USING LOGISTIC MODEL CALIBRATION TO ASSESS THE QUALITY OF PROBABILITY PREDICTIONS

Frank E. Harrell, Jr.
Kerry L. Lee

Division of Biometry, Duke University Medical Center
Box 3363, Durham, North Carolina 27710, USA

SUMMARY

We used a logistic calibration model (Cox, 1958a) to partition a logarithmic scoring rule (used to assess the quality of probability predictions) into indexes of discrimination and three indexes of unreliability. An index of overall quality that is not penalized for a prevalence correction is also proposed. Various tests for discrimination and unreliability arise immediately from these indexes. Power properties of a test for unreliability are studied.

I. INTRODUCTION

The assessment of predictive accuracy is of central importance in validating and comparing either subjective or model-based predictions of event outcomes. When one is predicting a continuous outcome measurement, using ordinary multiple linear regression for example, assessment of the quality of predictions can be carried out in a straightforward way using scatter diagrams, correlations (predicted with observed), and error estimates (predicted-observed outcomes). When, however, the prediction is the probability that a particular event will occur, assessment of predictive quality is much more difficult due to the binary nature of the outcome.

Two commonly used concepts of quality of predictions are the discrimination of a predictor (often called its refinement), i.e., the ability of the predictor to separate or rank-order observations with different outcomes, and its reliability (sometimes called validity or degree of being

calibrated) - the "correctness" of predictions on an absolute scale. If, for example, a predictor assigned a 20% probability of disease for each of a homogenous group of 100 patients and 20 patients were later diagnosed to have the disease, the predictions would be reliable. Even though reliability is a simpler concept, discrimination is easier to uniquely quantify. For example, we may calculate the concordance probability from the Wilcoxon-Mann-Whitney statistic - the proportion of pairs of subjects, one with and one without the outcome being predicted, such that the subject with the outcome had the higher predicted probability (see Harrell, et al, 1982). We refer to this concordance measure as the c-index. This measure is equivalent to the area under a "receiver operating characteristic" curve (Hanley and McNeil, 1982) and is a linear translation of Somers' rank correlation between predicted probabilities and the binary event indicator, which in the absence of ties on predicted values is also equivalent to the Goodman-Kruskal rank correlation coefficient (Goodman and Kruskal, 1975).

Reliability is traditionally assessed by estimating the prevalence of the event in question for each level of predicted probability. This method works well when either only a few unique predictions are made or the sample is extremely large. When the predictions vary continuously from 0 to 1, some grouping of the probabilities is usually necessary. This may be accomplished by rounding predicted probabilities into intervals or by constructing quantile groups. Once grouping is done, a variety of goodness of fit tests are available for detecting unreliability (Lemeshow and Hosmer, 1982).

The method of grouping predicted probabilities to assess reliability has several drawbacks when applied to predictions that range continuously from 0 to 1. The most serious of these is the fact that one's assessment of reliability can change significantly depending on how the groups are formed. In addition, when one wants to test whether the predictions are "significantly unreliable", i.e., whether the observed prevalence differs significantly from the predicted values, an ordinary chi-square goodness-of-fit test lacks power. If separate samples were used for deriving and assessing predictions, and the probabilities were divided into ten groups, the χ^2 statistic has 9 degrees of freedom (d.f.), which has a critical value of 16.9 at the 5% level.

If unreliability could be described with only 2 d.f., the critical value is reduced to 6.

Various indexes have been proposed for assessing the accuracy of probability forecasts [see DeGroot and Feinberg 1982, Hilden, Habbema, and Bjerregaard, 1978, and Spiegelhalter, 1986 for detailed discussions and bibliographies]. One commonly used accuracy index is Brier's (1950) quadratic scoring rule. This index is a "proper scoring rule", meaning that a predictor optimizes the Brier index by predicting the true probability of the event in question. Brier's index has been decomposed into reliability and discrimination components (Hilden et al., 1978, Spiegelhalter, 1986, Yates, 1982, Blattenberger and Lad, 1985, DeGroot and Feinberg, 1982) and a test for reliability based on one decomposition has been proposed (Hilden et al., 1978).

Logarithmic scoring rules (Good, 1952) are also popular, and Cox (1958b) has presented one related test for reliability based on linear log odds alternatives to perfect reliability. In contrast to Brier's index, less work has been done to decompose logarithmic scoring rules into indexes of reliability and discrimination. The method presented in Section 2 allows decomposition of an overall quality measure into discrimination and various unreliability components, each chance-corrected, and admits straightforward likelihood ratio and score tests for significant discrimination and unreliability.

Throughout the discussion we assume that predictions and outcomes are stochastically independent. This is true if, for example, a regression model (e.g., a logistic model) was derived from a "training" sample and predictive accuracy was assessed on an independent "test" sample, which is often the only way to obtain an unbiased validation of the entire modeling process.

2. DESCRIBING ACCURACY USING CALIBRATION

Suppose that one could estimate the "calibration curve" - the relationship between the predicted probability and the true probability of the event. Given an estimate of this relationship, one way to quantify the unreliability

of predictions is to measure what has to be done to make the calibration curve superimposed on the ideal curve (a 45 degree line). Discrimination, on the other hand, is related to whether or not the predictions are in any way related to the outcomes, i.e., whether or not the calibration curve is horizontal.

Thus the problem of quantifying unreliability and discrimination can be solved by estimating the relationship between predicted and observed values. Since observed values are binary, the relationship is stated in terms of the probability that the outcome occurs. The method of maximum likelihood can be used to estimate this relationship even when no two predictions are the same. The method only assumes that predictions are related to outcomes through a smooth curve that interpolates between different predictions.

For a simple predictor variable X, the logistic regression model (Cox, 1958a, 1966, Walker and Duncan, 1967) relates X to the probability of an event. Let the event or outcome variable be denoted by Y, where Y=1 when the event occurs and Y=0 otherwise. The model is as follows:

$$\text{Prob}(Y=1 \mid X) = \frac{1}{1 + \exp[-(a+bX)]}, \qquad (2.1)$$

where exp(x) is e^x, the natural antilogarithm. Cox (1958b, 1970) proposed using the linear logistic model to relate "subjective probabilities" to "objective probabilities". Let the predicted probabilities of $Y_1, Y_2, \ldots, Y_n = 1$ be denoted by $P_1, P_2, \ldots, P_n$ for n subjects, or cases. Let the true (calibrated) probabilities be denoted by $P_1', P_2', \ldots, P_n'$. We can estimate P_i' given P_i by estimating the relationship between P_i and Y_i. P_i is first transformed from a 0-1 scale to an unlimited scale to better fit the model. The logistic calibration model is:

$$\text{Prob}(Y_i=1 \mid P_i) = \frac{1}{1 + \exp[-(a+bL_i)]}, \qquad (2.2)$$

where $L_i = \text{logit}(P_i) = \log(P_i/(1-P_i))$. The model can be restated as

$$P_i' = \frac{1}{1 + \exp(-a)[P_i/(1-P_i)]^{-b}}. \qquad (2.3)$$

Figure 1 shows the shape of calibration curves for various values of a and b. The ideal relationship (no calibration required) is found on the curve marked a=0, b=1.

--- Figure 1 About Here ---

Note that when a=0 and b=1, $P_i' = P_i$, and no calibration is required. When no slope calibration is required (b=1),

$$P_i' = \frac{P_i}{P_i + (1-P_i)\exp(-a)}. \qquad (2.4)$$

In this case, where only a prevalence adjustment is made, exp(a) is the odds ratio of the corrected to the uncorrected overall prevalence, and the calibration model is identical to the simplest form of Bayes' rule. It should be recognized that a simpler calibration model such as $P_i' = a+bP_i$ cannot be used because this would allow P_i' to be less than 0 or greater than 1.

In many cases where a model has not been developed carefully on a dataset, predictions from the model will be found too extreme when they are validated in an independent sample because of overfitting the original dataset. For example, a probability of death of .1 may need to be calibrated to .25, and a prediction of .9 calibrated to .75. The corresponding logistic calibration for this example is obtained using a=0, b=.5 in equation (2.3). If predictions need to be shrunk symmetrically toward a probability of .5 and a predicted probability of P is calibrated to a value of P', the calibrating equation is derived from (2.3) using a=0, b=logit(P')/logit(P).

The parameters a and b can be estimated by maximizing the posterior likelihood of the observed data (P_i, Y_i, i=1,2,...,n), or equivalently by minimizing -2 times the log-likelihood function,

$$L = -2 \sum_{i=1}^{n} [Y_i \log(P_i') + (1-Y_i) \log(1-P_i')]. \qquad (2.5)$$

$$= -2 \sum_{i=1}^{n} [Y_i (a+bL_i) - \log(1+\exp(a+bL_i))]$$

L can be thought of as measuring information or quality of the predictions in relation to the outcomes, given a and b.

3. DERIVATION OF ACCURACY INDEXES

The following notation will be used:

L(a,b) = minimum L for all a,b
L(a,1) = minimum L for all a subject to b=1
L(a,0) = minimum L for all a subject to b=0
= $-2\Sigma[Y_i \log P+(1-Y_i)\log(1-P)]$
L(0,1) = value of L at a=0, b=1
= $-2\Sigma[Y_i \log P_i+(1-Y_i)\log(1-P_i)]$,

where $P=\Sigma Y_i/n$. We compute the unreliability U of the predictions from the difference in quality of the uncalibrated predictions and the quality of slope- and intercept-calibrated predictions:

$$U = [L(0,1) - L(a,b) - 2]/n \; . \qquad (3.1)$$

Since L(0,1) - L(a,b) is a likelihood ratio statistic for testing H_0: a=0,b=1 with an asymptotic χ^2 distribution having expected value 2 if H_0 is true, U has expected value 0 if the predictions are reliable. Division by the sample

size makes the range of U independent of n. U can be decomposed into $U = U_p + U_s$, where U_p is the unreliability due to the need for an overall prevalence correction (correction of intercept on logit scale) and U_s is unreliability due to the need for a slope correction given any needed prevalence correction:

$$U_p = [L(0,1) - L(a,1) - 1]/n \qquad (3.2)$$

$$U_s = [L(a,1) - L(a,b) - 1]/n \; . \qquad (3.3)$$

U_p is the difference in quality of the best uncalibrated predictor and an intercept-calibrated predictor. U_s is the difference in quality of the best intercept-calibrated predictor and the best slope- and intercept-calibrated predictor. The -1 term causes each index to have expected value 0 if the corresponding type of unreliability is truly absent. Large values of the unreliability indexes mean that the predictions are unreliable. Negative values indicate better reliability than one would expect by chance.

Likelihood ratio statistics are immediately available for testing each type of unreliability:

Null Hypothesis	Asymptotic χ^2	d.f.	
Significant total unreliability H_0:a=0,b=1	L(0,1)-L(a,b)	2	(3.4)
Significant unreliability due to overall prevalence error H_0:a=0 b=1	L(0,1)-L(a,1)	1	(3.5)
Significant unreliability due to slope error given prevalance correction H_0:b=1	L(a,1)-L(a,b)	1	(3.6)

Simple score tests are also available for testing the first two hypotheses above, avoiding the need for iterative calculations (Rao, 1973). A 2 d.f. asymptotic χ^2 score statistic for H_0: a=0, b=1 is given by

$$[\Sigma(Y_i-P_i) \quad \Sigma L_i(Y_i-P_i)] \begin{matrix} \Sigma P_i(1-P_i) & \Sigma L_i P_i(1-P_i) \\ \Sigma L_i P_i(1-P_i) & \Sigma L_i^2 P_i(1-P_i) \end{matrix}^{-1} \begin{matrix} \Sigma(Y_i-P_i) \\ \Sigma L_i(Y_i-P_i) \end{matrix} \quad (3.7)$$

A 1 d.f. test statistic for H_0: a=0 b=1 is

$$[\Sigma(Y_i-P_i)]^2/\Sigma P_i(1-P_i) \quad (3.8)$$

These score tests turn out to be identical to those proposed by Cox (1958b). Cox also presented a test for whether predicted probabilities are overly dispersed even though they are correct on the average (H_0: b=1 a=0).

The index of discrimination is derived by computing the difference in quality of the best constant predictor (one that on the average correctly predicts the overall prevalence of the event) and the best calibrated predictor:

$$D = [L(a,0) - L(a,b) - 1]/n. \quad (3.9)$$

D has expected value 0 if there is no discrimination (b=0). The likelihood ratio statistic for testing whether the predictions have any discriminatory ability (H_0:b=0) is L(a,0) - L(a,b), having asymptotically a chi-square distribution with 1 d.f. under H_0.

An overall summary index for the quality of predictions is derived by computing the difference in quality between the best constant predictor and the quality of the predictions as they stand (with no calibration):

$$Q = [L(a,0) - L(0,1) + 1]/n. \quad (3.10)$$

It can readily be seen that Q = discrimination - total unreliability = D - U. The summary index Q is a simple translation of the logarithmic scoring rule

Simple score tests are also available for testing the first two hypotheses above, avoiding the need for iterative calculations (Rao, 1973). A 2 d.f. asymptotic χ^2 score statistic for H_0: a=0, b=1 is given by

$$[\Sigma(Y_i-P_i) \quad \Sigma L_i(Y_i-P_i)] \begin{matrix} \Sigma P_i(1-P_i) & \Sigma L_i P_i(1-P_i) \\ \Sigma L_i P_i(1-P_i) & \Sigma L_i^2 P_i(1-P_i) \end{matrix}^{-1} \begin{matrix} \Sigma(Y_i-P_i) \\ \Sigma L_i(Y_i-P_i) \end{matrix} \quad (3.7)$$

A 1 d.f. test statistic for H_0: a=0 b=1 is

$$[\Sigma(Y_i-P_i)]^2/\Sigma P_i(1-P_i) \quad (3.8)$$

These score tests turn out to be identical to those proposed by Cox (1958b). Cox also presented a test for whether predicted probabilities are overly dispersed even though they are correct on the average (H_0: b=1 a=0).

The index of discrimination is derived by computing the difference in quality of the best constant predictor (one that on the average correctly predicts the overall prevalence of the event) and the best calibrated predictor:

$$D = [L(a,0) - L(a,b) - 1]/n. \quad (3.9)$$

D has expected value 0 if there is no discrimination (b=0). The likelihood ratio statistic for testing whether the predictions have any discriminatory ability (H_0:b=0) is L(a,0) - L(a,b), having asymptotically a chi-square distribution with 1 d.f. under H_0.

An overall summary index for the quality of predictions is derived by computing the difference in quality between the best constant predictor and the quality of the predictions as they stand (with no calibration):

$$Q = [L(a,0) - L(0,1) + 1]/n. \quad (3.10)$$

It can readily be seen that Q = discrimination - total unreliability = D - U. The summary index Q is a simple translation of the logarithmic scoring rule

(see Good (1952), Cox (1970, Eq. 4.34), and Shapiro (1977)). The reference point for Q is the best constant predictor, whereas other authors used as reference a predictor having constant probability 0.5.

The value of Q is invariant with respect to the form of the calibrating model. This result follows from 1) when a=0 and b=1, $P_i = P_i'$, making the log-likelihood function (here L(0,1)) dependent only on the observed data, and 2) when b=0, the calibrated probabilities do not make use of the predicted probabilities so that $P_i' = P$, the overall proportion of events that occurred ($\Sigma Y_i/n$). Hence (3.10) reduces to

$$Q = (2/n)\ \Sigma[Y_i \log (P_i/P) + (1-Y_i) \log (1-P_i)/(1-P)] + 1/n. \qquad (3.11)$$

An index of quality can also be constructed which does not penalize predictions for being wrong by a constant prevalence correction. This index is derived from the difference in quality of the best intercept-corrected predictions from the quality of the best constant predictor:

$$Q_S = [L(a,0) - L(a,1)]/n = D - U_S. \qquad (3.12)$$

Since score statistics can be used to approximate likelihood ratio statistics, simpler unreliability indexes can be constructed by substituting (3.8) for L(0,1)-L(a,1) in (3.2) and (3.7) for L(0,1)-L(a,b) in (3.1). This approach has two disadvantages, though. First, score statistics are not additive as are partitions of log-likelihood. Second, score statistics may not adequately quantify information content for situations far from the null hypothesis.

4. POWER OF TEST FOR UNRELIABILITY

The likelihood ratio test for total unreliability given in (3.4) is difficult to study because of the iterative calculations required. It has been shown in a similar situation that score tests have equivalent power functions as likelihood ratio tests (Lee et al, 1983). Therefore we study the power properties of the score test given by (3.7).

In general, $E(Y_i)=P_i'$ and the score vector $[\Sigma(Y_i-P_i),\ \Sigma L_i(Y_i-P_i)]$ is asymptotically normal with mean vector and covariance matrix given respectively by

$$\mu = \begin{matrix} \Sigma(P_i'-P_i) \\ \Sigma L_i(P_i'-P_i) \end{matrix} \qquad V = \begin{matrix} \Sigma P_i'(1-P_i') & \Sigma L_i P_i'(1-P_i') \\ \Sigma L_i P_i'(1-P_i') & \Sigma L_i^2 P_i'(1-P_i') \end{matrix} \tag{4.1}$$

It follows that the score statistic for testing H_0:a=0, b=1 has mean m and variance v given by

$$\begin{aligned} m &= \mathrm{tr}AV + \mu' A\mu \\ v &= 2\mathrm{tr}[(AV)^2] + 4\mu' AVA\mu \end{aligned} \tag{4.2}$$

where A is the matrix inverse in (3.7). The distribution of (3.7) can be approximated by a scaler multiple of a non-central χ^2 random variable $\beta\chi^2_2(\lambda)$ with 2 d.f. and noncentrality λ by equating the first two moments of such a distribution to (4.2) (Johnson and Kotz, 1970), yielding

$$\begin{aligned} \beta &= [m - (m^2 - v)^{1/2}]/2 \\ \lambda &= m/\beta - 2. \end{aligned} \tag{4.3}$$

If $\chi^2_{2;1-\alpha}$ represents the $1-\alpha$ quantile of a central χ^2 distribution having 2 d.f., the power of an α-level test for unreliability can be approximated by

$$\mathrm{Prob}(\beta\chi^2_2(\lambda) > \chi^2_{2;1-\alpha}) = \mathrm{Prob}\ (\chi^2_2(\lambda) > \chi^2_{2;1-\alpha}/\beta). \tag{4.4}$$

To test the adequacy of this approximation as well as the adequacy of the central χ^2 null distribution for (3.7), 2000 samples of varying sizes were simulated for each of a variety of setups in which only two distinct

predictions were made. Here k observations were assigned a predicted probability of p_1 and k were assigned a probability of p_2. Corresponding actual population probabilities were p_1' and p_2'. Power was estimated by computing the fraction of score statistics exceeding $\chi^2_{2;.95}=5.99$. Results for $\alpha=.05$ are found in Table 1. It can be seen that type I error is well controlled and that (4.4) appears to be a satisfactory power approximation.

--- Table 1 About Here ---

The power approximation in (4.4) can be used to estimate the sample size necessary to achieve a power of f for an α-level test of total unreliability. Suppose that k predictions are made at each of g probability levels, $p_1, p_2, \ldots, p_g$ and that the true probabilities are $p_1', p_2', \ldots, p_g'$. The total sample size is thus kg. Define $\gamma_j=\text{logit } p_j$. Then the quantities in (4.2) are given by

$$m = \text{tr } A^*V^* + k\mu^{*\prime}A^*\mu^*$$
$$v = 2\,\text{tr}[(A^*V^*)^2] + 4k\,\mu^{*\prime}A^*V^*A^*\mu^*, \tag{4.5}$$

where

$$A^* = \begin{bmatrix} \Sigma p_j(1-p_j) & \Sigma 1_j p_j(1-p_j) \\ \Sigma 1_j p_j(1-p_j) & \Sigma 1_j^2 p_j(1-p_j) \end{bmatrix}^{-1}$$

$$V^* = \begin{bmatrix} \Sigma p_j'(1-p_j') & \Sigma 1_j p_j'(1-p_j') \\ \Sigma 1_j p_j'(1-p_j') & \Sigma 1_j^2 p_j'(1-p_j') \end{bmatrix} \tag{4.6}$$

$$\mu^* = \begin{bmatrix} \Sigma(p_j'-p_j) \\ \Sigma 1_j(p_j'-p_j) \end{bmatrix},$$

and all summations are over $j=1, 2, \ldots, g$.

The following iterative algorithm converges quickly to a solution for k:

Initialize $k_{last} = 0$; $\beta=1$

loop X: Compute λ= non-centrality parameter of χ^2_2 distribution such that Prob $(\chi^2_2(\lambda) \leq \chi^2_{2;1-\alpha}/\beta) = 1-f$

Set $m = \beta(\lambda+2)$

Set $k = [m - trA^*V^*]/\mu^{*'}A^*\mu^*$

Set $\beta = [m - (m^2-v)^{1/2}]/2$

where $v = 2tr[(A^*V^*)^2] + 4k\mu^{*'}A^*V^*A^*\mu^*$

If $k-k_{last} < 1$ stop

Set $k_{last} = k$; go to X loop

5. EXAMPLES OF VALUES OF THE INDEXES

To help in interpreting the values of the indexes, consider a series of simple examples in which k subjects receive one prediction, p_1, and another k subjects receive a predicted probability of p_2. The first k subjects have an observed prevalence of the event of O_1 and the second k have a prevalence of O_2. The resulting calibration parameter estimates (a and b), accuracy indexes, and chi-square statistics are in Table 2 for k=100. For comparison, the c index is also given, along with a version of Brier's score defined by B = 1 - average of $(P_i-Y_i)^2$.

--- Table 2 About Here ---

Lines 1 and 2 in the table demonstrate the values of the indexes when there is perfect reliability and low to moderate discrimination, respectively. Similarly, lines 3-5 correspond to backwards predictions (e.g., predict .25 and .75, observe .75 and .25) with increasing discrimination. Total unreliability is statistically significant for lines 3-6, 8 and 9. Lines 6-9 are more typical examples of unreliability. The measure of overall quality, Q, is negative (lines 3-6) when the discrimination is not good enough to overcome serious unreliability. The index of discrimination, D, ranked the discrimination of predictions in the same order as the absolute value of c-.5. The rankings of Q and B are very similar but both differ from those of the absolute value of c-.5. They have similar rankings as c.

It appears that predictions for which U does not exceed about 0.05 are reliable for the most part. Statistical significance of U can also be used to quantify unreliability, although the power of this assessment depends on the sample size (significant unreliability is present at the α=.05 level if U>3.99/n; for U_p and U_s the critical levels are 2.84/n). It can be shown that for this situation (k predictions at each of two probabilities), the unreliability index is given by

$$U = O_1 \log[p_1/(1-p_1)]/[O_1/(1-O_1)] + O_2 \log[p_2/(1-p_2)]/[O_2/(1-O_2)] \qquad (5.1)$$

$$+ \log [(1-p_1)(1-p_2)]/[(1-O_1)(1-O_2)] \quad -2/n.$$

The analyst can use (5.1) to estimate acceptable levels of U for fixed p_1, p_2 by varying O_1 and O_2 and judging U by whether O_1 and O_2 are meaningfully different from p_1 and p_2. A plot of U with respect to O_1 and O_2 is shown in Figure 2 when k=100 (n=200) for the four combinations p_1=.25, p_2=.75; p_1=.85, p_2=.95; p_1=.6, p_2=.4; and p_1=.1, p_2=.2.

To show examples of the values of the new indexes as well as the resulting estimates of the calibration or reliability curves when the predictions are continuous, the predictive accuracy of two logistic regression models was considered. For both models, the outcome variable was complete response to treatment of non-Hodgkin's lymphoma, and the predictions were developed (Harrell et al., 1985) using a training sample of 110 patients (50 with

complete response, 60 without). These predictions were evaluated on a separate test sample of 116 patients. The first model was developed using a standard stepwise variable selection method with 25 candidate variables, far too many for only 50 cases of complete response. The second model used the incomplete principal component method, which effectively reduced the 25 variables to only 1. The accuracy indexes are found in Table 3. Corresponding p-values for significant unreliability or discrimination are in parenthesis. Reliability plots using the estimated a and b may be found in Figure 1. The results indicate significant unreliability (both types) and little discrimination ability for model 1, resulting in unacceptable predictions (Q=-.19). The extreme predictions from model 1 cannot be trusted, which is frequently the case when two many predictor variables are used with small sample sizes. Model 2 has moderate need for a prevalence correction (U_p*.05) but not for a slope correction, and has better discrimination than model 1, resulting in far better overall quality (Q=.03 vs. -.19). This improvement in predictive accuracy is due to the data reduction resulting from fitting principal components.

--- Table 3 About Here ---

6. COMPUTER SOFTWARE

A SAS (1985) macro is available from the authors for calculating all of the indexes mentioned in this paper as well as for drawing the reliability plot. Another SAS program is available for power and sample size calculations based on (4.4) and (4.7).

7. CONCLUSIONS

We sought a method of assessing predictive quality having the following properties: (1) no grouping of predictions is required, (2) an overall measure of the quality of predictions can be formally decomposed into a simple sum of indexes of unreliability and discrimination, (3) the index of unreliability can be further decomposed into an index of unreliability due to the need for an overall prevalence (constant) correction and unreliability due to a more complicated correction, (4) the method yielded as a byproduct an index of

overall predictive quality that was not penalized for a prevalence correction and (5) the method automatically yields formal statistical tests (with reasonable power) for significant unreliability (and its two components) and for significant discriminatory ability. The logistic regression model, when used to calibrate predicted probabilities to observed outcomes was useful in meeting these goals. The power approximation given in (4.4) is adequate for estimating the sample size needed to conduct studies such as those designed to test diagnostic accuracy of physicians or probability models.

ACKNOWLEDGEMENTS

This research was supported by the National Center for Health Services Research, and by the National Library of Medicine and National Heart, Lung, and Blood Institute of the National Institutes of Health. We thank Ms. Cristy Vollmar for the careful typing of the manuscript, Barbara Pollock for providing expert technical assistance, and Robert Rosati and David Pryor for motivating our work.

REFERENCES

1. Blattenberger G, Lad F (1985). Separating the Brier score into calibration and refinement components: a graphical example. Am Statistician 39:26-32.

2. Brier GW (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review 75:1-3.

3. Cox DR (1958a): The regression analysis of binary sequences (with discussion). J Roy Statist Soc B 20:215-242.

4. Cox DR (1958b): Two further refinements of a model for binary regression. Biometrika 45:562-565.

5. Cox DR (1966): Some procedures connected with the logistic qualitative response curve. In Research Papers in Statistics: Essays in Honor of J. Neyman's 70th Birthday, pp. 55-71, Ed. F.N. David. London:Wiley.

6. Cox DR (1970): The Analysis of Binary Data. London:Methuen, pp. 52-54.

7. DeGroot MH, Feinberg SE (1982): Assessing probability assessors: calibration and refinement. In Statistical Decision Theory and Related Topics III, Vol 1. Academic Press.

8. Good IJ (1952): Rational decisions. J Roy Statist Soc B 14:107-114.

9. Goodman LA, Kruskal WH (1979). Measures of Association for Cross - Classifications. New York: Springer-Verlag.

10. Hanley JA, McNeil BJ (1982): The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 50:23-36.

11. Harrell FE, Califf RM, Pryor DB, et al (1982): Evaluating the yield of medical tests. J Am Med Assoc 247:2543-6.

12. Harrell FE, Lee KL, Matchar DB, Reichert TA (1985): Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treatment Reports 69:1071-1077.

13. Hilden J, Habbema JDF, Bjerregaard B (1978): The measurement of performance in probabilistic diagnosis. III. - methods based on continuous functions of the diagnostic probabilities. Methods of Information in Medicine 17:238-246.

14. Johnson NL, Kotz S (1970): Distributions in Statistics: Continuous Univariate Distributions-2, pp. 165-166. New York: Wiley.

15. Lee KL, Harrell FE, Tolley HD, Rosati RA (1983): A comparison of test statistics for assessing the effects of concomitant variables in survival analysis. Biometrics 39:341-350.

16. Lemeshow S, Hosmer DW (1982): A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiology 115: 92-106.

17. Rao CR (1973): Linear Statistical Inference and Its Applications, Second Edition, pp. 418-419. New York: Wiley.

18. SAS Institute (1985): SAS User's Guide: Basics, Version 5 Edition. Cary, NC:SAS Institute, Inc., pp.643-727.

19. Shapiro AR (1977): The evaluation of clinical predictions: a method and initial application. New England J Med 296:1509-1514.

20. Speigelhalter DJ (1986): Probabilistic prediction in patient management and clinical trials. Statistics in Medicine 5:421-433.

21. Walker SH, Duncan DB (1967): Estimation of the probability of an event as a function of several independent variables. Biometrika 54:167-179.

22. Yates JF (1982): External correspondence: decompositon of the mean probability score. Organisational Behavior and Human Performance 30:132-156.

Table 1

Simulated and Approximated Power of
Score Test for Unreliability
α=.05

k	Predicted Probabilities		True Probabilities		Simulated Power	Power by (4.4)
	P_1	P_2	P_1	P_2		
10	.25	.75	.25	.75	.050	.050
			.10	.75	.063	.083
20	.25	.75	.25	.75	.050	.050
			.10	.75	.197	.191
30	.25	.75	.25	.75	.034	.050
			.10	.75	.293	.329
40	.25	.75	.25	.75	.055	.050
			.10	.75	.518	.475
100	.25	.75	.25	.75	.058	.050
			.15	.75	.546	.535
			.15	.85	.891	.873
	.02	.95	.02	.95	.042	.050
			.10	.95	.948	.950
	.10	.95	.10	.95	.051	.050
			.02	.95	.786	.783

Table 2

Examples of Values of the Indexes

	P_1	P_2	Q_1	Q_2	a	b	U_p	X^2	U_s	X^2	U	X^2	D	X^2	Q	c	B
1	.40	.60	.40	.60	0	1	-.005	0	-.005	0	-.01	0	.04	8	.04	.60	.76
2	.25	.75	.25	.75	0	1	-.005	0	-.005	0	-.01	0	.26	52	.27	.75	.81
3	.40	.60	.60	.40	0	-1	-.005	0	.160	32	.15	32	.04	8	-.12	.40	.72
4	.25	.75	.75	.25	0	-1	-.005	0	1.100	220	1.10	220	.26	52	-.83	.25	.56
5	.10	.90	.90	.10	0	-1	-.005	0	3.500	703	3.50	703	.73	147	-2.80	.10	.27
6	.40	.70	.60	.90	.99	1.43	.18	37	.005	2	.19	39	.12	25	-.07	.70	.80
7	.20	.70	.25	.75	.27	.98	.01	3	-.005	0	.004	3	.26	52	.25	.75	.81
8	.25	.70	.25	.90	.76	1.69	.04	10	.06	13	.11	23	.47	95	.36	.83	.84
9	.25	.55	.25	.90	1.69	2.54	.13	28	.15	31	.28	59	.47	95	.19	.83	.80

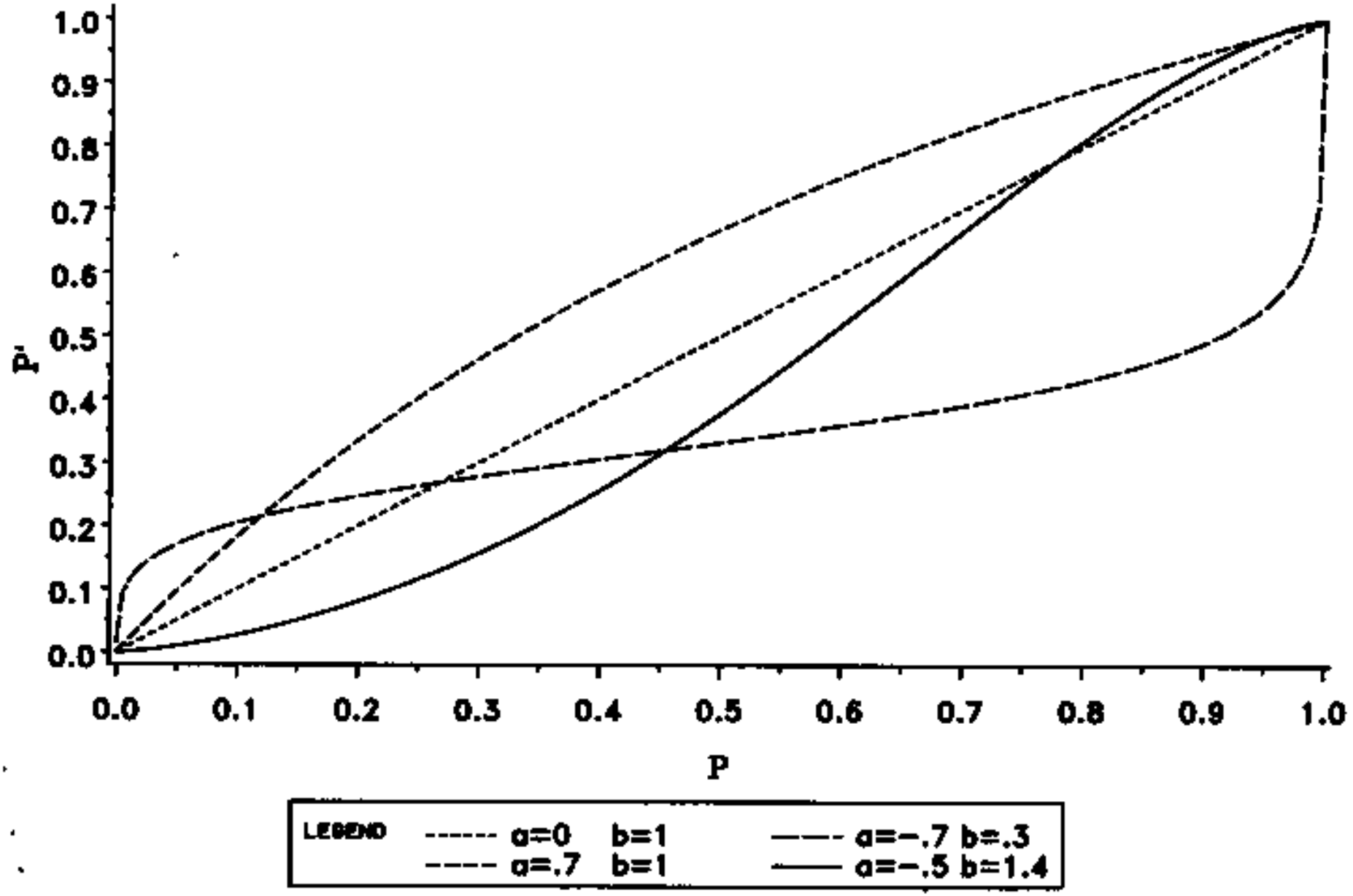
Table 3

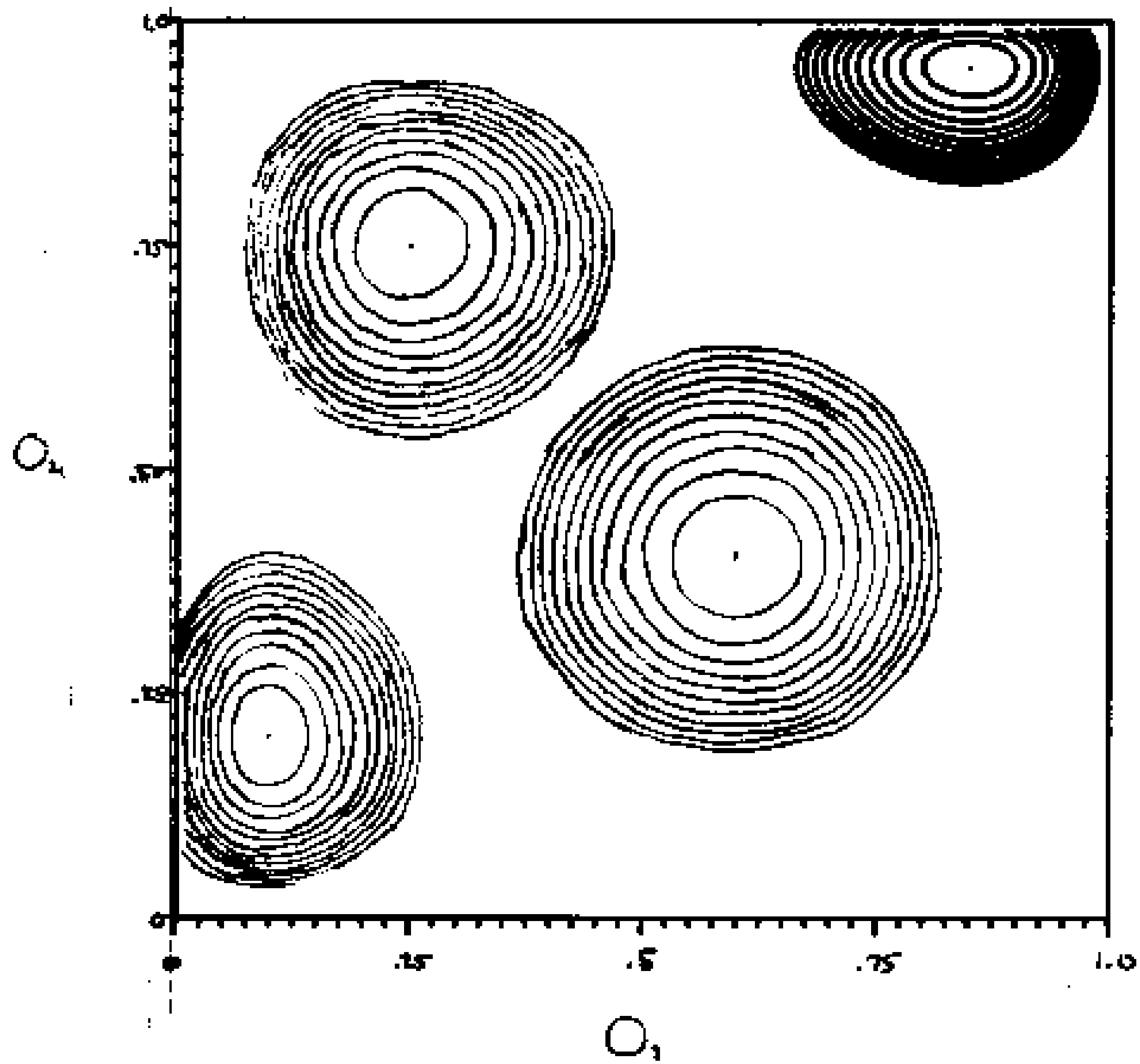
Comparing Predictive Accuracy of Two Logistic Regression Models

Quantity	Model 1	Model 2
a	-.7	-.5
b	.3	1.4
U_p	.14 (.0001)	.05 (.007)
U_s	.07 (.003)	-.003 (.4)
U	.21 (.0001)	.06 (.018)
D	.03 (.045)	.08 (.001)
Q	-.19	.03
Q_s	-.04	.08

Figure 1 Legend: Four logistic calibration (reliability) curves, including one for a reliable predictor (a=0, b=1).

Figure 2 Legend: Contour graphs of U (given by 5.1) as a function of observed proportions O_1 and O_2. The center of each set is (p_1, p_2), the true probabilities. The contours correspond to U = 0 (inner contour), .01, .02, ..., .10 (outer contour).


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
P'
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
P
LEGEND
a=0 b=1
a=.7 b=1
a=-.7 b=.3
a=-.5 b=1.4



1.0
.75
.50
.25
0
0
.25
.5
.75
1.0