for a positive, finite constant, k, which is independent of n.

The lemma states in words that the difference between the estimated and true sample variances converges uniformly in probability to zero as the sample size increases. This lemma in conjunction with (A3) will be used to show that the estimated sample variance converges in probability to $\sigma^2(\theta_o)$ (lemma 2). First we prove this lemma.

Proof: Let us rewrite for convenience the estimated sample variance as $\hat{S}^2_{n_2}$ and the sample variance for a given value of θ as $S^2_n(\theta|X)$. Because of property (A), $S^2_n(\theta|X)$ is a continuously differentiable function of θ with derivatives:

$$[\partial S^2_n(\theta|X)]/[\partial\theta_j]\,|_{\theta=\theta_o} = \sum_{i=1}^{n} (2/n)\Big\{G(\theta_o, X^{(i)}) - \sum_{m=1}^{n} G(\theta_o, X^{(m)}) \div n\Big\}\,[\partial G(\theta, X^{(i)})/\partial\theta_j]\,|_{\theta=\theta_o}$$

Because all of the elements in this sum are continuous for all values of $X = (X^{(1)},\ldots,X^{(n)})$, and by property (B) the $X^{(i)}$ values that can possibly occur are bounded, the addends are bounded by a number, M_j/n. Thus, the partial derivatives of $S^2_n(\theta|X)$ are uniformly bounded and for any θ in a small neighborhood of θ_o we can write:

$$|S^2_n(\theta|X) - S^2_n(\theta_o|X)| \leq \|\theta - \theta_o\|\, k; \|\theta - \theta_o\| < \delta'_o < \delta_o$$

for some $\delta'_o > 0$ and finite $k > 0$. The combination of this fact and (A4) proves the lemma (as long as $\epsilon'' \leq \delta'_o$). We now show that $\{\hat{S}^2_n\} \xrightarrow{p} \sigma^2(\theta_o)$.

Lemma 2

$$[\hat{S}^2_n] \xrightarrow{p} \sigma^2(\theta_o)$$

Proof: We shall prove that for any positive (ϵ, δ):

$$\Pr[|\hat{S}^2_n - \sigma^2(\theta_o)| \leq \epsilon] \geq 1 - \delta; \forall n \geq N(\epsilon, \delta) \qquad (A6)$$

where $N(\epsilon, \delta)$ is a finite positive integer that depends on ϵ and δ. Combining (A3) and (A5) we can write:

$$\begin{aligned}\Pr[|\hat{S}^2_n - \sigma^2(\theta_o)| \leq \epsilon' + k\epsilon''] &\geq \Pr[|\hat{S}^2_n - S^2_n(\theta_o|X)| \leq k\epsilon'' \text{ and } |S^2_n(\theta_o|X) - \sigma^2(\theta_o)| \leq \epsilon'] \\ &\geq 1 - (\delta' + \delta''), \text{ if } n \geq \max[N'(\epsilon', \delta'); N''(\epsilon'', \delta'')] \qquad (A7)\end{aligned}$$

It is now clear that Equation A6 follows from Equation A7 with $\delta' = \delta'' = \delta/2$, $\epsilon' = k\epsilon'' = \epsilon/2$, and $N(\epsilon, \delta) = \max\{N'[(\epsilon/2), (\delta/2)]; N''[(\epsilon/2k), (\delta/2)]\}$.

Lemma 3

Identical arguments show that the sample mean $\bar{G}_n$:

$$\bar{G}_n = (1/n)\sum_{i=1}^{n} G[\tilde{\theta}(X^{(1)},\ldots,X^{(n)}, i^{(1)},\ldots,i^{(n)}), X^{(i)}]$$

converges in probability to $f_1 = E_X[P_1(X)]$

Theorem: The D-statistic is a consistent estimator for d.

Proof: From the definition,

$$d = f[f_1, \sigma^2(\theta_o)] = \sqrt{\sigma^2(\theta_o)/f_1(1 - f_1)} \qquad (A8)$$

where the function $f(.,.)$ is continuous if $f_1 \epsilon (0, 1)$. Regularity condition (C) implies that $f_1 = E_X[G(\theta_o, X)]$ is in the open unit interval and therefore confirms the continuity of $f(.,.)$.

This continuity implies that if f_1 and $\sigma^2(\theta_o)$ are replaced by consistent estimators, $\bar{G}_n$ and $\hat{S}^2_n$, as the arguments of $f(.,.)$, the result:

$$f(\bar{G}_n, \hat{S}^2_n) = \sqrt{[\hat{S}^2_n/\bar{G}_n(1 - \bar{G}_n)]} = D$$

is a consistent estimator.

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

Evaluation of Usefulness of Two Standard Goodness-of-Fit Indicators for Comparing Non-Nested Random Utility Models

JOEL L. HOROWITZ

The likelihood ratio index and the percentage of correctly predicted choices in the estimation sample are two well-known goodness-of-fit indicators for logit and other random utility models. They are used frequently for comparing non-nested models (i.e., models such that neither can be obtained from the other by choosing suitable values of the estimated parameters) to determine which best explains the available data. The results of an investigation of the abilities of the two statistics to distinguish between correct and incorrect models in such comparisons are reported. It is shown that with estimation data sets of practical size, a slightly modified form of the standard likelihood ratio index has good ability to distinguish between correct and incorrect models when the root-mean-square (RMS) difference between the two models' choice probabilities exceeds 10 to 15 percent. In addition, very small differences between the values of two models' modified likelihood ratio indices indicate with high probability that the model with the lower index value is incorrect. The percent-correctly-predicted statistic is considerably less useful for comparing models. It can fail to distinguish between correct and incorrect models whose choice probabilities differ by at least 25 percent (RMS), even with arbitrarily large estimation samples. Moreover, there are no readily available criteria for determining how large the differences between the values of two models' percent-correctly-predicted statistics must be to justify a conclusion that the model with the lower value likely is incorrect. Several travel-related examples are given.

Travel decisions frequently entail choices among discrete sets of alternatives, such as frequencies, destinations, modes, and routes of travel, and it

often is necessary in travel behavior research and practical transportation studies to be able to predict the outcomes of such choices. In recent years, multinomial logit and multinomial probit models have begun receiving extensive use for this purpose. These models are examples of a broader class of models, called random utility models, that are derived from the behavioral principle of utility maximization. In models of this class, it is assumed that an individual's preferences among the available alternatives can be described with a utility function and that the individual selects the alternative with the greatest utility. The utility of an alternative is represented as the sum of a deterministic and a random component. The deterministic component accounts for systematic effects of observed factors that influence choice whereas the random component accounts for the effects of unobserved factors. The random utility model then predicts the probability that a randomly selected individual with given values of the observed factors will choose a particular alternative (i.e., the probability that the utility of the particular alternative exceeds the utilities of all other alternatives). [See Domencich and McFadden (1) and Hensher and Johnson (2) for detailed discussions of the behavioral foundations of random utility models.]

A logit, probit, or other random utility model constitutes a functional relation between the observed factors (or explanatory variables) and the probabilities that an individual chooses the various alternatives. These relations usually are not known a priori and must be estimated by fitting a model to observations of choices and the explanatory variables. The fitting process usually takes place in two steps. [See McFadden (3) and Daganzo (4) for detailed discussions of the fitting process.] In the first step, the functional form of the relation between the explanatory variables and the choice probabilities is specified up to a finite set of constant parameters. For example, it might be specified that a mode choice model has the multinomial logit functional form with a utility function that is a linear combination of a certain set of travel time and cost variables. The coefficients of these variables in the linear utility function would then constitute the set of constant parameters. In the second step of the fitting process, the values of the parameters are estimated statistically from the observations. The method of maximum likelihood usually is used for this purpose. The statistical theory on which this two-step process is based assumes that the first step is carried out without error. If this assumption is true and certain mild regularity conditions are satisfied, then the maximum likelihood estimates of the parameter values have a variety of desirable statistical properties. Most importantly, the estimated parameter values and the values of the choice probabilities computed from the estimated parameters approach the true values as the size of the estimation data set increases toward infinity.

In practice, of course, the correct functional forms of the relations between the choice probabilities and the explanatory variables are not known a priori, even up to a set of constant parameters. Not surprisingly, use of an incorrect functional form in the second estimation step can lead to models that produce highly erroneous forecasts of travel behavior (5-7). Consequently, the development of empirical random utility models, like the development of most other statistically based models, usually includes comparing several models with different functional forms in an effort to distinguish forms that are likely incorrect from ones that may be correct.

This paper is concerned with comparisons that involve two models. Several procedures are available for carrying out such pairwise comparisons, depending on whether the models being compared are nested or non-nested. Two models are nested if one model can be obtained from the other by assigning appropriate values to the latter model's parameters. In non-nested 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. (Examples of nested and non-nested models are given in the following section of this paper.) Comparisons of nested models usually are carried out by using likelihood ratio or t-tests (3). The statistical properties of these tests are well known, and numerical experiments with the tests have indicated that they have good ability to distinguish between correct and seriously erroneous random utility models in the nested case (6,7).

Comparisons of non-nested models usually are made with so-called goodness-of-fit statistics, such as the likelihood ratio index and the percentage of choices correctly predicted if each individual in the estimation data set is assumed to choose the alternative with the highest probability in the estimated model. The basis for using these statistics for comparing models is largely intuitive. Plausibly, it is assumed that a correct model is likely to have larger values of these statistics than are incorrect models. Thus, models with relatively large values of these statistics are presumed more likely to be correct (or to be less erroneous) than models with relatively low values of these statistics. However, there has been no systematic investigation of the abilities of these statistics to distinguish reliably between correct and incorrect models. This paper reports the results of such an investigation. It is shown that with estimation data sets of practical size, a slightly modified form of the standard likelihood ratio index statistic has good ability to distinguish between correct and incorrect models when the root-mean-square (RMS) difference between the choice probabilities of the two models exceeds 10-15 percent. The percent-correctly-predicted statistic is considerably less reliable. It can fail to distinguish between models whose choice probabilities differ by at least 25 percent (RMS), even if the size of the estimation data set is allowed to increase without bound.

EXAMPLES OF NESTED AND NON-NESTED MODELS

The results presented in this paper apply only to comparisons of non-nested models. Serious errors can result from attempting to apply the results in the nested case. Therefore, it is important to understand the distinction between nested and non-nested models. This distinction is illustrated by the following examples. [See Horowitz (8) for a precise mathematical definition of non-nestedness.]

Example 1 (Nested Models)

Suppose that two logit models are being considered for possible use in forecasting mode choice. For simplicity, assume that there are only two modes, automobile and transit. The two models differ in the specifications of their utility functions. (In the following discussion the term "utility function" will refer to the deterministic component unless otherwise noted.) The utility function for mode i (i = automobile or transit) in model 1 is

$$U_i = \alpha_1 (\text{Time}_i) + \beta_1 (\text{Cost}_i) \quad (1)$$

where

U_i = utility of mode i,
$Time_i$ = travel time of mode i,
$Cost_i$ = travel cost of mode i, and
α_1 and β_1 = constant parameters.

In model 2 the utility function is

$$U_i = \alpha_2 (Time_i) \quad (2)$$

In either model, the probability of choosing mode i is given by the binomial logit function

$$P_i = \exp(U_i)/[\exp(U_1) + \exp(U_2)] \quad (3)$$

where P_i is the choice probability. Then model 2 is nested with model 1, since model 2 can be obtained from model 1 by setting $\alpha_1 = \alpha_2$ and $\beta_1 = 0$.

Example 2 (Non-Nested Models)

Suppose that in the same mode choice study a third model (model 3) is being considered and that its utility function for mode i is

$$U_i = \alpha_3 \log(Time_i) + \beta_3 \log(Cost_i) \quad (4)$$

As in models 1 and 2, the choice probabilities in model 3 are related to the utility function by Equation 3. Then models 1 and 3 and models 2 and 3 form non-nested pairs. Apart from the degenerate case in which all of the parameter values are zero, it is not possible to choose parameter values for models 1 and 3 so that the two models coincide, nor is it possible to choose parameter values for models 2 and 3 so that those models coincide.

CRITERIA FOR EVALUATING COMPARISON PROCEDURES

In this paper the likelihood ratio index and percent-correctly-predicted goodness-of-fit statistics will be evaluated according to their abilities to distinguish between correctly and incorrectly specified models. Two factors must be taken into account in making these evaluations: the abilities of the statistics to distinguish between correct and incorrect models in the absence of random sampling error, and the effects of random sampling error on the comparisons. 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, owing to the effects of unobserved factors (i.e., the random component of the utility function). 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, even if the model involved is correctly specified. 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. 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 insignificantly small by making the estimation sample for a model sufficiently large. Moreover, if the estimation sample is large enough to make the effects of sampling error insignificant, then it always is possible to determine unambiguously if 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 of comparison statistics, such as the goodness-of-fit statistics discussed here, 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 practice, of course, it usually is not possible to work with samples sufficiently large to eliminate the effects of random sampling error. As has already been noted, random sampling error can cause comparison procedures to give misleading results (e.g., to cause an incorrectly specified model to have a more favorable goodness-of-fit statistic than a correctly specified one), even if the procedures always would give correct results in samples large enough to eliminate sampling error. In general, it is not possible to guarantee that a comparison procedure always will distinguish correctly between correct and incorrect models when sampling error is present, particularly if the choice probabilities of the two models would not be greatly different in the absence of sampling error. However, to be useful, a comparison procedure should be capable of making the distinction correctly most of the time (i.e., with high probability) if the two models would have significantly different choice probabilities in the absence of sampling error.

The foregoing discussion suggests that to be useful, a test procedure should have the following two characteristics:

1. It should be consistent. In other words, in the absence of random sampling error it should always distinguish correctly between correctly and incorrectly specified models.
2. It should not be highly sensitive to random sampling error. In other words, in samples of practical size (typically 100-1000 observations) the procedure should have a high probability of rejecting an incorrect model in a comparison with a correct one if the two models would yield substantially different values of the choice probabilities in the absence of sampling error.

It also is desirable that the value of the test statistic associated with a comparison procedure be easy to compute. However, the two procedures discussed in this paper have roughly equal computational requirements so that computational considerations do not provide a basis for a comparative evaluation of the procedures.

In the following two sections, the likelihood-ratio-index and percent-correctly-predicted statistics will be evaluated according to the foregoing two criteria.

LIKELIHOOD RATIO INDEX

The most commonly used form of the likelihood ratio index, and the only form that will be discussed here, is defined as follows. [See Daganzo (4) and Tardiff (9) for definitions and discussions of other forms.] Let L denote the value of a model's log-likelihood function when the values of the model's parameters equal their maximum likelihood estimates. Let L_0 denote the value of the log-likelihood function of a model that assigns equal values to the choice probabilities of all alternatives, regardless of the values of the explanatory variables. Then

the likelihood ratio index, ρ^2, is defined as

$$\rho^2 = 1 - L/L_0 \quad (5)$$

If there are N individuals in the estimation sample and each individual chooses among J alternatives, then L_0 is given by

$$L_0 = -N \log J \quad (6)$$

(Throughout this paper it will be assumed that all individuals face the same number of alternatives. Allowing different individuals to face different numbers of alternatives would add complexity to the presentation without changing the results significantly.)

The likelihood ratio index is a goodness-of-fit statistic for random utility models that is similar in many respects to the coefficient of multiple determination, R^2, in regression models. The larger the value of ρ^2 for a model, the better the model fits the given data. Therefore, two nonnested models P and Q can be compared by comparing the likelihood ratio indices ρ_P^2 and ρ_Q^2 for the two models. If $\rho_P^2 - \rho_Q^2 > 0$, this suggests that model P is superior to model Q, whereas $\rho_P^2 - \rho_Q^2 < 0$ suggests that model Q is superior.

Suppose that model P is correct and model Q is incorrect. Then, to evaluate the likelihood ratio index according to the criteria given in the previous section it is necessary to determine (a) if $\rho_P^2 - \rho_Q^2 > 0$, always, when the sample size N is large enough to make the effects of random sampling error insignificant and (b) if $\rho_P^2 - \rho_Q^2 > 0$ is a high probability outcome in samples of practical size if models P and Q would yield substantially different values of the choice probabilities in the absence of random sampling error. These determinations can be made if the probability distribution of $\rho_P^2 - \rho_Q^2$ is known.

The following notation will be used in describing the probability distribution of $\rho_P^2 - \rho_Q^2$ and in the subsequent discussion. Let P(i,X) denote the true probability that an individual chooses alternative i when the explanatory variables have the value X. (Here, X denotes the entire set of values of all of the explanatory variables of both models. If some elements of X are not variables of one of the models, then the choice probabilities of this model are independent of the values of these elements.) Let Q(i,X) denote the choice probability for alternative i that model Q would yield when the explanatory variables have the value X if there were no random sampling error. P(i,X) and Q(i,X), respectively, are the large-sample limits (i.e., the limits as the sample size approaches infinity) of the maximum likelihood estimates of the choice probabilities of models P and Q. Let $p_x(X)$ denote the proportion of individuals in the population being studied for whom the values of the explanatory variables equal X. Let k_P and k_Q, respectively, denote the numbers of estimated parameters in models P and Q. As before, let N denote the number of individuals in the estimation data set, and let J denote the number of alternatives available to each individual. Finally define Δ^2 by

$$\Delta^2 = \sum_{i,X} \left\{[P(i,X) - Q(i,X)]/P(i,X)\right\}^2 P(i,X)p_x(X) \quad (7)$$

Δ^2 is the weighted mean square fractional error in the choice probabilities that would result from using the incorrect probabilities Q(i,X) in place of the correct probabilities P(i,X). The weight for given i and X values equals the proportion of the population that has explanatory variable values X and selects alternative i.

Note that Δ^2 always exceeds zero unless Q(i,X) = P(i,X) for all i and X. In other words, Δ^2 exceeds zero unless model Q is identical to model P and, therefore, is correctly specified.

The probability distribution of $\rho_P^2 - \rho_Q^2$ is derived in Horowitz (8). It is shown there that $\rho_P^2 - \rho_Q^2$ has approximately the normal distribution with the following mean (μ) and variance (σ^2):

$$\mu = \Delta^2/2 \log J + (k_P - k_Q)/2N \log J \quad (8)$$

$$\sigma^2 = \Delta^2/N (\log J)^2 \quad (9)$$

The accuracy of the approximation increases as the sample size, N, increases. It follows from Equations 8 and 9 that $\rho_P^2 - \rho_Q^2$ exceeds zero, thereby indicating that the correct model is superior to the incorrect one, with the following probability:

$$\Pr(\rho_P^2 - \rho_Q^2 > 0) = \Phi\left[N^{1/2}\Delta/2 + (k_P - k_Q)/2N^{1/2}\Delta\right] \quad (10)$$

where Φ is the cumulative standard normal distribution function.

It is easy to see from Equation 10 that the likelihood ratio index satisfies the consistency criterion given in the previous section. As N approaches infinity, $\Pr(\rho_P^2 - \rho_Q^2 > 0)$ approaches 1. Since the limit of an infinite sample corresponds to eliminating random sampling error, this result implies that in the absence of sampling error, ρ_P^2 always exceeds ρ_Q^2. Thus, if there is no sampling error, the likelihood ratio index always indicates that the correct model is superior to the incorrect one.

It also can be seen from Equations 8-10 that with finite samples, adding parameters to an incorrect model (i.e., increasing k_Q) tends to decrease the value of $\rho_P^2 - \rho_Q^2$ and of $\Pr(\rho_P^2 - \rho_Q^2 > 0)$, even if the variables associated with the added parameters are incorrectly specified or irrelevant to the choices being studied. This clearly is an undesirable characteristic of the likelihood ratio index because it means that in finite samples the index tends to favor models with large numbers of parameters, regardless of whether these models are correct. However, this characteristic can be removed by making a simple modification in the definition of the likelihood ratio index. Define $\bar{\rho}^2$, the modified likelihood ratio index, for a model with k estimated parameters by

$$\bar{\rho}^2 = \rho^2 - k/2N \log J \quad (11)$$

or, equivalently,

$$\bar{\rho}^2 = 1 - (L - k/2)/N \log J \quad (12)$$

The modified statistic $\bar{\rho}^2$ is used in the same way as ρ^2 for comparing two models. Thus, $\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0$ indicates that model P is superior to model Q, and $\bar{\rho}_P^2 - \bar{\rho}_Q^2 < 0$ indicates that model Q is superior.

Equations 10 and 11 imply that the probability that $\bar{\rho}_P^2 - \bar{\rho}_Q^2$ exceeds zero is given by

$$\Pr(\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0) = \Phi(N^{1/2}\Delta/2) \quad (13)$$

It can be seen from Equation 13 that $\bar{\rho}^2$ is consistent and, in contrast to the unmodified likelihood ratio index, is not biased in favor of models with large numbers of parameters. This makes $\bar{\rho}^2$ more

Table 1. Probabilities that modified likelihood ratio index selects correct model (P) in comparison with incorrect model (Q).

N	Δ	$Pr(\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0)$	N	Δ	$Pr(\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0)$
100	0.05	0.60	250	0.15	0.88
	0.10	0.69		0.20	0.94
	0.15	0.77	500	0.05	0.71
	0.20	0.84		0.10	0.87
250	0.05	0.66		0.15	0.95
	0.10	0.79		0.20	0.99

Notes: N = size of the estimation data set.
Δ = RMS difference between the large sample limiting values of the choice probabilities of models P and Q.
$Pr(\bar{\rho}_P^2 - \bar{\rho}_Q^2 > 0)$ = probability that the correct model is selected.

useful than ρ^2 for comparing models. Accordingly, only the modified index $\bar{\rho}^2$ will be used in the remainder of this paper.

The performance of $\bar{\rho}^2$ according to the second criterion given in the previous section can be assessed by computing $Pr(\rho_P^2 - \rho_Q^2 > 0)$ for various values of N and Δ. Table 1 shows the results of such a computation. It can be seen that if the sample size exceeds roughly 250, a comparison of two models using $\bar{\rho}^2$ has a probability of at least 0.80 of selecting the correct model when the RMS percentage difference between the two models' choice probabilities (i.e., 100Δ) exceeds 10 to 15 percent.

The probability distribution of $\bar{\rho}_P^2 - \bar{\rho}_Q^2$ can be used to derive a simple upper bound on the probability that $\bar{\rho}^2$ for an incorrect model (Q) exceeds $\bar{\rho}^2$ for a correct model (P) by an arbitrary amount z. The bound is [see Horowitz (8)]

$$Pr(\bar{\rho}_Q^2 - \bar{\rho}_P^2 > z) \leq \Phi[-(2Nz \log J)^{1/2}] \quad (14)$$

This inequality implies that in moderate size samples, very small differences between the $\bar{\rho}^2$ values of two models indicate with high probability that the model with the lower $\bar{\rho}^2$ value is incorrect. For example, if $N \geq 250$, $z \geq 0.01$, and $J \geq 2$, inequality (15) yields

$$Pr(\bar{\rho}_Q^2 - \bar{\rho}_P^2 > z) \leq 0.03 \quad (15)$$

In other words, if $N \geq 250$ and the $\bar{\rho}^2$ values of two models differ by 0.01 or more, the model with the lower $\bar{\rho}^2$ value almost certainly is incorrect.

PERCENT-CORRECTLY-PREDICTED STATISTIC

The percent-correctly-predicted statistic for a model is obtained by "predicting" that each individual in the model's estimation data set chooses the alternative that has the highest choice probability according to the estimated model. The predictions are compared with the observed choices of the individuals in the estimation data set, and the percentage of correct predictions (i.e., the percentage of individuals for which the predicted and observed choices coincide) is computed. The result yields the percent-correctly-predicted statistic. In this section an example is presented in which the percent-correctly-predicted statistic fails to satisfy either of the previously defined evaluation criteria. The implications of the example for the usefulness of the percent-correctly-predicted statistic are discussed following the presentation of the example.

Notation and Formulas

The example is based on binomial logit models (i.e., logit models of choice between two alternatives). Before presenting the example, it is necessary to present the notation and formulas that will be used in computing the percent-correctly-predicted statistic for the example.

Let model P be correct and model Q be incorrect. As before, let P(i,X) denote the true probability that an individual with explanatory variables X chooses alternative i (i = 1 or 2), and let Q(i,X) denote the large-sample limit of the model Q estimate of the probability that this individual chooses alternative i. Let the number of individuals in the estimation sample be N, and let these individuals be indexed by n (n = 1, 2,..., N). Let $\hat{P}(i,X)$ and $\hat{Q}(i,X)$, respectively, denote the maximum likelihood estimates of the model P and model Q choice probabilities obtained from the estimation sample. Thus, for example $\hat{P}(i,X_n)$ and $\hat{Q}(i,X_n)$ are the estimated probabilities that individual n chooses alternative i by using models P and Q. Note that as N approaches infinity, $\hat{P}$ approaches P and $\hat{Q}$ approaches Q.

For each individual n in the estimation sample define j_n by

j_n = 1 if individual n was observed to choose alternative 1, 0 otherwise (16)

Thus, j_n indicates the observed choice of individual n. Also, for each individual n in the estimation sample define $\hat{S}_P(n)$ and $\hat{S}_Q(n)$ by

$$\hat{S}_P(n) = \begin{cases} 100 \text{ if } \hat{P}(1,X_n) > 1/2 \text{ and } j_n = 1 \text{ or } \hat{P}(1,X_n) < 1/2 \\ \quad \text{and } j_n = 0; \\ 50 \text{ if } \hat{P}(1,X_n) = 1/2; \text{ and} \\ 0 \text{ if } \hat{P}(1,X_n) > 1/2 \text{ and } j_n = 0 \text{ or } \hat{P}(1,X_n) < 1/2 \\ \quad \text{and } j_n = 1 \end{cases} \quad (17)$$

$$\hat{S}_Q(n) = \begin{cases} 100 \text{ if } \hat{Q}(1,X_n) > 1/2 \text{ and } j_n = 1 \text{ or } \hat{Q}(1,X_n) < 1/2 \\ \quad \text{and } j_n = 0; \\ 50 \text{ if } \hat{Q}(1,X_n) = 1/2; \text{ and} \\ 0 \text{ if } \hat{Q}(1,X_n) > 1/2 \text{ and } j_n = 0 \text{ or } \hat{Q}(1,X_n) < 1/2 \\ \quad \text{and } j_n = 1 \end{cases} \quad (18)$$

Then $\hat{S}_P(n)$ equals 100 if individual n was observed to choose the alternative with the larger value of $\hat{P}(i,X_n)$ (i = 1 or 2), $\hat{S}_P(n)$ equals 0 if individual n was observed to choose the alternative with the lower P value, and $\hat{S}_P(n) = 50$ if the two P values for individual n are equal. Thus, $\hat{S}_P(n)$ is the percent-correctly-predicted statistic for model P by using the single individual n. An analogous interpretation applies to $\hat{S}_Q(n)$. The percent-correctly-predicted statistics for models P and Q by using the entire estimation sample, $\hat{S}_P$ and $\hat{S}_Q$, are obtained by averaging $\hat{S}_P(n)$ and $\hat{S}_Q(n)$ over all of the individuals in the sample. Thus,

$$\hat{S}_P = (1/N) \sum_n \hat{S}_P(n) \quad (19)$$

$$\hat{S}_Q = (1/N) \sum_n \hat{S}_Q(n) \quad (20)$$

$\hat{S}_P$ and $\hat{S}_Q$ coincide with the usual definitions of the percent-correctly-predicted goodness-of-fit statistics for models P and Q.

The large-sample limits of $\hat{S}_P$ and $\hat{S}_Q$ can be computed by applying the strong law of large numbers to Equations 19 and 20. This yields the following result:

Table 2. Values of the explanatory variables and choice probabilities for the example.

ΔT (min)	ΔC (cents)	P (auto, ΔT)[a]	P (transit, ΔT)[a]	Q (auto, ΔC)[a]	Q (transit, ΔC)[a]
5	20	0.62	0.38	0.71	0.29
10	5	0.73	0.27	0.56	0.44
20	40	0.88	0.12	0.86	0.14

[a]P and Q are computed as large-sample limits and, therefore, are free of random sampling error. The large-sample limit of the maximum likelihood estimate of β is 0.045.

$$S_P = 100 \sum_x p_x(X) \max[P(1, X), P(2, X)] \quad (21)$$

$$S_Q = 100 \sum_x p_x(X) \left\{ I(X) P(1, X) + [1 - I(X)] P(2, X) \right\} \quad (22)$$

where S_P and S_Q, respectively, denote the large-sample limits of $\hat{S}_P$ and $\hat{S}_Q$ and I(X) is defined by

$$I(X) = \begin{cases} 1 \text{ if } Q(1, X) > 1/2; \\ 1/2 \text{ if } Q(1, X) = 1/2; \text{ and} \\ 0 \text{ if } Q(1, X) < 1/2 \end{cases} \quad (23)$$

Equation 21 has been derived previously by Daganzo (10). S_P and S_Q are the values that the percent-correctly-predicted statistics for models P and Q would have if there were no random sampling error.

Example

Suppose that two binomial logit models, P and Q, of mode choice between automobile and transit are being considered. In model P the probability of choosing automobile is given by

$$P(\text{auto}, \Delta T) = 1/[1 + \exp(-\alpha \Delta T)] \quad (24)$$

where Δ T is transit travel time minus automobile travel time, and α is a constant. In model Q the probability of choosing automobile is given by

$$Q(\text{auto}, \Delta C) = 1/[1 + \exp(\beta \Delta C)] \quad (25)$$

where Δ C is transit travel cost minus automobile travel cost and β is a constant. The transit choice probabilities are equal to one minus the automobile choice probabilities.

Suppose that model P is correct and that the true value of α is 0.10. Suppose also that in the population being studied there are only three possible combinations of values of Δ T and Δ C, as shown in Table 2, and that these combinations occur with equal probability. Thus $p_X(X) = 1/3$ for all values of the explanatory variables of both models. Given the values of α and Δ T, the values of the true choice probabilities P(i, Δ T) (i = automobile or transit) can be computed from Equation 25, and the large-sample-limit of $\hat{Q}$(i, Δ C) can be computed by using methods described in Horowitz (6). The results are shown in Table 2. It can be seen from the table that the differences between the true choice probabilities P and large-sample-limit model Q choice probabilities vary from 2 to 63 percent. The RMS percentage difference, as computed from Equation 7, is 25 percent. Thus, the true and erroneous models yield substantially different values of the choice probabilities in the absence of random sampling error.

The large-sample limits of the percent-correctly-predicted statistics for models P and Q can be computed from Equations 21 and 22. The result is $S_P = S_Q = 74$ percent. Thus, the large-sample limits of the percent-correctly-predicted statistics for the two models are equal. This means that when there is no random sampling error, these statistics provide no information useful for distinguishing between the correct model P and the incorrect model Q, even though there are large differences between the choice probabilities of the two models. Clearly, the percent-correctly-predicted statistic fails to satisfy the first of the previously defined evaluation criteria in this case.

It also can be shown that the percent-correctly-predicted statistic is not useful for distinguishing between models P and Q with finite samples. (This is to be expected because reducing the sample size from infinity reduces the information content of the sample.) With a finite sample, the percent-correctly-predicted statistics of models P and Q can differ only if there are values of Δ T and Δ C for which the alternative with the highest estimated model P choice probability is different from the alternative with the highest estimated model Q choice probability. In terms of the notation developed in the previous subsection, this means that there must be pairs of values of Δ T and Δ C such that either $\hat{P}$(auto, Δ T) > 0.5 and $\hat{Q}$(auto, Δ C) < 0.5 or $\hat{P}$(auto, Δ T) < 0.5 and $\hat{Q}$(auto, Δ T) > 0.5, as can be seen from Equations 18 and 19. The probability that either or both of these events occurs can be computed from the information in Table 2 by using methods described by Daganzo (4,11) and Horowitz (12). The result is that with an estimation data set of 100 or more observations, the probability that either or both of these events occurs is virtually zero (i.e., less than 5×10^{-6}). Thus, the percent-correctly-predicted statistics of models P and Q are virtually certain to be equal with any reasonable estimation sample size. It follows that the percent-correctly-predicted statistic almost certainly will fail to distinguish between the two models with any reasonably sized estimation sample. Thus, the statistic fails to satisfy the second of the previously defined evaluation criteria.

The performance of the percent-correctlypredicted statistic in this example may be contrasted with that of the modified likelihood ratio index. The Δ value for comparing models P and Q is 0.25. It follows from Equation 13 that $\bar{\rho}_P^2 - \bar{\rho}_Q^2$ exceeds zero with probability 0.89 if the sample size is 100 and probability 0.98 if the sample size is 250. The probability approaches 1.0 as the sample size increases further. Thus, in contrast to the percent-correctly-predicted statistic, the modified likelihood ratio index has a high probability of distinguishing correctly between models P and Q with any reasonably sized estimation sample.

Discussion of Example

The foregoing example shows that the percent-correctly-predicted statistic may fail to distinguish between a correct and an incorrect model, even if the choice probabilities of the two models differ substantially. Of course, this does not mean that the statistic always will fail to make such distinctions. On the contrary, it is possible to construct examples in which the statistic distinguishes between correct and incorrect models quite satisfactorily. However, the example shows that the statistic is an unreliable diagnostic tool. The fact that two models have similar values of the percent-correctly-predicted statistic does not necessarily

imply that the two models have similar abilities to explain the available data or are equally likely to be correct.

Although small differences between the values of two models' percent-correctly-predicted statistics do not necessarily imply that the two models are equally satisfactory, the possibility remains that a large difference between the values of the statistics implies with high probability that the model with the lower value is incorrect. This possibility clearly is fulfilled qualitatively. However, it does not appear to be possible to develop for the percent-correctly-predicted statistic an inequality analogous to inequality 14 for the modified likelihood ratio index. Thus, it does not appear possible, at least analytically, to specify quantitatively a minimum difference between the values of two models' percent-correctly-predicted statistics that enables one to conclude with high probability that the model with the lower value is incorrect. As was discussed in connection with inequality 14, it is possible to specify such a minimum difference for the modified likelihood ratio index. Although the apparent lack of a "minimum significant difference" for the percent-correctly-predicted statistic is disappointing and clearly impairs the statistic's usefulness, it should not be considered surprising or unusual. For example, there also is no known minimum significant difference for the R^2 values of two linear regression models.

CONCLUSIONS

The results presented here indicate that the modified likelihood ratio index provides a powerful method for comparing non-nested random utility models. It has been shown that very small differences between the values of the modified likelihood ratio indices of two models indicate with high probability that the model with the lower index value is incorrect. Moreover, if one of the models being compared is correct and there are substantial differences between the choice probabilities of the correct and incorrect models, the modified likelihood ratio index has a high probability of indicating that the correct model is superior to the incorrect one with estimation samples of practical size.

The percent-correctly-predicted statistic is much less useful for comparing models. The statistic may fail to distinguish between correct and incorrect models, regardless of sample size, even if the choice probabilities of the two models are very different. Moreover, there are no readily available quantitative criteria for determining how large the differences between the values of two models percent-correctly-predicted statistics must be to justify a conclusion that the model with the lower value likely is incorrect.

In summary, the modified likelihood ratio index is a much more useful tool for comparing non-nested random utility models than is the percent-correctly-predicted statistic.

ACKNOWLEDGMENT

The views expressed in this paper are mine. They are not necessarily endorsed by the U.S. Environmental Protection Agency.

REFERENCES

1. T.A. Domencich and D. McFadden. Urban Travel Demand: A Behavioral Analysis. American Elsevier Publishing Company, New York, 1975.
2. D.A. Hensher and L.W. Johnson. Applied Discrete-Choice Modelling. Croom-Helm, London, 1981.
3. D. McFadden. Analysis of Qualitative Choice Behavior. In Frontiers in Econometrics (P. Zarembka, ed.), Academic Press, New York, 1974.
4. C. Daganzo. Multinomial Probit: The Theory and Its Application to Demand Forecasting. Academic Press, New York, 1979.
5. J.L. Horowitz. Sources of Error and Uncertainty in Behavioral Travel Demand Models. In New Horizons in Travel-Behavior Research (P.R. Stopher, A.H. Meyburg, and W. Brog, eds.), Lexington Books, Lexington, MA, 1981.
6. J. Horowitz. Identification and Diagnosis of Specification Errors in the Multinomial Logit Model. Transportation Research, Vol. 15B, 1981, pp. 345-360.
7. J. Horowitz. The Accuracy of the Multinomial Logit Model as an Approximation to the Multinomial Probit Model of Travel Demand. Transportation Research, Vol. 14B, 1980, pp. 331-341.
8. J.L. Horowitz. Statistical Comparison of Non-Nested, Discrete-Choice, Random-Utility Models. U.S. Environmental Protection Agency, 1981.
9. T.J. Tardiff. A Note on Goodness-of-Fit Statistics for Probit and Logit Models. Transportation, Vol. 5, 1976, pp. 377-388.
10. C.F. Daganzo. Goodness-of-Fit Measures and the Predictive Power of Discrete Choice Models. TRB, Transportation Research Record 874, 1982, pp.13-19.
11. C.F. Daganzo. The Statistical Interpretation of Predictions with Disaggregate Demand Models. Transportation Science, Vol. 13, 1979, pp. 1-12.
12. J. Horowitz. Confidence Intervals for the Choice Probabilities of the Multinomial Logit Model. TRB, Transportation Research Record 728, 1979, pp. 23-30.

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