Goodness-of-Fit Measures and the Predictive Power of Discrete Choice Models

CARLOS F. DAGANZO

Although a number of goodness-of-fit measures for discrete choice models have been proposed and are widely in use, there have been few attempts at interpreting their physical meaning at a practical level. This paper presents a family of goodness-of-fit measures, which contains currently used measures such as the pseudo-correlation coefficient and the percent right, and shows how its members are related. More important, it is shown that one of these measures has an interpretation identical to the correlation coefficient of multiple regression in that it can be used to calculate the expected reduction in the root-mean-square prediction error afforded by a model. This goodness-of-fit measure, d, is uniquely defined for binary models, and can be approximated by an easy to calculate consistent statistic, D. For models with more than two alternatives, the same can be done, but the measure depends on the alternative under consideration. The d-measure usually takes values in between the commonly used normalized percent right measure and the pseudo-correlation coefficient.

A discrete choice model is a formula that relates some explanatory variables, X (sometimes called attributes), to the probability that an individual chooses one of a set of alternatives. The expression given below is the discrete choice model itself. It assumes that there are I alternatives, i, and is called the choice function:

$$\Pr[\text{choose } i|X] = P_i(X),\ i = 0, 1, \ldots, I-1 \tag{1}$$

The closer the probabilities given by Equation 1 are to either zero or one, the more unequivocal will be the predictions of the model. A true model--it is assumed that Equation 1 does not contain specification errors--that gives values close to either zero or one for most of the values of X that are likely to be encountered in practice is superior to a model that gives probabilities that barely depend on X.

The distinction of good and bad models can be captured by many measures of goodness (of fit), and a purpose of this paper is to discuss these in a unified way. The multiple linear regression analog of these measures is the correlation coefficient. Ultimately, because a model is as good as its predictions, it will be shown that there is a goodness-of-fit measure, d, which is related to prediction error reduction in the same way as the correlation coefficient of regression analysis. Thus, the model goodness can be readily interpreted by analysts used to working with regression models.

The d-measure will be related to other commonly used goodness-of-fit indicators such as the "percent predicted right" indicator of success tables and the pseudo-correlation coefficient of McFadden (1). Hauser (2) provides a good review of goodness-of-fit measures to which the reader is referred.

ON MS PREDICTION ERRORS FOR REGRESSION MODELS

The linear regression model is

$$Y = \beta \cdot X^T + \epsilon \tag{2}$$

where Y is a continuous dependent variable, β and X are row vectors of constants and explanatory variables, and ε is a zero mean random error term that is independent of X and of the error terms of other observations.

The correlation coefficient, ρ, of a regression model (see Equation 2) can be expressed as follows:

$$1 - \rho^2 = [\text{var}(\epsilon)]/[\text{var}(Y)] \tag{3}$$

This is well known, and we also note that

$$\begin{aligned}\text{var}(Y) &= E_X[\text{var}(Y|X)] + \text{var}_X[E(Y|X)] \\ &= [\text{var}(\epsilon)] + \text{var}\,[(\beta \cdot X^T)]\end{aligned} \tag{4}$$

In the regression terminology, var(ε) is called the unexplained variance and var(β • X^T) the explained variance. It should be noted that the ratio of explained to total variance (ρ^2) can be increased for the same real world phenomena by increasing the spread in the sampling distribution of X. It cannot be increased, however, by increasing the sample size.

The correlation coefficient can also be related to an average reduction in prediction error that is afforded by the model if the sampling distribution of X coincides with the distribution of values for which predictions are desired.

For a model without explanatory variables, the predictions will be constant, y_o. The mean squared prediction error for a model without explanatory variables, MSE, when the prediction is y_o is

$$MSE = E(Y - y_o)^2 = [\text{var}(Y)] + [E(Y) - y_o]^2$$

Because E(Y) is the value of y_o that minimizes the MSE, we assume that $y_o = E(Y)$ and denote it by $\bar{Y}$. Then,

$$MSE = \text{var}(Y) \tag{5}$$

We note that the MSE for any given value of X, MSE_X, is

$$\begin{aligned}MSE_X &= E(Y - \bar{Y}^2|X) \\ &= \text{var}(Y - \bar{Y}|X) + (\beta \cdot X^T - \bar{Y})^2 \\ &= \text{var}(Y|X) + [\beta \cdot (X - \bar{X})^T]^2 \\ &= \text{var}(\epsilon) + [\beta \cdot (X - \bar{X}^T)]^2\end{aligned} \tag{6}$$

That is, MSE_X is made up of a random component, var(ε), and a systematic component, $[\beta \cdot (X - \bar{X})^T]^2$. As shown below, the systematic component is removed by the model with explanatory variables. Because Equation 5 is the average of Equation 6 over the distribution of X: $MSE = E_X(MSE_X)$. MSE will be called the average mean squared prediction error.

For the model with explanatory variables, the mean squared prediction error for a given value of X, MSE_X^W, can be shown to be independent of X. Since the model prediction is β • X^T, we may write

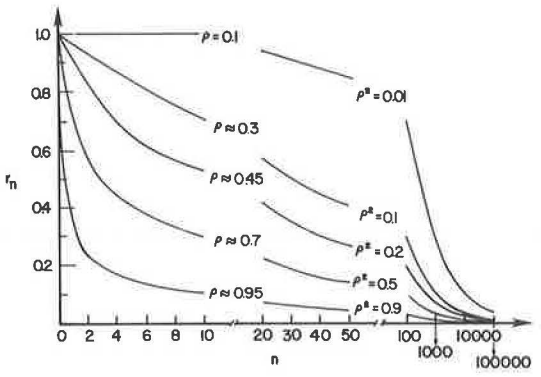
$$MSE_X^W = E[(Y - \beta \cdot X^T)^2|X] = E[\epsilon^2|X] = \text{var}(\epsilon) \tag{7}$$

Thus, the average mean squared prediction error, MSE^W, is also given by var(ε). Clearly, then, the interpretation of the correlation coefficient, ρ, that is related to prediction errors is

$1 - \rho^2$ = average mean squared prediction error with model ÷ average mean squared prediction error without model

and one can interpret $\sqrt{1 - \rho^2}$ as the after/

Figure 1. Residual prediction error as a fraction, r_n, of the prediction error without a model for different values of the correlation coefficient, ρ, and the prediction group size, n.



before ratio of prediction errors that can be expected. This ratio will be noted by r.

Instead of predicting the dependent variable for one observation, if one wants the average for a set of n observations with the same value of X, the prediction errors will be reduced. A bar will be used over the MSE symbol to denote that Y is an average of n observations. The same arguments, as before, yields

$$\overline{MSE^W} = \overline{MSE_X^W} = (1/n)\text{var}(\epsilon) \tag{8}$$

This is the equivalent of Equation 7. For the model without explanatory variables, only the random component of the prediction errors is diminished:

$$\overline{MSE_x} = (1/n)\text{var}(\epsilon) + [\beta \cdot (X - \bar{X})^T]^2 \tag{9}$$

Of course,

$$\overline{MSE} = E_x(\overline{MSE_x}) = (1/n)\text{var}(\epsilon) + \text{var}(\beta \cdot X^T) \tag{10}$$

which is the equivalent of Equation 4. The before/after prediction error ratio for an average of n observations is defined as above:

$$r_n = \sqrt{MSE^W/\overline{MSE}} = \sqrt{(1-\rho^2)/(1-\rho^2) + n\rho^2} \tag{11}$$

This well-known expression is plotted on Figure 1 to clarify further the relationship between ρ and prediction errors.

The next section contains a similar discussion for binary discrete choice models and introduces a measure, d, of the difference in prediction accuracy with and without a model. The rest of the paper attempts to relate the different goodness-of-fit measures that exist.

THE d-MEASURE

Let us consider binary discrete choice models. For these models $P_1(X) = 1 - P_0(X)$, and the dependent variable is either zero or one. It is assumed, however, that $P_i(X)$ is strictly between zero and one. For any given value of X, the expected value of the Bernoulli dependent variable, i, is $P_1(X)$. With this in mind, we can repeat the steps noted in the earlier section on MS prediction errors for regression models.

If y_o is the predicted value of i for a model without explanatory variables, the MSE is

$$MSE = E(i - y_o)^2 = \text{var}(i) + [E(i) - y_o]^2$$

The value of y_o that minimizes MSE is E(i), the fraction of the population of individuals for which predictions are desired that chooses alternative 1. This value will be called f_1 and clearly $\text{var}(i) = f_1(1 - f_1)$. The trivial cases with $f_1 = 0$ or $f_1 = 1$ will not be considered in this paper. This is consistent with the postulate that $P_1(X)$ is different from zero and one since there are no specification errors. From now on, we assume that $y_o = f_1$. Thus,

$$MSE = f_1(1 - f_1) \tag{12}$$

which is the equivalent of Equation 5.

Analogously,

$$MSE_x = E[(i - f_1)^2|X] = \text{var}(i|X) + [E(i|X) - f_1]^2 = P_1(X)[1 - P_1(X)] + [P_1(X) - f_1]^2 \tag{13}$$

As in Equation 6, $P_1(X)[1 - P_1(X)]$ is a random component and $[P_1(X) - f_1]^2$ is a removable systematic component. The difference is that the random component is not fixed. Nevertheless, it is true that $MSE = E_X(MSE_X)$, and MSE can still be called legitimately the average mean squared prediction error.

For models with explanatory variables, MSE_x^W is no longer independent of X. Let $y_o(X)$ be the prediction, and write

$$MSE_x^W = E\{[i - y_o(X)]^2|X\} = \text{var}(i|X) + [E(i|X) - y_o(X)]^2 = P_1(X)[1 - P_1(X)] + [P_1(X) - y_o(X)]^2$$

Since a prediction of $y_o(X) = P_1(X)$ minimizes the mean squared error, it will be assumed to be that way from now on. Thus,

$$MSE_x^W = P_1(X)[1 - P_1(X)] \tag{14}$$

which as with regression is the random component of MSE_X.

Unlike in regression, however, the average mean squared prediction error must be calculated. It is

$$MSE^W = E_X(MSE_x^W) = E_X[\text{Var}(i|X)] = E\{P_1(X)[1 - P_1(X)]\} \tag{15}$$

The mean of the (removed) systematic component is

$$E_X\{[P_1(X) - f_1]^2\} = \text{var}_X[P_1(X)]$$

which is similar to the expression of the systematic error of the regression model.

By analogy to the correlation coefficient of regression, let us define a measure of goodness-of-fit, d, which will capture the difference that the model makes:

$$1 - d^2 = MSE^W/MSE = E_X\{P_1(X)[1 - P_1(X)]\}/f_1(1 - f_1) \tag{16a}$$

Alternatively,

$$d^2 = \text{var}_X[P_1(X)]/f_1(1 - f_1) \tag{16b}$$

It is not difficult to realize that Figure 1 also applies to the d-measure. This interpretation is even more important for discrete choice models because (at least in most transportation applications) one is usually interested in predicting the behavior of groups of people with more than one person ($n > 1$).

ESTIMATION OF d: THE D-STATISTIC

The value of d can be approximated from the data used for estimation. Assume that a random, attribute-

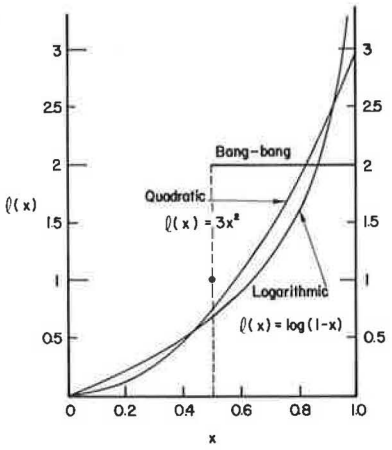


Figure 2. Quadratic, logarithmic, and bang-bang loss functions.

based sample is used and define $\bar{P}_1$ equals the mean estimated choice probability of alternative 1 in the sample, and S_p^2 equals the sample variance of the estimated choice probability for alternative 1 (or 0--they are the same) in the sample.

Then, the following statistic,

$$D = \sqrt{S_p^2/\bar{P}_1(1-\bar{P}_1)} \tag{17}$$

intuitively approximates d for large samples. Appendix A proves that under regularity conditions, D is consistent.

To show that $D \in [0, 1]$, consider the set of estimated alternative 1 choice probabilities for all the observations in the sample. The sample mean of these probabilities was denoted $\bar{P}_1$ and the second sample moment will be denoted $\bar{\bar{P}}_1$. Because all the probabilities are between zero and one, we may write $\bar{\bar{P}}_1 \leq \bar{P}_1$, or subtracting $\bar{P}_1^2$ on both sides, $\bar{P}_1 - \bar{P}_1^2 \leq \bar{P}_1(1 - \bar{P}_1)$. Since the sides of this inequality are the numerator and denominator of Equation 17, it is clear that $D \leq 1$. That $D \geq 0$ is obvious.

FRAMEWORK FOR COMPARISON

Let us now investigate how the d-measure is related to other measures currently in use. To do this more easily, the measures will be derived from a common model.

Let us assume that when an observation is taken, one experiences the following penalty for predicting y as the choice probability for alternative 1 [$y \in (0, 1)$] when the choice is actually i (i = 0, 1):

$$L(y, i) = \ell(|y - i|),$$

where $\ell(\cdot)$ is a real-valued, nondecreasing loss function defined in the (0, 1) interval, with $\lim_{x \to 0} \ell(x) = 0$. The following are three possibilities for $\ell(x)$; they are labeled A, B, and C since they will be used in the sequel:

(A) Quadratic loss: $\ell(x) = x^2$.

(B) Logarithmic loss: $\ell(x) = -\log(1 - x)$

(C) Bang-bang loss:
$$\ell(x) = \begin{cases} 0 & \text{if } x < (1/2) \\ 1/2 & \text{if } x = (1/2) \\ 1 & \text{if } x > (1/2). \end{cases}$$

Figure 2 depicts these functions. (Because the scaling of the loss functions does not affect the goodness-of-fit measures that one derives, the scales in the figure are such that $\int_0^1 \ell(x)dx = 1$ in all three cases.)

The goodness of a model can, therefore, be measured by comparing the expected loss for a prediction with and without the model.

If we do not have a model, a prediction, $y = P_1$, that minimizes the expected loss (if it exists) is found by

$$\min_{y \in (0,1)} \left\{E_i[\ell(|y - i|)]\right\} = \min_{y \in (0,1)} \left\{f_1 \ell(1 - y) + (1 - f_1)\ell(y)\right\} \tag{18}$$

where, as before, f_1 is the fraction of the population choosing 1. As is commonly done in Bayesian analysis, only loss functions for which $P_1 = f_1$ are considered. Otherwise, there would be a reward for providing the wrong prediction, and this should be avoided. Loss functions for which $P_1 = f_1$ will be called regular.

For regular loss functions that are differentiable in [0, (1/2)], the derivative, $\ell'(x)$, of the loss function must satisfy the following (necessary) symmetry condition in (0, 1) if the minimum of Equation 18 is to be equal to f_1:

$$[\ell'(x)]/x = \ell'(1 - x)/1 - x, x \neq (1/2) \tag{19}$$

Cases A, B, and C all satisfy Equation 19 and are regular.

The minimum of Equation 18 is unique only for cases A and B, however.

A regular loss function that is nondecreasing and differentiable in [0, (1/2)] may have a discontinuity at x = (1/2). (See case C.)

For regular loss functions, the expected loss without a model, L, is calculated with a prediction, $P_1 = f_1$:

$$L = \psi(f_1) = f_1 \ell(1 - f_1) + (1 - f_1)\ell(f_1) \tag{20}$$

The next theorem establishes some important properties of $\psi(x)$.

Theorem 1: Assume $\ell(x)$ is real-valued, nonnegative, nondecreasing, regular, differentiable in [0, (1/2)], and such that $\lim_{x \to 0} \ell(x) = 0$. Then, $\psi(x)$ is continuous, bounded, concave, reaches a maximum at x = 1/2, and is such that $\lim_{x \to 0,1} \psi(x) = 0$.

Proof: Because it is nonnegative, $\psi(x)$ is bounded and must satisfy

$$\psi(x) \leq x\ell(1/2) + (1 - x)\ell(1/2) = \ell(1/2)$$

since by regularity $\psi(x)$ is the g.l.b. of Equation 18. Thus, $\ell(1/2)$ bounds $\psi(x)$ if $\psi(x)$ is regular. Since $\psi(1/2) = \ell(1/2)$, $\psi(x)$ reaches a maximum at x = 1/2.

To prove that $\psi(x)$ is continuous, it suffices to show that $\psi(x) \to \psi(1/2)$ as $x \to (1/2)$.

Because $\ell(x)$ is regular and $\psi(1/2) = \ell(1/2)$, Equation 18 with $f_1 = (1/2)$ enables us to write:

$$\psi(1/2) = \ell(1/2) \leq (1/2)\left\{\ell[(1/2) + \epsilon] + \ell[(1/2) - \epsilon]\right\}, \forall \epsilon \in [0,(1/2)] \tag{21}$$

For $f_1 = (1/2) + \epsilon$, Equation 18 yields

$$\psi[(1/2) + \epsilon] = 1/2\left\{\ell[(1/2)+\epsilon] + \ell[(1/2) - \epsilon]\right\} - \epsilon\left\{\ell[(1/2) + \epsilon] - \ell[(1/2) - \epsilon]\right\} \leq \psi(1/2), \forall \epsilon \in [0, (1/2)] \tag{22}$$

The inequality stems from the fact that $\psi(1/2)$ is a maximum of $\psi(x)$. It is clear from Equations 21 and 22 that

$$|\psi[(1/2)] - \psi[(1/2) + \epsilon]| \leq \epsilon\left\{\ell[(1/2)+\epsilon] - \ell[(1/2) - \epsilon]\right\}, \forall \epsilon \in [0,(1/2)]$$

and since the RHS goes to zero when $\epsilon \to 0$, $\psi[(1/2) + \epsilon] \to \psi[(1/2)]$. This establishes continuity. Concavity is next.

By using Equation 19 we see that $\psi'(x) = \ell(1 - x) - \ell(x)$, $x \in [0, (1/2)]$. Clearly, $\psi'(x)$ is nonnega-

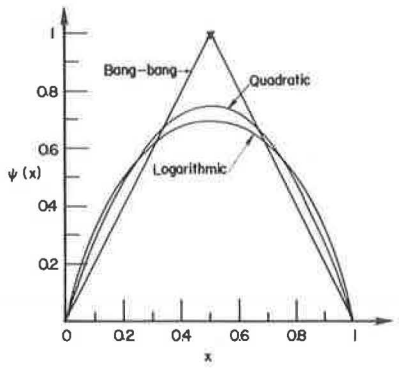


Figure 3. $\psi(x)$ for quadratic, logarithmic, and bang-bang loss functions.

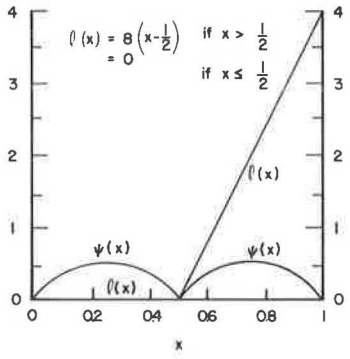


Figure 4. Example of irregular loss function.

tive and nonincreasing. Thus, $\psi(x)$ is concave and nondecreasing in [0, (1/2)].

Since $\psi(x)$ is continuous at 1/2 it is also continuous, concave, and nondecreasing in [0, (1/2)]. The concavity of $\psi(x)$ in (0, 1) follows immediately from the above properties and the symmetry of $\psi(x)$ around 1/2. (Alternatively, the reader can verify that $\{\partial\psi(x) = \psi'(x)$ if $x \neq (1/2)$, $\partial\psi(x) = 0$ if $x = (1/2)\}$ is a subgradient of $\psi(x)$.

To show that $\lim_{x\to 0,1} \psi(x) = 0$, it suffices to show that $(1 - x)\ell(x) \to 0$ as $x \to 1^-$. If $\ell(x)$ is bounded in a neighborhood of 1, this is obvious. Otherwise, $\ell(x) \to \infty$ as $x \to 1^-$ and the limit is the same as

$$\lim_{x\to 1^-} (1 - x)\ell(x) = \lim_{x\to 1^-} (1 - x)^2\ell'(x)$$

because of l'Hôpital's rule. ($[\ell(x)]$ is differentiable in [(1/2), 1] because of the symmetry condition implied by regularity.) By using Equation 19,

$$\lim_{x\to 1^-} (1 - x)^2\ell'(x) = \lim_{x\to 1^-} x(1 - x)\ell'(1 - x) = 0 \cdot \text{QED}$$

Figure 3 plots $\psi(x)$ for cases A, B, and C. Note in particular that $\psi(x)$ is continuous at x = (1/2) for the bang-bang case. Figure 4 illustrates that the theorem does not hold for loss functions that are not regular.

The next step will be to calculate the expected prediction loss when we use the true model $P_1(X)$ and to compare that with $\psi(f_1)$ to derive a goodness-of-fit measure that will capture the reduction in loss achieved by the model.

Goodness-of-Fit Measures from Regular Loss Functions

Given X, the probability that i = 1 is $P_1(X)$ and because the loss function is regular, the optimal prediction is $P_1(X)$. The expected loss conditional on X, with the model is:

$$L_x^W = E_i\{\ell[|i - P_1(X)|] \mid X\} = P_1(X)\ell[1 - P_1(X)] + [1 - P_1(X)]\ell[P_1(X)] = \psi[P_1(X)]$$

The unconditional expected loss with the model is

$$L^W = E_X\{\psi[P_1(X)]\} \tag{23}$$

Note that $E_X[P_1(X)] = f_1$.

It makes sense to define the goodness-of-fit measure, g, to be the fractional reduction in expected loss from the introduction of the model:

$$g = 1 - (L^W/L).$$

But before one can do that, it is important to find which loss functions result in values of g that range from zero to one; for otherwise, the loss function would not be reasonable.

Theorem 2: If Theorem 1 holds, g ranges from zero to one.

Proof: First we show the $L^W \in [0, L]$. Note that (a) $L^W \geq 0$ because $\psi(x) \geq 0$, and (b) $L^W = E_X\{\psi(P_1(X))\} \leq \psi\{E_X[P_1(X)]\} = \psi(\bar{f}_1) = L$ because, by virtue of theorem 1, $\psi(\cdot)$ is concave. Thus, $L^W \in [0, L]$.

That the limit values 0 and L can be reached is clear because if $P_1(X)$ is constant, it must equal f_1, and $L_X^W = L^W = L$. On the other hand, if $P_1(X)$ takes values differing from zero and one by less than ε except for a set of X-values with probability measure less than ε, $L_X^W \to 0$ as $\varepsilon \to 0$ except for a set of X-values of measure zero. Thus, as $\varepsilon \to 0$, $L^W \to 0$, and g can be made arbitrarily close to one by letting $P_1(\cdot)$ resemble a simple function, $\delta(\cdot)$, sufficiently well.

It follows that any value of g in the [0, 1] interval can be attained with a convex combination of $P_1(X) = f_1$ and $P_1(X) = \delta(X)$. QED

The quadratic, logarithmic, and bang-bang losses yield three well-known goodness-of-fit measures.

(A) Quadratic Loss, g_q

From Equations 12 and 20 and by using the quadratic loss formula in the latter, we see that $L \equiv MSE$. Similarly, Equations 15 and 23 yield that $L^W \equiv MSE^W$. Thus, $g_q = d^2$, and the quadratic loss function generates the d-measure.

(B) Logarithmic loss, g_ℓ

Replacing the logarithmic loss function, $\ell(x)$, into Equations 20 and 23, we find that the expected loss, L, is

$$L = -f_1 \log f_1 - (1 - f_1)\log(1 - f_1),$$

and that L^W is

$$L^W = E_X\{-p_1(X)\log P_1(X) - [1 - P_1(X)]\log P_1(X)\}.$$

If, as stated earlier, the distribution of X used represents its sampling distribution (it is assumed that sampling is not choice-based), this expression is the negative expected value of log-likelihood function (for one observation):

$$\mathcal{L}(X) = \log p_1(X) \quad \text{if } i = 1$$
$$= \log[1 - P_1(X)] \quad \text{if } i = 0,$$

and

$$L^W = E_X[\mathcal{L}(X)].$$

Thus, L^W/L represents the reduction in log-likelihood that is achieved by introduction of the model and g_ℓ is the square of the pseudo-corre-

lation coefficient that is commonly used [1, 2].

(C) Bang-bang loss, q_b

Repeating the steps with the new $\ell(x)$, we find that

$$L = \min(f_1, 1 - f_1),$$

and

$$L^W = E_X\{\min[P_1(X), 1 - P_1(X)]\}.$$

The quantity, $\min\{P_1(X), 1 - P_1(X)\}$, represents the fraction of times that we will be wrong if for a choice-maker with attributes X we predict choice 1 if $P_1(X) > 0.5$ and choice 0 if $P_1(X) < 0.5$. Thus, L^W represents the "% wrong" measure that is used with success tables, and q_b is the reduction in the percentage of "wrong" predictions that is achieved by introduction of the model.

It should be clear that other goodness-of-fit measures can be derived by proper redefinition of $\ell(x)$ and that different measures may be called for depending on the application. The discussion at the outset of this paper and in the section on the d-measure argued in favor of the d-measure when the application aim is the prediction of market shares for large groups of individuals.

The next subsection discusses the relative magnitudes of q_q, q_ℓ, and q_b and gives a formula relating q_q and q_ℓ when they are small.

RELATIONSHIP AMONG GOODNESS-OF-FIT MEASURES

Recall that L is the height of the curve on Figure 3 (up to a factor of 2) when the abscissa is f_1, and that L^W is the average of the height when the abscissa varies with $P_1(X)$. This scaling does not affect the value q_b that is ultimately obtained because q_b depends on the ratio L^W/L.

If the variation in the abscissa is small and f_1 is substantially different from (1/2), the bang-bang function will yield $L \simeq L^W$ (and $q_b \simeq 0$) since the function is linear over the relevant range of $P_1(X)$. On the other hand [and also for small $\text{var}\ P_1(X)$], if $f_1 = (1/2)$, L is 1 and $L^W = 1 - 2E_X[|P_1(X) - (1/2)|]$. Then, $q_b = 2E_X[|P_1(X) - (1/2)|]$. This illustrates that for the bang-bang function and for models with low variance for $P_1(X)$ (as usually occurs), q_b will be smaller when f_1 is substantially different from (1/2).

Since the quadratic and logarithmic losses yield smooth $\psi(x)$'s, we approximate L^W taking expectations on a two-term Taylor series expansion:

$$L^W \cong L + \psi''(f_1)\text{var}_X[P_1(X)]; \text{var}_X[P_1(X)] \to 0$$

$$g \cong |\psi''(f_1)|/L\ \text{var}_X[P_1(X)]; \text{var}_X[P_1(X)] \to 0 \quad (24)$$

Replacing in Equation 24 $\psi''(f_1)$ and L by the appropriate values we find

$$g_q/g_\ell \cong 2[-(1 - f_1)\log(1 - f_1) - f_1 \log f_1] = 2L_\ell; \text{var}_X[P_1(X)] \to 0 \quad (25)$$

In this expression, L_ℓ is the logarithmic loss without model. It is clear from Figure 3 that $q_q > q_\ell$ if f_1 is between (approximately) 0.2 and 0.8, and that $q_q < q_\ell$ otherwise. If the probabilities are very unbalanced and $\text{var}_X[P_1(X)]$ is small: $0 \simeq q_b < q_q < q_\ell$, but if $f_1 \simeq 0.5$, $q_\ell < q_q < q_b$. This is because if $f_1 \simeq 0.5$, q_b is comparable with $(\text{var}_X[P_1(X)])^{(1/2)}$ but q_q and q_ℓ are comparable with $\text{var}_X[P_1(X)]$.

While Equation 25 allows us to approximate d from the output of most computer programs with

$$d \cong \sqrt{g_q} \cong \sqrt{2[(\hat{\mathcal{L}} - \bar{\mathcal{L}})/N]} = \ell \quad (26)$$

where $\bar{\mathcal{L}}$ is the background log-likelihood [3], $\hat{\mathcal{L}}$ is the maximum log-likelihood, and N is the sample size, it is recommended to calculate d exactly by simple modification of the computer programs.

Equation 26, thus, suggests that estimated log-likelihood values can actually be used to assess prediction errors. It should be remembered, however, that the approximation is only valid when $(\text{var}_X[P_i(X)])^{(1/2)}$ is so small that values of $P_i(X)$ outside the range where the two-term Taylor series expansion of $\psi(x)$ is a good approximation are rare. For this to happen, the d-measure must be small, which happens quite often in practice.

EXAMPLE

Assume that we have the following binary probit model:

$$P_1 = 1 - P_2 = \Phi(\theta X),$$

where Φ is the standard normal c.d.f., and θ and X are scalars. Furthermore, assume that the sampling distribution of X, F(x), is normal with zero mean and variance one.

As the value of θ increases, most of the individuals in the population face choices that can be predicted with less and less uncertainty because unless $X = 0$, p_1 and p_2 approach either zero or one as $\theta \to \infty$. In the limit choices can be predicted deterministically and the d-measure is 1.

This example illustrates this phenomenon. It will calculate $d(\theta)$ and show how it increases from zero to one as θ goes from zero to ∞. It will also demonstrate that for small values of θ, the approximation discussed in the preceding section holds.

Let us first calculate f_i and $\text{var}_X[P_i(\theta, X)]$. Letting $\phi(x)$ denote the standard normal p.d.f., we write

$$f_i = \int_{-\infty}^{\infty} \Phi(\theta x)\phi(x)dx = \int_{-\infty}^{\infty} \Phi(-\theta x)\phi(x)dx$$

$$= \int_{-\infty}^{\infty} [1 - \Phi(\theta x)]\,\phi(x)dx = 1 - \int_{-\infty}^{\infty} \Phi(\theta x)\phi(x)dx$$

$$= 1 - f_i \Rightarrow f_i = 0.5$$

Also,

$$\text{var}_X[P_i(\theta, X)] = \int_{-\infty}^{\infty} \Phi^2(\theta x)\phi(x)dx - f_i^2$$

Consequently,

$$d(\theta) = [4\int_{-\infty}^{\infty} \Phi^2(\theta x)\phi(x)dx - 1]^{(1/2)} \quad (27)$$

It is clear that if $\theta = 0$, $d = 0$ and that $\lim_{\theta \to \infty} d(\theta) = 1$. Figure 5 plots $d(\theta)$.

The approximation given by Equation 26 can also be calculated:

$$(1/N)\,\hat{\mathcal{L}} = 2 \times 0.5 \log 0.5 = -0.693$$

$$(1/N)\,\hat{\mathcal{L}} = 2\int_{-\infty}^{\infty} \Phi(\theta x)\,[\log \Phi(\theta x)]\,\phi(x)dx$$

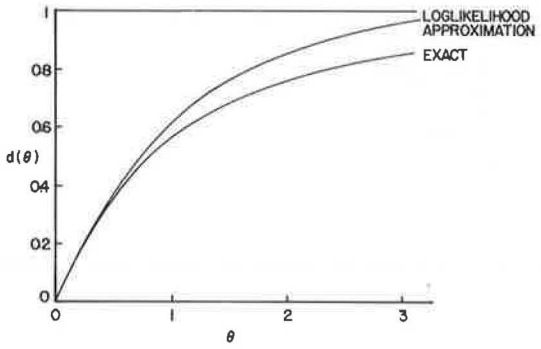
$$d(\theta) \approx [4\int_{-\infty}^{\infty} \Phi(\theta x)\,[\log \Phi(\theta x)]\,\phi(x)dx + 1.386]^{(1/2)}$$

Figure 5 also plots this equation, which, as expected, tracks well low-to-moderate correlation values.

CONCLUSION AND EXTENSIONS

The d-measure and the D-statistic have been shown to be the binary model equivalent of the correlation coefficient of regression in the sense that they can be related to predictive RMS error in the same way. It was shown that D was a consistent estimator for d and that for the typical low-correlation models en-

Figure 5. D-measure and its log-likelihood approximation.



countered in social sciences, d is in between the two other measures of goodness-of-fit that are commonly used: the likelihood ratio goodness-of-fit measure, g_ℓ, and the success table measure of goodness-of-fit, g_b.

If, as is usually the case, the goodness-of-fit measures are not high, it is possible to obtain an approximation for d that is based on the log-likelihood function, Equation 26. For models with low correlation, the measure ℓ itself may be regarded as a goodness-of-fit measure but care must be exercised because ℓ may exceed one.

Equation 26 is also useful because under the usual regularity conditions, $\hat{\ell}$ (the estimator for ℓ from the sample log-likelihood) is such that $(N\hat{\ell}^2)$ is χ^2 distributed with as many degrees of freedom as parameters. Of course, D has approximately the same distribution if small.

For models with more than two alternatives (i = 0,1,...,I-1), a prediction goodness-of-fit measure can be developed in the same way. In this instance, however, the goodness-of-fit measure is not the same for all the alternatives because $\text{var}_X[P_i(X)]$ is not the same for all. Consideration shows that the measure for the ith alternative, d_i, is

$$d_i = \sqrt{\{\text{var}_X[P_i(X)]\}/[f_i(1 - f_i)]},$$

and that a D-statistic can be derived in the same way. The interpretation in terms of prediction RMS error reduction (Figure 1) is unchanged.

ACKNOWLEDGMENT

The research reported in this paper was supported by NSF grants CEE-7812004 and CEE-81-11681 to the University of California, Berkeley.

REFERENCES

1. D. McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. In Frontiers of Econometrics (P. Zarembka, ed.), Academic Press, New York, 1970, pp. 105-142.
2. J.R. Hauser. Testing the Accuracy, Usefulness and Significance of Probabilistic Choice Models: An Information-Theoretic Approach. Operations Research, Vol. 26, 1978, pp. 406-421.
3. C.F. Daganzo. Multinomial Probit: The Theory and Its Application to Demand Forecasting. Academic Press, New York, 1979.

Appendix

CONSISTENCY OF THE D-STATISTIC

Let us assume that a random, attribute-based sampling process was used to gather a sample for estimation of a binary choice model. We denote by $X^{(n)}$ the attribute vector of the nth observation, by $i^{(n)}$, the choice of said observation and by N the number of observations. We assume that there is a set of parameters, θ, that are entered in the specification of the model:

$$0 < p_I = G(\theta., X) = 1 - p_o < 1 \tag{A1}$$

and that for a true (unknown), θ_o, $G(\theta_o, X)$ coincides with $P_1(X)$.

Let $F_X(x)$ denote the sampling distribution of X and write $\sigma^2(\theta)$ for the variance of $G(\theta, X)$; $\sigma^2(\theta_o)$ represents the variance of the choice probabilities in the numerator of Equation 16b. We assume that some regularity conditions to make the MLE of θ_o, $\hat{\theta}$, consistent hold, and in particular that

(A) $G(\theta, X)$ is a continuously differentiable function of θ_o in a δ_o neighborhood of it;
(B) The sample space of X is bounded, and
(C) $0 < G(\theta, X) < 1$.

These three properties will be used in the consistency proof of D.

Let $S_N^2[...]$ denote the sample variance function of the N elements in brackets. Then, Property (C) ensures that $\sigma^2(\theta_o)$ is finite and the consistency of the sample variance statistic enables us to write:

$$\left\{[S_N^2[G(\Theta_o, X^{(1)}), \ldots, G(\Theta_o, X^{(N)})] - \sigma^2(\Theta_o)]\right\} \xrightarrow{p} 0 \tag{A2}$$

where $\xrightarrow{p}$ denotes the limit in probability of the sequence in braces as $N \to \infty$. Note that the $X^{(n)}$ are i.i.d. drawings from $F_X(x)$.

Because $\hat{\theta}$ is consistent and because of Equation A2, we can write for arbitrarily small values of ϵ', ϵ'', δ', and δ'':

$$\Pr\left\{|S_n^2[G(\theta_o, X^{(1)}), \ldots, G(\theta_o, X^{(n)})] - \sigma^2(\theta_o)| \leq \epsilon'\right\} \geq 1 - \delta'; \forall n'' \geq N'(\epsilon', \delta') \tag{A3}$$

and

$$\Pr\left[\| \hat{\Theta}(X^{(1)} \ldots X^{(n)}; i^{(1)} \ldots i^{(n)}) - \theta_o \| \leq \epsilon''\right] \geq 1 - \delta''; \forall n'' \geq N''(\epsilon'', \delta'') \tag{A4}$$

In these two expressions, the random variables $(X^{(1)} \ldots X^{(n)})$ are the same, $N'(\epsilon', \delta')$ and $N''(\epsilon'', \delta'')$ represent finite positive numbers that depend on (ϵ', δ') and (ϵ'', δ''), $\| \cdot \|$ represents the Euclidean norm, and $\hat{\theta}(...)$ is the function that relates the data, $(X^{(1)} \ldots X^{(n)}; i^{(1)} \ldots i^{(n)})$, to the MLE of θ_o.

Before proceeding, we need the following preliminary result:

Lemma 1

Under the regularity conditions previously stated, we can write for sufficiently small but positive values of (ϵ'', δ''):

$$\Pr\Big(|S_n^2\{G[\hat{\Theta}(X^{(1)} \ldots X^{(n)}; i^{(1)} \ldots i^{(n)}), X^{(1)}], \ldots, G[\hat{\Theta}(X^{(1)} \ldots X^{(n)}; i^{(1)} \ldots i^{(n)}), X^{(n)}]\} - S_n^2[G(\theta_o, X^{(1)}), \ldots, G(\theta_o, X^{(n)})]| \leq k\epsilon''\Big) \geq 1 - \delta''; \forall n'' \geq N''(\epsilon'', \delta'') \tag{A5}$$

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