STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:1771–1781

Summarizing the predictive power of a generalized linear model

Beiyao Zheng[1,*,†] and Alan Agresti[2]

[1] *Wake Forest University School of Medicine, Department of Public Health Sciences, Medical Center Boulevard, Winston-Salem, NC 27157-1051, U.S.A.*
[2] *Department of Statistics, University of Florida, Gainesville, FL 32611-8545, U.S.A.*

SUMMARY

This paper studies summary measures of the predictive power of a generalized linear model, paying special attention to a generalization of the multiple correlation coefficient from ordinary linear regression. The population value is the correlation between the response and its conditional expectation given the predictors, and the sample value is the correlation between the observed response and the model predicted value. We compare four estimators of the measure in terms of bias, mean squared error and behaviour in the presence of overparameterization. The sample estimator and a jack-knife estimator usually behave adequately, but a cross-validation estimator has a large negative bias with large mean squared error. One can use bootstrap methods to construct confidence intervals for the population value of the correlation measure and to estimate the degree to which a model selection procedure may provide an overly optimistic measure of the actual predictive power. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

The generalized linear model (GLM) provides a very broad and popular family for statistical analysis. For a particular choice of GLM, a measure of the model's predictive power can be useful for evaluating the practical importance of the predictors and for comparing competing GLMs, for example, models with different link functions or with different linear predictors. In ordinary regression for a normal response, the multiple correlation R and the coefficient of determination R^2 serve this purpose.

No summary measure of predictive power yet proposed for GLMs seems to have achieved the strong acceptance that R and R^2 have for normal regression models. Section 2 summarizes some measures that have been proposed. Most of these have serious limitations, such as lack of discriminatory power, restrictions in the response distribution, or poor interpretability. This article discusses a measure that is applicable to all types of GLMs: the correlation between the response

*Correspondence to: Beiyao Zheng, Wake Forest University School of Medicine, Department of Public Health Sciences, Medical Center Boulevard, Winston-Salem, NC 27157-1051, U.S.A.
†E-mail: bzheng@wfubmc.edu

Contract/grant sponsor: NIH; contract/grant number: GM 43824.

Received December 1998
Accepted August 1999

and its conditional expectation given the predictors. Its sample value is the sample correlation between the observed response and the model predicted value. Section 3 introduces this measure and discusses some of its properties. Section 4 uses simulation studies to compare four estimators of the measure in terms of bias, mean squared error, and behaviour in the presence of overparameterization. Section 5 illustrates the measure and its use as an aid in evaluating models, with data from a study of risk factors associated with low infant birth weight. Section 6 discusses the use of the bootstrap to obtain confidence intervals for the true measure and to estimate the degree to which a model selection procedure may provide an overly optimistic measure of the actual predictive power.

Let Y represent a response variable, and let X represent a vector of predictors. We treat both X and Y as random, and for statistical inference we assume that the sample (X_i, Y_i), $i = 1, 2, \ldots, n$, is a random sample. When only Y is random, as in many applications, the expected value of the sample estimator discussed here (like most other measures of association) depend on the selection of X values. For a GLM, let $E(Y|X = x)$ denote the conditional expectation of Y given $X = x$ and let $g(.)$ denote the link function, with $g[E(Y|X = x)] = \alpha + \beta x$. Finally, let $\hat{Y}$ denote the maximum likelihood (ML) estimate of $E(Y|X)$ based on this model.

2. SOME MEASURES OF PREDICTIVE POWER

Many summary measures of predictive power have been proposed [1] for GLMs. We now describe three of the main types of these measures and their shortcomings.

2.1. Measures Based On Ranking Information

These statistics measure the association between the ordered values of the response outcomes and the fitted values. The most popular measure of this type is the concordance index [2], denoted by c. Consider those pairs of observations that are untied on Y. The index c equals the proportion of such pairs for which the predictions $\hat{Y}$ and the outcomes Y are concordant, the observation with the larger Y also having the larger $\hat{Y}$. For a binary response, c is related to a widely used measure of diagnostic discrimination, the area under a receiver operating characteristic (ROC) curve [2, 3]. Various software packages, including S-plus [4], STATA and SAS (PROC LOGISTIC), report this measure.

Appealing features of c are its simple structure and its generality of potential application. Because c utilizes ranking information only, however, it cannot distinguish between different link functions, linear predictors, or distributions of the random components that yield the same orderings of the fitted values. For a binary response with a single linear predictor, for instance, the concordance index c assumes the same value for logit and complementary log-log link functions, even though the models are quite different; as long as the predicted values remain monotonic, c also remains the same when polynomial terms are added to the linear predictor. See Ash and Shwartz [5] for other criticisms.

2.2. Measures Based On a Variation Function

In ordinary linear regression with the normal model assuming constant variance, the coefficient of determination, R^2, describes the proportion of variance in Y explained by the model. It has been applied to other types of responses. For binary outcomes, for instance, let $\hat{\pi}_i$ denote the model-based

ML estimate of the probability of a positive response for subject i, and let $\bar{y}$ denote the sample proportion of positive responses. The sample measure [6] is defined as $R^2 = 1 - [\sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2]/[\sum_{i=1}^{n}(y_i - \bar{y})^2]$. Some have criticized the use of R^2 for non-normal GLMs because of restrictions in possible values to the lower end of the usual $[0, 1]$ scale and sensitivity to the prevalence of the outcome [7]. Others have argued, however, that sensitivity to prevalence is a strength [8], that a model with a low value of R^2 may still be helpful for prediction [5], and that R^2 captures information [5] not reflected by c.

For an arbitrary measure of variation $D(.)$, a natural extension [6, 9] of R^2 takes the form

$$\frac{\sum_{i=1}^{n} D(Y_i) - \sum_{i=1}^{n} D(Y_i|X_i)}{\sum_{i=1}^{n} D(Y_i)}$$

where $D(Y_i)$ denotes the variation for the ith observation and $D(Y_i|X_i)$ denotes the variation for the ith observation given the fixed value X_i of X. For a binary response, the proposed variation functions include squared error, prediction error, entropy and linear error [6]. For a categorical response, proposed variation functions include the Gini concentration measure and the entropy measure [9, 10]. Variation measures have also been proposed for other variants of the usual continuous response, such as a variety of measures for censored responses in survival analysis [11, 12].

Like c, an appealing aspect of measures based on variation functions is their simple structure, one that is well familiar to those who use R^2 for normal data. A disadvantage is that their numerical values can be difficult to interpret, depending on the choice of variation function. Although the measures may be useful in a comparative sense, many biostatisticians and most of the medical scientific community would find it difficult to envision what a 50 per cent reduction in entropy represents, for instance.

2.3. Measures Based on the Likelihood Function

Let ℓ denote the likelihood function and let $L = \log \ell$ denote the log-likelihood. Let $L_M = \log \ell_M$ denote the maximized log-likelihood under the model of interest. Let L_S denote the maximized log-likelihood under the saturated model, which has as many parameters as observations, and let L_0 denote the maximized log-likelihood under the null model, which has only an intercept term. Let $D_M = -2(L_M - L_S)$ and $D_0 = -2(L_0 - L_S)$ denote the deviances for the model of interest and the null model. A summary measure based on the likelihood function is [10, 14, 15]

$$D = \frac{L_M - L_0}{L_S - L_0} = \frac{D_0 - D_M}{D_0}$$

the proportional reduction in deviance due to the model of interest. Because of the monotone increase in L_M as a function of the complexity of the model, D shares this property with R^2 for normal models. This is appealing, particularly for comparing predictive power of nested models.

One can also regard this measure as a proportional reduction in variation measure (Section 2.2), identifying a subject's contribution to the deviance as the component of the variation measure. Its numerical value can be difficult to interpret, however, since the log-likelihood is often not a natural scale for interpretation. The population value of this measure, namely the limit of D as $n \to \infty$, can be interpreted (B. Zheng, unpublished dissertation, 1997) [16, 17] as the ratio of the expected Kullback–Leibler distance between the model of interest and the null model to that between the saturated model and the null model, where the expectation is taken with respect to the distribution of X.

As with all association measures, the magnitude of D depends on the level of aggregation of the observations. We mention this here to caution how one obtains the deviance for the saturated model. For a binary response with categorical predictors, for instance, D assumes different values according to whether one specifies the individual observations as the binary (Bernoulli) outcomes for individual subjects or the grouped binomial counts. This is due to the difference in L_S in the two cases. For the individual binary outcomes, $L_S = 0$, whereas for the grouped data L_S is a function of the sample proportions. The population values of D can be considerably different. For grouped observations with categorical predictors, for instance, this value equals 1 when model-specified probabilities are equal to the true population probabilities even though predictions for individual binary observations are far from perfect. For most purposes, D based on the grouped data is not an appropriate measure of predictive power, and we recommend basing D on the likelihood for the ungrouped data.

A variety of other likelihood-based measures have also been proposed. For instance, for a normal response with constant variance, $R^2 = 1 - (\ell_0/\ell_M)^{2/n}$, and one might consider using the measure $1-(\ell_0/\ell_M)^{2/n}$ more generally [18, 19]. For the Cox model for survival data with censored response, simulation studies have suggested that this measure performs well [19], whereas the measure [20] D (with a correction for the number of parameters in the model) may increase with the amount of censoring [12]. An adjustment to $1 - (\ell_0/\ell_M)^{2/n}$ has been proposed [21] so it can attain the value 1. For binary outcomes, yet other likelihood-based measures have been proposed [22], such as $1-(L_0/L_M)^{(2/n)L_0}$. Again, these measures have the disadvantage that the log-likelihood or likelihood raised to some power may not be a natural scale for interpretation.

2.4. Other Measures

Other measures apply only for specific sorts of data rather than all GLMs. For a binary response, for instance, a measure of predictive ability is the fraction of incorrectly classified responses, called the prediction error rate [6]. One obtains the predicted probability from a fitted model and predicts $Y = 1$ if it is greater than a cut-off point and predicts $Y=0$ otherwise. Similar indices could be defined for any type of data, but an obvious disadvantage is the dependence on the choice of the cut-off point and the failure to distinguish between widely dissimilar predicted values with reference to that cut-off point.

3. A CORRELATION MEASURE OF PREDICTIVE POWER FOR GLMs

This section presents a measure of predictive power for a GLM that we feel is often more useful than the measures just described. Regarding a model as providing good predictions of a response Y if $\hat{Y}$ correlates strongly with Y, we suggest the correlation between Y and $\hat{Y}$ as a simple measure of predictive power for a GLM. The corresponding population measure is the correlation between Y and the conditional mean of Y, which we denote by $\text{cor}(Y, E(Y|X))$.

In ordinary linear regression, for which the conditional variance $\text{var}(Y|X)$ is assumed to be constant, the multiple correlation coefficient between the response and the predictors is $\mathbf{R} = [1 - \frac{\text{var}(Y|X)}{\text{var}(Y)}]^{1/2}$. In that case [23], $\text{cor}(Y, E(Y|X)) = \mathbf{R}$. Thus, $\text{cor}(Y, E(Y|X))$ then relates to the proportion of variability explained by a model. For the single predictor case, $\text{cor}(Y, E(Y|X)) = |\beta|\sqrt{\{\frac{\text{var}(X)}{\text{var}(Y)}\}}$. The measure then has intuitive appeal: the larger the effect size, the stronger the correlation.

We prefer $\text{cor}(Y, E(Y|X))$ to its square because of the appeal of working on the original scale and, in particular, its proportionality to the effect size. Although our emphasis in this paper is on the correlation scale, we realize that many statisticians prefer to summarize predictive power using its square, and we note that analogous results hold for that measure. Although restrictions on values of response variables may imply limitations on possible values for either measure [7], we suggest using them in a comparative sense for different models applied to the same data set, for which such limitations have less relevance.

For an arbitrary GLM, it is straightforward to see that $\text{cor}(Y, E(Y|X))$ is invariant to a location-scale transformation on X. In addition

$$\text{cov}(Y, E(Y|X)) = E[Y E(Y|X)] - EY E(E(Y|X)) = \text{var}(E(Y|X))$$

$$\text{cor}(Y, E(Y|X)) = \frac{\text{cov}(Y, E(Y|X))}{[\text{var}(Y)\text{var}(E(Y|X))]^{1/2}} = \left[1 - \frac{E[\text{var}(Y|X)]}{\text{var}(Y)}\right]^{1/2}$$

That is, $\text{cor}(Y, E(Y|X))$ equals the positive square root of the average proportion of variance explained by the predictors, which generalizes the corresponding relationship between $\text{cor}(Y, E(Y|X))$ and $\mathbf{R}$ in ordinary linear regression.

Some properties that hold for simple (normal) linear regression do not hold more generally. In linear regression, this measure equals the square root of R^2 as defined in Section 2.2, but for an arbitrary GLM, $\sqrt{R^2}$ need not equal the correlation between the response and the model predicted value. Also, for an arbitrary GLM the measure $\text{cor}(Y, E(Y|X))$ is not guaranteed to be monotone increasing in the complexity of the linear predictor, although this almost always seems to happen in practice. Also, the proportional relationship between β and $\text{cor}(Y, E(Y|X))$ for univariate X does not hold for an arbitrary GLM, although one can show (B. Zheng, unpublished dissertation, 1997) that an approximate relationship of this type exists when β is close to 0.

Compared to the measures introduced in Section 2, the correlation measure has the advantage of being applicable to all types of GLMs and having a familiar interpretation. It provides greater information about predictive power than the index c because it uses the actual response instead of its ranking in evaluating the predictive power. It has the deficiency of being potentially sensitive to outliers, from which c does not suffer. It differs from D in that the latter depends on the choice of distribution for the random component through the form of its likelihood as well as the fitted values. That is, two distributions that provide the same fitted values necessarily have the same $\text{cor}(Y, \hat{Y})$ values, but they will typically not have the same value for D because of the difference in likelihoods.

4. COMPARING ESTIMATORS OF THE CORRELATION MEASURE

We now study the performance of four estimators of the correlation measure: the sample estimator; a jack-knife estimator; a modified jack-knife estimator, and a leave-one-out cross-validation estimator. The sample estimator is $\text{cor}(Y, \hat{Y})$. We expected this estimator to show positive bias, since $\hat{Y}$ is a function of Y, and other estimators were studied as ways of potentially reducing that bias. A fifth possible estimator, the ML estimator of $\text{cor}(Y, E(Y|X))$, requires distributional assumptions on X. This is usually difficult to justify in practice, so we do not consider it here.

Table I. Summary of estimated bias and root MSE for estimators of correlation measure: logistic regression with a normal predictor.

True correlation	n	$\text{cor}(Y, \hat{Y})$		$\text{cor}(Y, \hat{Y})_{\text{jack}}$		$\text{cor}(Y, \hat{Y})^0_{\text{jack}}$		$\text{cor}(Y, \hat{Y}_{\text{crs}})$	
		Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$
0	50	0.117	0.146	0.060	0.201	0.109	0.146	−0.312	0.436
	100	0.078	0.098	0.036	0.154	0.074	0.098	−0.285	0.399
	200	0.059	0.074	0.033	0.117	0.058	0.074	−0.220	0.334
0.04	50	0.070	0.115	0.007	0.206	0.063	0.118	−0.372	0.486
	100	0.045	0.080	0.006	0.154	0.042	0.081	−0.291	0.406
	200	0.025	0.055	0.005	0.104	0.024	0.056	−0.223	0.332
0.40	50	−0.001	0.130	−0.001	0.130	−0.001	0.130	−0.112	0.221
	100	−0.000	0.089	−0.000	0.088	−0.000	0.088	−0.048	0.115
	200	0.001	0.066	0.001	0.066	0.001	0.066	−0.020	0.074
0.80	50	0.009	0.074	0.009	0.072	0.009	0.072	−0.024	0.082
	100	0.003	0.048	0.000	0.048	0.000	0.048	−0.011	0.051
	200	−0.000	0.034	−0.002	0.034	−0.002	0.034	−0.007	0.036

Let $\text{cor}(Y, \hat{Y})(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \text{cor}(Y, \hat{Y})^{(-i)}$, where $\text{cor}(Y, \hat{Y})^{(-i)}$ is the correlation for the sample with the ith observation removed. The jack-knife estimator [24] is

$$\text{cor}(Y, \hat{Y})_{\text{jack}} = n\,\text{cor}(Y, \hat{Y}) - (n-1)\text{cor}(Y, \hat{Y})(\cdot)$$

The jack-knife estimator of the bias of $\text{cor}(Y, \hat{Y})$ is the difference between $\text{cor}(Y, \hat{Y})$ and $\text{cor}(Y, \hat{Y})_{\text{jack}}$. The estimator $\text{cor}(Y, \hat{Y})_{\text{jack}}$ can take negative values, and in practice one would likely replace such values by 0. Hence, practical usage corresponds to the modified jack-knife estimator $\text{cor}(Y, \hat{Y})^0_{\text{jack}} = \max[0, \text{cor}(Y, \hat{Y})_{\text{jack}}]$.

Next, we define a leave-one-out cross-validation estimator as follows. We fit the model to the sample without the ith observation. The fitted value for the ith observation, denoted by $\hat{Y}^{(-i)}$, then depends on the parameters estimated from the remaining $n-1$ observations. The estimator is $\text{cor}(Y, \hat{Y}_{\text{crs}})$, where $\hat{Y}_{\text{crs}} = (\hat{Y}^{(-1)}, \hat{Y}^{(-2)}, \ldots, \hat{Y}^{(-n)})$.

We conducted a small simulation study to compare these four estimators in terms of bias and mean squared error. The study considered models with a single predictor, distributed as either N(0, 1) or log-normal(0, 1). The response was generated according to either a logistic regression model or a Poisson regression model. The true correlation $\text{cor}(Y, E(Y|X))$ was determined by the model parameters; α was set at the arbitrary value 1.0 and β was selected to provide zero, low, medium and high correlations. The sample size equalled 50, 100 or 200. We obtained results for all combinations of the above factors, using a Sun SPARCstation 20 (Model 60) with GLIM. To keep the study computationally manageable, we selected 1000 Monte Carlo samples, which was adequate for discerning general patterns of behaviour of the estimators.

Table I, which displays the estimated bias and root MSE for logistic regression with a normally distributed predictor, is typical of the results. For this case, the model parameter values $\beta = (0, 0.1, 1.0, 4.0)$ correspond to $\text{cor}(Y, E(Y|X)) = (0, 0.04, 0.40, 0.80)$. As an indicator of the Monte Carlo error, in the null case the estimated standard error for the values provided in this table is roughly (0.003, 0.002, 0.001) for $n = (50, 100, 200)$, both for the bias and for $\sqrt{\text{MSE}}$ of the sample correlation measure. Table I shows that in the null and low correlation cases $\text{cor}(Y, \hat{Y})$ has

Table II. Summary of estimated bias and root MSE for sample and jack-knife estimators for overparameterized logistic regression models (X_2 unrelated to Y).

True correlation	n		$\text{cor}(Y, \hat{Y})$				$\text{cor}(Y, \hat{Y})_{\text{jack}}$			
			X_1	$X_1 + X_2$	$X_1 \times X_2$	X_2	X_1	$X_1 + X_2$	$X_1 \times X_2$	X_2
0.04	50	Bias	0.076	0.143	0.194	0.117	0.019	0.049	0.075	0.060
		$\sqrt{\text{MSE}}$	0.119	0.173	0.220	0.146	0.202	0.185	0.186	0.201
	100	Bias	0.043	0.086	0.122	0.078	0.006	0.021	0.041	0.036
		$\sqrt{\text{MSE}}$	0.078	0.111	0.143	0.098	0.147	0.131	0.131	0.154
0.40	50	Bias	−0.002	0.026	0.047	0.118	−0.002	−0.002	−0.004	0.061
		$\sqrt{\text{MSE}}$	0.131	0.129	0.133	0.148	0.131	0.138	0.145	0.206
	100	Bias	0.002	0.013	0.025	0.077	0.002	0.000	0.001	0.033
		$\sqrt{\text{MSE}}$	0.093	0.092	0.091	0.097	0.092	0.094	0.095	0.163
0.80	50	Bias	0.007	0.018	0.025	0.118	−0.001	−0.003	−0.004	0.082
		$\sqrt{\text{MSE}}$	0.072	0.075	0.077	0.144	0.071	0.072	0.078	0.199
	100	Bias	0.003	0.008	0.010	0.076	−0.000	−0.001	−0.001	0.035
		$\sqrt{\text{MSE}}$	0.050	0.051	0.051	0.095	0.050	0.050	0.050	0.163

a larger bias but smaller $\sqrt{\text{MSE}}$ than $\text{cor}(Y, \hat{Y})_{\text{jack}}$; for the medium and high correlation cases, the two estimators have little bias with similar $\sqrt{\text{MSE}}$. In terms of both bias and $\sqrt{\text{MSE}}$, $\text{cor}(Y, \hat{Y})^0_{\text{jack}}$ behaves similarly to $\text{cor}(Y, \hat{Y})$. By contrast, the cross-validation correlation has a negative bias that is surprisingly severe for the null and low correlation cases. It also has much larger MSE than the other estimators, except when the correlation is very large.

We now describe the reason for the poor performance of the cross-validation estimator, for a broad class of GLMs. First, we show that $\text{cor}(Y, \hat{Y}_{\text{crs}}) = -1$ for a null GLM with a canonical link. The null GLM contains only an intercept term and has $\text{cor}(Y, E(Y|X)) = 0$. With the canonical link, the likelihood equation is $\sum_{i=1}^{n} Y_i = n\alpha$, so the fitted value $\hat{Y} = \bar{Y}$. For the ith observation, its fitted value based on cross-validation is $\hat{Y}^{(-i)} = \frac{n}{n-1}\bar{Y} - \frac{1}{n}Y_i$. Hence, a perfect linear negative relationship exists between $\{Y_i\}$ and $\{\hat{Y}^{(-i)}\}$, and $\text{cor}(Y, \hat{Y}_{\text{crs}}) = -1$. This extreme bias extends to nearby non-null models. For instance, when we add a predictor to the model, $\hat{Y}^{(-i)}$ approaches $\frac{n}{n-1}\bar{Y} - \frac{1}{n}Y_i$ as the estimated effect approaches 0. So, when the association is weak, $\text{cor}(Y, \hat{Y}_{\text{crs}})$ can assume a large negative value and tends to have a strong negative bias.

Our second simulation study compared $\text{cor}(Y, \hat{Y})^0_{\text{jack}}$ and $\text{cor}(Y, \hat{Y})$ in terms of bias and MSE when the model is overparameterized. The predictors $X_1 \sim \text{N}(0, 1)$ and $X_2 \sim \text{N}(0, 1)$ were taken as independent, and the response Y was taken to depend on X_1 alone. When X_2 is added to the model the true summary measures do not change and the sample values should ideally be relatively insensitive to this overparameterization. We denote the models studied by the highest-order terms for the predictors for the model, for example, $X_1 + X_2$ denotes the model with X_1 and X_2 as predictors, and $X_1 \times X_2$ denotes the model also containing their interaction.

Table II, which displays results for the logistic regression model with $n = 50$ and 100, shows typical results. For models X_1, $X_1 + X_2$, and $X_1 \times X_2$, $\text{cor}(Y, \hat{Y})$ has a positive bias that increases as the model becomes more complex, an indication of susceptibility to overparameterization. The bias is severe for the low correlation case but relatively minor in the medium and high correlation cases. By contrast, $\text{cor}(Y, \hat{Y})_{\text{jack}}$ has a much smaller bias in the low correlation case and is essentially unbiased for the medium and high correlation cases. In addition, its bias remains almost constant as

Table III. Summary measures for various logistic regression models for low birth weight data set.

Model	Predictor	$\text{cor}(Y, \hat{Y})$	$\text{cor}(Y, \hat{Y})_{\text{jack}}$
M1	WT	0.18	0.18
M2	M1 + HT	0.28	0.27
M3	M2 + PL	0.32	0.30
M4	M3 + RACE	0.35	0.32
M5	M4 + SM	0.39	0.35
M6	M5 + AGE	0.39	0.35
M6a	M6 but with WTD and PLD	0.41	0.37
M6b	M6a but c-log-log link	0.41	0.37
M7	M6 + UI	0.41	0.35
M8	M7 + FTV	0.41	0.35
M9	M6a + AGE × PLD	0.42	0.37
M10	M6a + SM × RACE	0.42	0.36
M11	M6a + UI + AGE × WTD + SM × WTD	0.43	0.36

more terms are added to the model, indicating robustness to overparameterization. For the model $X_2, \text{cor}(Y, E(Y|X_2)) = 0$ since Y is dependent on X_1 only, but both $\text{cor}(Y, \hat{Y})$ and $\text{cor}(Y, \hat{Y})_{\text{jack}}$ show some bias, the bias being smaller for $\text{cor}(Y, \hat{Y})_{\text{jack}}$. Note, however, that the MSE is often considerably larger for $\text{cor}(Y, \hat{Y})_{\text{jack}}$ than $\text{cor}(Y, \hat{Y})$. Thus, although $\text{cor}(Y, \hat{Y})_{\text{jack}}$ can be useful for indicating when the addition of a term to the model is not vital, it is not generally as good an estimator of the population correlation as $\text{cor}(Y, \hat{Y})$.

Again, $\text{cor}(Y, \hat{Y})_{\text{jack}}$ achieves the goal of bias reduction at the cost of possible negative values when $\text{cor}(Y, E(Y|X))$ is zero or close to it. Our numerical evaluations, not reported here, showed that the modified version $\text{cor}(Y, \hat{Y})^0_{\text{jack}}$ is biased and acts much like the sample correlation. Bias reduction without generating negative values is not possible for $\text{cor}(Y, E(Y|X))$ or any measure that assumes non-negative values but can equal 0, since an unbiased estimator that assumes only non-negative values does not exist.

5. AN EXAMPLE: MODELS PREDICTING LOW BIRTH WEIGHT

This section illustrates the correlation measures and their use in helping to evaluate logistic regression models. We use data from a study of risk factors associated with low infant birth weight [25]. The data were collected on 189 women, 59 of whom had low birth weight babies and 130 of whom had normal birth weight babies. Risk factors included age of the mother (AGE), weight at the last menstrual period (WT), race (RACE), smoking status during pregnancy (SM), history of premature labour (PL), history of hypertension (HT), presence of uterine irritability (UI), and number of physician visits during the first trimester (FTV). The response is an indicator of low birth weight, namely birth weight $\geqslant 2500$ g versus birth weight < 2500 g. (If available, of course, one would prefer to utilize the continuous measure itself.)

Table III displays the measures for a variety of models, where + represents adding a predictor to the preceding model. Models M6a and M6 have the same predictors, but following a model suggested by Hosmer and Lemeshow (Reference [25], p. 98), WT and PL are dichotomized into

WTD and PLD in M6a. Table III shows a variety of models. We selected the best single-predictor model based on the summary measures $\text{cor}(Y, \hat{Y})$ and $\text{cor}(Y, \hat{Y})_{\text{jack}}$. We denote it by M1. We then added the predictor whose addition leads to the maximum improvement in the summary measures. We proceeded in this fashion until the improvement in the summary measures was minor. In addition to the risk factors mentioned above, we also considered two-way interactions. We do not intend this as a suggested model selection device, but used it merely as a way to exhibit a variety of models that an analyst might consider for these data.

Except for M1, $\text{cor}(Y, \hat{Y})$ is consistently higher than $\text{cor}(Y, \hat{Y})_{\text{jack}}$, indicating a potential positive bias. Viewing the jack-knife estimate helps to suggest when the addition of a new term may not truly improve predictive power. Among the main effect models M1–M8, the measures improve with the complexity of the model until M5. They show only minor changes afterwards, particularly in terms of $\text{cor}(Y, \hat{Y})_{\text{jack}}$. Thus, M5 is a reasonable tentative choice of model, or one might consider M6 simply because AGE is biologically important and might interact with other risk factors (Reference [25], p. 95).

Dichotomizing WT and PL (that is, M6a) leads to a slight improvement, but adding interaction terms to M6 or M6a does not lead to substantive improvement (see models M9, M10, M11). Similarly, using alternative link functions did not help (for example, model M6b has the same predictors as M6a but uses the complementary log-log link). In summary, the measures lead us in the direction of M5, M6 and M6a as tentative choices to use as the basis of further model building; these have the interpretive advantage of containing only main effect terms.

Hosmer and Lemeshow selected model M11, based on statistical significance considerations. Although it includes three more predictors than M6a, its summary measures (particularly $\text{cor}(Y, \hat{Y})_{\text{jack}}$) are almost identical to those for M6a, and thus it gains little, if any, predictive power. This illustrates the well known adage that statistical significance does not imply practical importance. The summary measures can help us use practical importance as well as statistical significance as a criterion in evaluating models.

6. BOOTSTRAP CONFIDENCE INTERVALS AND OPTIMISM DETERMINATION

In some applications it may be useful to construct an interval estimate for the population value of the correlation measure, $\text{cor}(Y, E(Y|X))$. Although it seems difficult to obtain analytical results in this direction, it is simple to use bootstrap methods for this purpose. For each type of bootstrap interval method [26] (for example, percentile, BCa), one can adopt two approaches to Monte Carlo sampling [27] to generate the bootstrap samples.

In the parametric bootstrap approach, predictors are considered fixed and the response for a Monte Carlo sample is generated according to the model, using the parameter values estimated from the observed sample. A bootstrap sample thus generated conforms to the model specification. A caution is in order for this approach. When the model is misspecified, the intervals can be misleading. For instance, suppose that one uses a Poisson regression model but there is marked overdispersion. Then, using the parametric bootstrap with the Poisson model will imply a stronger correlation than actually exists, since the parametric bootstrap samples will tend to exhibit much less dispersion than the sample counts. In the non-parametric approach, a Monte Carlo sample is generated from the empirical distribution having mass $1/n$ at each (X, Y) observation. Such a bootstrap sample does not rely on any assumption about the model, which is advantageous when the model exhibits lack of fit.

We illustrate bootstrap interval estimation for model M6a with the low birth weight data. For those data, the sample correlation equals 0.41. Based on using 10,000 Monte Carlo samples to form percentile bootstrap intervals, the non-parametric interval is (0.32,0.56) and the parametric interval is $(0.34, 0.58)$.

Another potential use of bootstrap methods is to estimate the amount of potential overfitting, or *optimism*, in $\text{cor}(Y, \hat{Y})$ associated with a model selection process [28]. Optimism summarizes the discrepancy between the model performance on a new subject and that on the observed sample. Let (X_0, Y_0) denote a new subject drawn from the same population as the observed sample, and let $\hat{Y}_0$ denote the prediction based on the model. The model performance on a new subject is summarized by $E\{\text{cor}(Y_0, \hat{Y}_0)\}$, where the expectation is taken over the distribution of the new subject, with the sample values fixed. The amount of overfitting is summarized by the optimism, defined as $E\{\text{cor}(Y_0, \hat{Y}_0)\} - \text{cor}(Y, \hat{Y})$. The use of the bootstrap to estimate the optimism proceeds as follows [4]. For each Monte Carlo sample drawn with replacement from the observed data, one applies the same stopping rule used for the observed sample and selects a model [4]. Let cor_b denote the correlation measure for this Monte Carlo sample and let cor_s denote the same measure but applied to the observed sample, that is, cor_s is calculated using the observed responses and fitted values based on coefficients from the aforementioned model. Repeat this procedure 100 to 200 times. The bootstrap optimism, denoted by Op, is the average difference between cor_b and cor_s. The optimism-corrected correlation measure is then $\text{cor}(Y, \hat{Y}) - \text{Op}$. This reflects the performance of the model, or rather of the entire model selection process, on a new subject, given the observed data.

Section 4 showed that cross-validation can behave poorly in adjusting sample values for bias. Efron [28] showed that for the prediction error rate measure, the bootstrap optimism-corrected version has a larger bias but a smaller MSE while the cross-validation version is almost unbiased but has a larger MSE. A useful future project would be to study the performance of bootstrap optimism-corrected measures for a variety of GLMs.

7. CONCLUSIONS

This article has studied the correlation measure as a summary of the predictive power of a GLM. It generalizes the multiple correlation coefficient, and although it does not maintain the guaranteed monotonicity property, in our experience this property almost always holds in practice. It has the advantage of using the original scale, being numerically simple to interpret regardless of the choice of probability distribution for the GLM, and it is thus comparable in numerical value across GLMs with different links and choices of probability distribution. Of the estimators of the correlation, the sample estimator and the jack-knife estimator behave well while the cross-validation estimator does not. In future work, it might be useful to generalize the correlation measure to other cases, such as models for a multinomial response, and multivariate models for longitudinal data.

ACKNOWLEDGEMENTS

The work of Agresti on this article was partially supported by NIH grant GM 43824. The authors are grateful to a referee for many helpful suggestions.

REFERENCES

1. Mittlböck M, Schemper M. Explained variation for logistic regression. *Statistics in Medicine* 1996; **15**:1987–1997.
2. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Journal of the American Medical Association* 1982; **247**:2543–2546.
3. Hanley JA, McNeil BJ. The meaning of and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
4. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing error. *Statistics in Medicine* 1996; **15**:361–387.
5. Ash A, Shwartz M. R^2: a useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine* 1999; **18**:375–384.
6. Efron B. Regression and anova with zero-one data: measures of residual variation. *Journal of the American Statistical Association* 1978; **73**:113–121.
7. Cox DR, Wermuth N. A comment on the coefficient of determination for binary responses. *American Statistician* 1992; **46**:1–4.
8. Hilden J. The area under the ROC curve and its competitors. *Medical Decision Making* 1991; **11**:95–101.
9. Haberman SJ. Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association* 1982; **77**:568–580.
10. Theil H. On the estimation of relationships involving qualitative variables. *American Journal of Sociology* 1970; **76**:103–154.
11. Korn EL, Simon R. Measures of explained variation for survival data. *Statistics in Medicine* 1990; **9**:487–503.
12. Schemper M. The explained variation in proportional hazards regression (correction in 1994; **81**:631). *Biometrika* 1990; **77**:216–218.
13. Shannon CE. A mathematical theory of communication. *Bell System Technical Journal* 1948; **27**:379–423, 623–656.
14. Goodman LA. The analysis of multinomial contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* 1971; **13**:33–61.
15. McFadden D. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*. Academic Press: New York, 1974; 105–142.
16. Kullback S. *Information Theory and Statistics.* Wiley: New York, 1959.
17. Hastie T. A closer look at the deviance. *American Statistician* 1987; **41**:16–20.
18. Magee L. R^2 measures based on Wald and likelihood ratio joint significance tests. *American Statistician* 1990; **44**:250–253.
19. Schemper M. Further results on the explained variation in proportional hazards regression. *Biometrika* 1992; **79**:202–204.
20. Harrell FE. The PHGLM procedure. In *SUGI Supplemental Library User's Guide*, *Version 5 ed.*, Hastings RP (ed.). SAS institute Inc.: Cary, North Carolina, 1986; 437–466.
21. Cragg JG, Uhler R. The demand for automobiles. *Canadian Journal of Economics* 1970; **3**:386–406.
22. Estrella A. A new measure of fit for equations with dichotomous dependent variables. *Journal of Business & Economic Statistics* 1998; **16**:198–205.
23. Kendall SM, Stuart A. *The Advanced Theory of Statistics*: *Inference and Relationship*; volume 2. Macmillan Publishing Co.: New York, 1979; 335.
24. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993; 141.
25. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* Wiley: New York, 1989; 247.
26. Efron B. Better bootstrap confidence intervals. *Journal of the American Statistical Association* 1987; **82**:171–185.
27. Hall P. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag: New York, 1992; 7, 170.
28. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* 1983; **78**:316–331.

