently large to make the effects of random sampling errors negligibly small. Accordingly, it is concluded that prediction success tables and indices should not be used for model selection. Alternative selection procedures that are both reliable and easy to use are described.

The development of an empirical random utility travel demand model (e.g., a logit or probit model), like the development of most other statistically based models, typically includes testing and comparing several different functional specifications of the model to determine which specification best explains the available data. For example, in developing a logit mode choice model, a specification in which the utility function is linear in the travel time might be compared with a specification in which the utility function is linear in the logarithm of the travel time. A variety of formal statistical procedures for testing and comparing alternative specifications of models is available (1-4). The discussion in this paper is concerned with comparison procedures based on prediction success tables and indices (3). These procedures have no formal justification, but they have greater intuitive appeal than do many of the formal procedures and therefore are attractive in practice.

A prediction success table for a model of choice among J alternatives contains J rows and J columns The entry in the (i,j) cell of the table is the number (or proportion) of individuals in the available data set who are observed to choose alternative i and predicted by the model under consideration to choose alternative j. Intuition suggests that a model with relatively large diagonal elements in its prediction success table is likely to provide a better explanation of the available data than is a model with relatively small diagonal elements because the former model gives a higher proportion of correct predictions of choice than does the latter. A single indicator of a model's prediction success can be obtained by forming a suitable average of the diagonal elements of its prediction success table. The resulting prediction success index provides an unambiguous criterion for discriminating among several models when no model dominates the others in terms of all of the diagonal elements of the models' prediction success tables.

The purpose of this paper is to show by means of examples that an erroneously specified model can have larger diagonal elements in its prediction success table and a larger prediction success index than does a correctly specified model. This can happen even with data sets sufficiently large to make the effects of random sampling errors negligibly small. Thus, contrary to intuition, prediction success tables and indices do not provide reliable means for comparing models with different specifications. Alternative comparison techniques that are both reliable and easy to use are described in the final section of the paper.

DEFINITIONS OF PREDICTION SUCCESS TABLES AND INDICES

Prediction success tables and indices were proposed originally by McFadden (3) as goodness-of-fit indicators for random utility models. A prediction success table for a model can be developed as follows. Let the available data consist of observations of N individuals who choose among J alternatives. Let P_{jn} denote the probability that individual n in the data set (n = 1, ..., N) chooses alternative j (j = 1, ..., J) according to the model under consideration. Let S_{jn} equal 1 if individual n is observed to choose alternative j and 0 otherwise. For each pair of alternatives (i,j) (i,j = 1, ..., J) define N_{ij} as

$$N_{ij} = \sum_{n=1}^{N} S_{in}P_{jn} \tag{1}$$

and define $\hat{\pi}_{ij}$ by

$$\hat{\pi}_{ij} = N_{ij}/N \tag{2}$$

Then N_{ij} and $\hat{\pi}_{ij}$ respectively represent the number and proportion of individuals in the data set who are observed to choose alternative i and predicted by the model to choose alternative j. N_{ii} and $\hat{\pi}_{ii}$ respectively represent the number and proportion of indivi-

duals who are correctly predicted to choose alternative i. A prediction success table for the model is the J x J array whose (i,j) element is either N_{ij} or $\hat{\pi}_{ij}$. Either form of the table contains the same diagnostic information, and it is a matter of convenience which is used. In this paper, it will be convenient to use the form based on $\hat{\pi}_{ij}$.

The total proportion of choices successfully predicted by the model under consideration is

$$\hat{\pi} = \sum_{j=1}^{J} \hat{\pi}_{jj} \tag{3}$$

This constitutes a goodness-of-fit index for the model. However, a better index can be achieved by averaging the differences between the proportions of correct predictions for each alternative obtained from the model and the proportions of correct predictions that would be obtained if each alternative were assumed to be chosen by each individual with a probability equal to the alternative's observed aggregate share. The resulting prediction success index is

$$\hat{\sigma} = \sum_{j=1}^{J} [\hat{\pi}_{jj} - (\hat{\pi}_{j\cdot})^2] \tag{4}$$

where

$$\hat{\pi}_{j\cdot} = \sum_{k=1}^{J} \hat{\pi}_{jk} \tag{5}$$

[Equation 4 corrects a typographical error in a previous study (3) that has the effect of exchanging the order of the subscripts on the right-hand side of Equation 5.] When $\hat{\pi}$ (or $\hat{\sigma}$) is used to compare models, the model with the largest $\hat{\pi}$ (or $\hat{\sigma}$) value is preferred to the others because this model yields the largest proportion of correct predictions in the case of $\hat{\pi}$ or, in the case of $\hat{\sigma}$, the largest increase in the proportion of correct predictions relative to the proportion implied by the observed aggregate shares.

CRITERION FOR EVALUATING USEFULNESS OF PREDICTION SUCCESS TABLES AND INDICES

In this paper, prediction success tables and indices will be evaluated according to their abilities to distinguish between correctly and incorrectly specified models. Before this can be done, it is necessary to consider the effects of random sampling errors on the ability of any statistical procedure to distinguish between correct and incorrect models and to identify a method for dealing with these effects. Random sampling error arises because different individuals with the same observable characteristics (i.e., the same values of a model's explanatory variables) and the same sets of alternatives may make different choices because of the effects of unobserved factors. As a result, the estimated parameter values, choice probabilities, and goodness-of-fit statistics for a model tend to have different values in different finite samples of individuals. These random fluctuations in estimation results can cause a goodness-of-fit statistic for an incorrectly specified model to be more favorable than that for a correctly specified model on occasion, even if the statistic usually or on the average favors the correct model. Random sampling error therefore constitutes a "noise factor" that impairs the ability of test statistics to distinguish correct models from incorrect ones.

Random sampling error always can be made negligibly small by making the sample used for estimating and testing models sufficiently large. Moreover, if the sample is large enough to make the effects of sampling error negligible, then it always is possible to determine unambiguously whether a model is correct by comparing the values of its choice probabilities for each set of values of the explanatory variables with the observed choices of individuals with the same values of the explanatory variables. A model whose choice probabilities for the available alternatives differ from the observed proportions of individuals choosing these alternatives is incorrect. Accordingly, it is reasonable to demand for comparison statistics, such as prediction success tables and indices, that they be capable of distinguishing without error between correct and incorrect models in the absence of random sampling error. In formal statistical terms, this property of a test is called consistency. Statistical test procedures that are not consistent usually are considered to be unacceptable.

In the next section, it will be shown by example that prediction success tables and indices are not consistent when used to discriminate among models. In other words, prediction success tables and indices can result in the selection of an incorrect model in a comparison with a correct one, even if the sample used for estimating and testing the models is large enough to make random sampling errors negligibly small. To show this, it is necessary to be able to evaluate the limits of the entries in a prediction success table and of $\hat{\pi}$ and $\hat{\sigma}$ as the sample size approaches infinity (large-sample limits). It follows from the strong law of large numbers that as the sample size N approaches infinity, the entries $\hat{\pi}_{ij}$ in a prediction success table approach

$$\pi_{ij} = E[Q_i(X)P_j(X)] \tag{6}$$

where $Q_i(X)$ denotes the true probability that a randomly selected individual for whom the values of the explanatory variables are X chooses alternative i (i.e., the probability according to the correctly specified model and the true parameter values), $P_j(X)$ denotes the large-sample limit of the probability according to the model under consideration that a randomly selected individual for whom the values of the explanatory variables are X chooses alternative j, and E denotes the expectation over the distribution of explanatory variables X in the population being sampled. The large-sample limits of $\hat{\pi}$ and $\hat{\sigma}$ are obtained by substituting Equation 6 into Equations 3 and 4. These limits will be denoted by π and σ, respectively.

TWO EXAMPLES OF INCONSISTENCY

Suppose that a model of choice among two alternatives (e.g., mode choice between automobile and transit) is being developed. Then J = 2, and

$$\pi_{11} = E(Q_1P_1) \tag{7}$$

$$\pi_{22} = \pi_{11} - E(P_1) - E(Q_1) + 1 \tag{8}$$

where the argument X of P_1 and Q_1 has been suppressed to simplify the notation. If $P_1(X) = Q_1(X)$ for all X (i.e., the model under consideration is correctly specified), Equations 7 and 8 become

$$\pi_{11} = E(Q_1^2) \tag{9}$$

$$\pi_{22} = \pi_{11} - 2E(Q_1) + 1 \quad (10)$$

By subtracting Equation 9 from Equation 7 and Equation 10 from Equation 8, one obtains the large-sample limits of the differences between the diagonal elements of the prediction success tables of an arbitrary model P and the correctly specified model Q. Denote the limits of these differences by $\Delta\pi_{jj}(j = 1,2)$. Then

$$\Delta\pi_{11} = E\,Q_1(P_1 - Q_1) \quad (11)$$

$$\Delta\pi_{22} = \Delta\pi_{11} - E(P_1 - Q_1) \quad (12)$$

Now suppose that models P and Q yield the same predictions of the aggregate shares of alternatives 1 and 2. Then $E(P_1 - Q_1)$ and

$$\Delta\pi_{11} = \Delta\pi_{22} = E\,Q_1(P_1 - Q_1) \quad (13)$$

Equivalently,

$$\Delta\pi_{11} = \Delta\pi_{22} - E\{[Q_1 - E(Q_1)](P_1 - Q_1)\} \quad (14)$$

Finally, suppose that in addition to satisfying $E(P_1 - Q_1) = 0$, P_1 has the property that

$$P_1(X) = \begin{cases} 1 \text{ if } Q_1(X) > E(Q_1) \\ 0 \text{ otherwise} \end{cases} \quad (15)$$

In other words, model P assigns individuals deterministically to alternative 1 if $Q_1(X) > E(Q_1)$ and deterministically to alternative 2 otherwise. Model P is misspecified because $P_1(X) \neq Q_1(X)$ whenever $Q_1(X)$ differs from 1 or 0. However, it can be seen from Equation 14 that $\Delta\pi_{11} > 0$ and $\Delta\pi_{22} > 0$. Therefore, if the sample size is large enough to make random sampling errors negligibly small, the diagonal elements of the prediction success table of the erroneous model P will exceed the corresponding elements of the prediction success table of the correct model Q. Similarly, the goodness-of-fit indices $\hat{\pi}$ and $\hat{\sigma}$ will be larger for model P than for model Q when the sample size is sufficiently large. Thus, the prediction success tables and indices will lead to selection of the wrong model in large samples and are inconsistent. The following example illustrates this result numerically.

Example 1

In a model of mode choice between automobile and transit let mode 1 be automobile and mode 2 be transit. Let the correctly specified model be

$$Q_1(T) = 1/[1 + \exp(-0.1T)] \quad (16)$$

where T denotes transit travel time minus automobile travel time in minutes. Let the distribution of T in the sampled population be uniform on the interval [-10,10]. Then $E(Q_1) = E(Q_2) = 0.5$ in this population, and $E(Q_1^2) = E(Q_2^2) = 0.27$. It follows from setting $P_i = Q_i$ in Equation 6 that the large-sample limit of model Q's prediction success table is

$$\text{Table}(Q) = \begin{bmatrix} 0.27 & 0.23 \\ 0.23 & 0.27 \end{bmatrix} \quad (17)$$

The values of π and σ for model Q are $\pi(Q) = 0.54$ and $\sigma(Q) = 0.04$.

Now define the misspecified model P by

$$P_1(T) = \begin{cases} 1 \text{ if } T > 0 \\ 0 \text{ otherwise} \end{cases} \quad (18)$$

Then $E(P_1) = E(Q_1)$, and $E(P_1Q_1) = 0.31$. It follows from Equation 6 that the large-sample limit of model P's prediction success table is

$$\text{Table}(P) = \begin{bmatrix} 0.31 & 0.19 \\ 0.19 & 0.31 \end{bmatrix} \quad (19)$$

The values of π and σ for model P are $\pi(P) = 0.62$ and $\sigma(P) = 0.12$. Thus, if the sample size is sufficiently large and models Q and P are compared by using their prediction success tables or their π- or σ-values, the erroneous model P will be accepted and the correct model Q rejected. This is true despite the fact that model P yields predictions that can be both unreasonable and highly erroneous. For example, suppose that T = 1 for a certain population group (i.e., transit travel time exceeds automobile travel time by 1 min). Then model Q yields the result that 48 percent of the members of this group use transit, whereas model P yields the unreasonable and erroneous result that no members of the group use transit.

Example 1 shows that the use of prediction success tables and indices for model selection can lead to selection of an erroneously specified model and rejection of a correctly specified one. However, the erroneous model P used in this example cannot be obtained through maximum-likelihood estimation, which is the standard method for estimating empirical choice models. This suggests the possibility that prediction success tables and indices may discriminate correctly among models when the sample size is large if consideration is restricted to models that can be obtained through maximum-likelihood estimation. The next example shows that even when this restriction is imposed, prediction success tables and indices can select the wrong model.

Example 2

As in Example 1, let individuals choose between the modes automobile (mode 1) and transit (mode 2). Let the correctly specified model be given by Equation 16. Assume that the values of T in the sampled population are restricted to those shown in Table 1 (e.g., because the sample is stratified) and that each of these values occurs with probability 1/9. Let the erroneous model be specified as

$$P_1 = 1/[1 + \exp(-\alpha C)] \quad (20)$$

where α is a positive constant and C is the cost of transit travel minus the cost of automobile travel in dollars. Assume that in the sampled population, there is a unique value of C associated with each value of T (e.g., because of the stratification procedure that is used) and that the C-values corresponding to the T-values are as shown in Table 1.

TABLE 1 Values of Explanatory Variables for Example 2

T (min)	C ($)	T (min)	C ($)
-80.0	-1.00	10.0	0.29
-60.0	-0.97	20.0	0.52
-20.0	-0.52	60.0	0.97
-10.0	-0.29	80.0	1.00
0.0	0.0		

The large-sample limit of the prediction success table of the correct model Q is

$$\text{Table (Q)} = \begin{bmatrix} 0.4046 & 0.0954 \\ 0.0954 & 0.4046 \end{bmatrix} \quad (21)$$

The π- and σ-values of model Q are $\pi(Q) = 0.8092$ and $\sigma(Q) = 0.3092$. The large-sample limit of the maximum-likelihood estimate of α, which can be computed by using methods described elsewhere (4), is 4.2877. The large-sample limit of the prediction success table of the erroneous model P can be obtained from Equation 6 by using Equation 16 to evaluate the Q probabilities and Equation 20 with $\alpha = 0.2332$ to evaluate the P probabilities. The result is

$$\text{Table (P)} = \begin{bmatrix} 0.4060 & 0.0940 \\ 0.0940 & 0.4060 \end{bmatrix} \quad (22)$$

The π- and σ-values of model P are $\pi(P) = 0.8120$ and $\sigma(P) = 0.3120$. It can be seen that the prediction success tables and π- and σ-values all favor the erroneous model P. Although the differences between the prediction success tables and π- and σ-values of the two models are small, a comparison of the models based on any of these criteria will lead to acceptance of the erroneous model and rejection of the correct one if the sample size is large enough to make random sampling errors negligible. As an example of the prediction errors that can result from selection of the incorrect model, suppose that transit improvements cause T to decrease from 20.0 to 10.0 for a certain population group while C remains unchanged. Then model Q yields the result that transit ridership in this group increases by 126 percent, whereas model P yields the result that there is no change in transit ridership.

DISCUSSION

The examples presented here show that prediction success tables and indices are unreliable means for discriminating among models. They can result in acceptance of an incorrect model and rejection of a correct one, even when the sample used for estimation and testing is large enough to make random sampling errors negligible. Because, as will now be discussed, comparison procedures that do not have this deficiency are readily available, prediction success tables and indices should not be used for model selection.

The appropriate procedure to use for comparing two models depends on whether the models are nested or nonnested. Two models are nested if one model can be obtained from the other by assigning appropriate values to the latter model's parameters. In nonnested models, this cannot be done; given the values of either model's parameters, it is not possible to choose values of the other model's parameters so that the two models become identical. Models P and Q in Example 1 are nested because P can be obtained from Q by setting the coefficient of T in Q equal to $-\infty$. Models P and Q in Example 2 are nonnested. See the discussion by Horowitz (1) for further examples of nested and nonnested models.

Comparisons of nested models can best be carried out by using likelihood ratio or t-tests (3). In a comparison of a correctly specified model with an incorrectly specified one, these tests always select the correct model in the absence of random sampling error (i.e., they are consistent). When random sampling error is present (as it always is in practice), likelihood ratio and t-tests have high probabilities of selecting the correct model when a correctly specified model is compared with one that is seriously erroneous (4). Likelihood ratio and t-tests are easily implemented because they rely on information that is virtually always included in the outputs of computer programs used for estimating random utility travel demand models.

Nonnested models can be compared easily by using the likelihood ratio index statistic modified to account for the effects of any differences in the numbers of estimated parameters in the models being compared (1,2). Like the likelihood ratio and t-tests for nested models, comparisons based on the modified likelihood ratio index are consistent and with samples of practical size, where random sampling error is present, have high probabilities of selecting the correct model when a correctly specified model is compared with a seriously erroneous one (1,2). Comparisons based on the modified likelihood ratio index can be implemented by using information that is included in the outputs of existing computer programs for estimating random utility travel demand models.

ACKNOWLEDGMENT

The research reported in this paper was supported in part by an assistance agreement between the Urban Mass Transportation Administration and the University of Iowa.

REFERENCES

1. J.L. Horowitz. Evaluation of Usefulness of Two Standard Goodness-of-Fit Indicators for Comparing Non-Nested Random Utility Models. In Transportation Research Record 874, TRB, National Research Council, Washington, D.C., 1982, pp. 19-25.
2. J.L. Horowitz. Statistical Comparison of Non-Nested Probabilistic Discrete Choice Models. Transportation Science, Vol. 17, 1983, pp. 319-350.
3. D. McFadden. Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments. In Behavioural Travel Modelling, (D.A. Hensher and P.R. Stopher, eds.), Croom Helm, London, 1979.
4. J.L. Horowitz. Identification and Diagnosis of Specification Errors in the Multinomial Logit Model. Transportation Research, Vol. 14B, 1980, pp. 331-34.

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