STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:1265–1275

Summarizing the goodness of fit of generalized linear models for longitudinal data

Beiyao Zheng*,†

Wake Forest University School of Medicine, Department of Public Health Sciences, Medical Center Boulevard, Winston-Salem, NC 27157-1051, U.S.A.

SUMMARY

This paper extends four goodness-of-fit measures of a generalized linear model (GLM) to random effects and marginal models for longitudinal data. The four measures are the proportional reduction in entropy measure, the proportional reduction in deviance measure, the concordance correlation coefficient and the concordance index. The extended measures satisfy the basic requirements for measures of association. Two examples illustrate their use in model selection. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Goodness-of-fit measures of a generalized linear model (GLM) is of great interest to applied statisticians [1–3] (B. Zheng, unpublished dissertation, 1997) for at least two reasons. First, when used as criteria for model selection, they generally reflect practical importance and thus complement statistical significance consideration. An immediate example is the use of the coefficient of determination, R^2, for variable selection in linear regression. A statistically significant predictor might not be included in the model if its addition leads to little change in the fitted values, and thus resulting in little increase in R^2. Second, applying these measures is often straightforward because they mostly have a simple structure, a familiar interpretation, and a range of $(0, 1)$ or $(-1, 1)$ that allows easy evaluation of their magnitude. This paper proposes extensions of four such measures to GLMs for longitudinal data, in particular, a marginal model and a random effects model. The proposed measures satisfy the basic requirements for measures of association and provide practical tools for model selection. Section 2 reviews three GLM measures, including the proportional reduction in entropy measure, the proportional reduction in deviance measure, and the concordance index. It then studies the properties of the concordance correlation coefficient as a measure of the goodness of fit of a GLM. Sections 3 and 4 extend the four measures to the marginal model and the random effects model, respectively. Section 5 uses two examples to illustrate the proposed measures and their use in model selection. The paper ends in a discussion.

*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

Received April 1999
Accepted August 1999

For notation, the subscript t $(t=1,2,\ldots,T)$ refers to time and the subscript i $(i=1,2,\ldots,n)$ refers to an observed subject. Let Y represent a response variable and X represent the predictors, both treated as random. Here X could be a vector or a scalar. Let μ represent a prediction based on a model of interest and $\hat{Y}$ represent the fitted value. In a longitudinal setting, let $\mathbf{Y}$ denote the vector of repeated responses by a subject and $\boldsymbol{\mu}$ denote the corresponding vector of predictions. Throughout the paper, the null model fits only an intercept. Its prediction is $\mu_0 = E(Y)$, the marginal expectation of the response. The sample estimator of μ_0 is denoted by $\bar{Y}$. In a GLM, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. In a longitudinal setting, $\bar{Y} = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} Y_{it}$.

2. GOODNESS-OF-FIT MEASURES OF A GENERALIZED LINEAR MODEL

In a GLM, the model-based prediction is $\mu = E(Y|X)$, the conditional expectation of the response given the predictors. It depends on the predictors through parameters β and a link function g, as in $g(\mu) = X\beta$. The fitted value $\hat{Y}$ is the maximum likelihood (ML) estimate of μ, obtained by substituting in μ the ML estimate of β. In this section, we introduce four measures of the goodness of fit of a GLM: the proportional reduction in entropy measure; the proportional reduction in deviance measure; the concordance index and the concordance correlation coefficient.

2.1. Proportional reduction in entropy H

Entropy is often used to measure the uncertainty associated with the probability distribution of a categorical variable [4, 5]. For an integral value k $(1 \leqslant k \leqslant K)$, let $\pi_k = P(Y=k|X)$. The entropy [6] is defined as $\mathrm{En} = -\sum_{k=1}^{K} \pi_k \log(\pi_k)$. It lies between 0 and $\log(K)$, with a large value indicating great uncertainty. Let $p_k = P(Y=K)$. Let $\hat{\pi}_k$ and $\hat{p}_k$ be the ML estimates of π_k and p_k.

The measure H is defined as [4, 5]

$$H = 1 - \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{\pi}_{ik}\log(\hat{\pi}_{ik})}{n\sum_{k=1}^{K} \hat{p}_k \log(\hat{p}_k)}$$

It can be interpreted as the proportional reduction in entropy due to the model of interest. It has a range of $(0, 1)$. It equals 0 when there is no association between the response and the predictors and it equals 1 when the response falls into one category with estimated probability 1, which implies perfect prediction when the fitted model is correct. It is monotone increasing in the complexity of the model. Its population value, namely the limit of H as $n \to \infty$, is $1 - \frac{E[\sum_{k=1}^{K}\pi_k \log(\pi_k)]}{\sum_{k=1}^{K} p_k \log(p_k)}$.

2.2. Proportional reduction in deviance D

We denote the log-likelihood function under the model of interest by $l(\mu, \phi; Y)$, where ϕ is a dispersion parameter which equals 1 for a binomial or count response, and equals the variance parameter of a normal response. The maximum achievable log-likelihood is $l(Y, \phi; Y)$, which corresponds to perfect prediction. For a fixed value of ϕ, define [7, 8] the unit deviance $d(Y, \mu)$ as $d(Y, \mu)/\phi = -2(l(\mu, \phi; Y) - l(Y, \phi; Y))$. It measures the discrepancy between the response and the model-based prediction. Define the sample deviances under the model of interest and the null model as $\sum_{i=1}^{n} d_i(Y_i, \hat{Y}_i)$ and $\sum_{i=1}^{n} d_i(Y_i, \bar{Y})$, respectively.

The measure D is defined as [4, 9, 10]

$$D = 1 - \frac{\sum_{i=1}^{n} d_i(Y_i, \hat{Y}_i)}{\sum_{i=1}^{n} d_i(Y_i, \bar{Y})}$$

It gives the proportional reduction in deviance due to the model of interest. It has a range of (0,1). It equals 0 when there is no association between the response and the predictors and it equals 1 when there is perfect prediction. It is monotone increasing as a function of the complexity of the model. It is particularly useful for comparing the prediction of nested models. Its numerical value may or may not be difficult to interpret, depending on the form of the log-likelihood for the chosen distribution. In linear regression, when the response is normally distributed with a constant variance, it equals the sample coefficient of determination, $R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$. As the sample sizes increases, D approaches its population (B. Zheng, unpublished dissertation, 1997) value $1 - \frac{E[E\{d(Y,\mu)|X\}]}{E[E\{d(Y,\mu_0)|X\}]}$. The measure D (with a correction for the number of parameters in the model) has the deficiency of depending on the amount of censoring when applied to the Cox model for survival data with a censored response [11, 12]. More appropriate measures [13] should be used for such a model.

As can be seen, D and H have many properties in common because they both summarize proportional reduction in variation due to the model of interest. In fact, $D = H$ in a logistic regression model for a binary response.

2.3. Concordance index c

Consider those pairs of observations that are untied on Y. The index [14] c equals the proportion of such pairs for which $\hat{Y}$ and Y are concordant, the observation with the larger Y also having the larger $\hat{Y}$. For a binary response, c is identical to a widely used measure of diagnostic discrimination, the area under a receiver operating characteristic (ROC) curve [14, 15]. Since it depends on ranking information only, the index c cannot distinguish between models that yield the same orderings of the fitted values. For a binary response with a single linear predictor, for instance, it assumes the same value for models with logit and complementary log-log link functions, even though the models are quite different.

2.4. Concordance correlation coefficient r_c

The concordance correlation coefficient was originally proposed to evaluate the reproducibility of measurements generated by different instruments or assays. Let Y_1 and Y_2 be two readings on a subject. Let $\bar{Y}_1$ and $\bar{Y}_2$ denote the respective sample means. The concordance correlation coefficient ρ_c is defined as [16]

$$\rho_c(Y_1, Y_2) = \frac{2\mathrm{cov}(Y_1, Y_2)}{\mathrm{var}(Y_1) + \mathrm{var}(Y_2) + \{E(Y_1) - E(Y_2)\}^2}$$

It assesses the agreement between Y_1 and Y_2 under the constraints [16, 17] that the intercept is 0 and the slope is 1. It is less than or equal to the Pearson correlation coefficient which does not impose such constraints. It equals 1 when Y_1 and Y_2 are in perfect agreement (the point (Y_1, Y_2) falls on the 45 degree line through the origin) and it equals -1 when Y_1 and Y_2 are in perfect

reversed agreement (the point (Y_1, Y_2) falls on the 135 degree line through the origin). It equals 0 when there is no linear association between Y_1 and Y_2. Its sample estimator is defined as

$$r_c(Y_1, Y_2) = \frac{2\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)^2 + \sum_{i=1}^{n}(Y_{i2} - \bar{Y}_2)^2 + n(\bar{Y}_1 - \bar{Y}_2)^2}$$

It approaches ρ_c as $n \to \infty$.

A multivariate version of ρ_c has been proposed for a repeated measures design [17]. Let $\mathbf{Y}_1$ and $\mathbf{Y}_2$ denote two vectors of readings of length Q for a subject. Let $\Sigma = \mathrm{var}\binom{\mathbf{Y}_1}{\mathbf{Y}_2}$ denote the covariance matrix for the combined vector of readings $\binom{\mathbf{Y}_1}{\mathbf{Y}_2}$. Under the assumption that the response follows a random-coefficient growth curve model, Σ varies across subjects. For subject i, let (Y_{i1q}, Y_{i2q}) denote the qth pair of readings, $q=1,\ldots,Q$. Define $\rho_{ci} = \frac{1}{Q}\sum_{q=1}^{Q}\rho_c(Y_{i1q}, Y_{i2q})$, which depends upon the elements of Σ. The summary measure is defined as $\rho_c = (\sum_{i=1}^{n} w_i)^{-1}\sum_{i=1}^{n} w_i\rho_{ci}$, with weights w_i inversely related to the variations in $\mathbf{Y}_1$ and $\mathbf{Y}_2$ for subject i. The sample estimator is obtained by replacing Σ with its unbiased estimator.

Next we study the properties of ρ_c as a measure of the goodness of fit of a GLM. It is easy to see that

$$\rho_c(Y, \mu) = \frac{2\mathrm{cov}(Y, \mu)}{\mathrm{var}(Y) + \mathrm{var}(\mu)} = \frac{2\mathrm{var}(\mu)}{\mathrm{var}(Y) + \mathrm{var}(\mu)}$$

Thus ρ_c has a range of (0,1). It equals 0 when there is no association between the response and the predictors. It equals 1 when there is perfect prediction. In linear simple regression, its value increases with the magnitude of the slope and the variance of the predictor. If the variance of the response around the linear regression line is a constant, say σ^2, then $\rho_c = \frac{2R^2}{1+R^2}$, where R^2 is the coefficient of determination defined as $R^2 = 1 - \sigma^2/\mathrm{var}(Y)$. The measure ρ_c has wider application than R^2 by allowing non-constant variance. Its simple structure and correlation-type interpretation are appealing.

The sample estimator of $\rho_c(Y, \mu)$ is defined as

$$r_c(Y, \hat{Y}) = \frac{2\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}$$

In linear regression, it has a range of (0,1) and is related to the sample coefficient of determination by $r_c = \frac{2R^2}{1+R^2}$. In a GLM, it has a range of $(-1, 1)$. It equals -1 when the response and the fitted value are in perfect reversed agreement and equals 1 when there is perfect fit. It equals 0 when there is no linear association between the response and the fitted value. As $n \to \infty$, $r_c(Y, \hat{Y})$ approaches $\rho_c(Y, \mu)$.

In general, all four measures share the desirable properties mentioned in the introduction section. The goal of the next two sections is to extend them to a marginal model and a random effects model while retaining these properties.

3. GOODNESS-OF-FIT MEASURES OF A MARGINAL MODEL

Similar to a GLM, a marginal model specifies that $g(\mu) = X\beta$, with $\mu = E(Y|X)$. The variance of the response and the covariance matrix describing the correlations among the repeated responses

by a subject are functions of the mean and maybe some additional parameters [18]. Estimation is done by the generalized estimating equations (GEE) method. When the model for μ is correctly specified, the GEE estimate $\hat{\beta}$ approaches the underlying parameter β as $n\to\infty$. The fitted value $\hat{Y}$ is obtained by substituting $\hat{\beta}$ for β in μ.

A marginal model does not specify any likelihood function and therefore we consider extensions of non-likelihood-based measures, H, r_c and c. We also extend R^2, since it is a proportional reduction in variation measure and is applicable to both continuous and binary responses [19].

3.1. Measures of proportional reduction in variation H_{marg} and R^2_{marg}

For a marginal model, we summarize its cross-sectional variation following the way variation is summarized for a GLM in the previous section and define its total variation as the sum of the cross-sectional variation. The summary measures are defined analogous to their GLM counterparts. We illustrate with two measures H and R^2.

Let $\pi_{tk} = P(Y_t = k|X)$ denote the model-based probability that a categorical response at time t equals k and $\hat{\pi}_{tk}$ denote its estimate. Let $\alpha_k = P(Y = k)$ denote the marginal probability of response k and $\hat{\alpha}_k$ denote its estimate. The extension of H to a marginal model is defined as

$$H_{\text{marg}} = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\pi}_{itk}\log(\hat{\pi}_{itk})}{nT\sum_{k=1}^{K}\hat{\alpha}_k\log(\hat{\alpha}_k)}$$

Here the denominator is based on the estimated marginal probability $\hat{\alpha}_k$. When based on the cross-sectional probability $\alpha_{tk} = P(Y_t = k)$ and its estimate $\hat{\alpha}_{tk}$, it equals $n\sum_{t=1}^{T}\sum_{k=1}^{K}\hat{\alpha}_{tk}\log(\hat{\alpha}_{tk})$, which is the entropy associated with the model M that uses time as a single predictor. The resulting summary measure assumes the value of 0 regardless of fit when M itself is the model of interest. Owing to this observation, we define the denominator in its current form.

The measure H_{marg} can be interpreted as the proportional reduction in entropy due to the model of interest and reduces to H for $T = 1$. It equals 0 when there is no association between the response and the predictors and it equals 1 when the response at each time point falls into one category with estimated probability 1, which means perfect prediction if the fitted model is correct. It assumes a negative value when there is greater uncertainty in prediction under the model of interest than under the null model. When the mean model is correctly specified, by Taylor's theorem it approaches the population value, $1 - \frac{E\{\sum_{k=1}^{K}\pi_{tk}\log(\pi_{tk})\}}{\sum_{k=1}^{K}\alpha_k\log(\alpha_k)}$, as $n\to\infty$.

Although its population value is bounded by 0 and 1, H_{marg} has only an upper bound of 1. To achieve a lower bound of 0, a modified measure uses $nT\log(K)$ as the denominator and essentially changes the reference model from the null model to the one that assigns equal probabilities to the response categories. Such rescaling has a potential negative effect. It might make a considerable reduction in entropy appear small if $nT\log(K)$ is much greater than $-nT\sum_{k=1}^{K}\hat{\alpha}_k\log(\hat{\alpha}_k)$. Our recommendation is to obtain both measures and use the modified measure when the problem does not occur.

The extension of R^2 is similar. Here the summary measure takes the form

$$R^2_{\text{marg}} = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(Y_{it} - \hat{Y}_{it})^2}{\sum_{t=1}^{T}\sum_{i=1}^{n}(Y_{it} - \bar{Y})^2}$$

For the same reason mentioned earlier, the denominator measures the distance of the fitted value from the marginal mean $\bar{Y}$ rather than the cross-sectional mean. The measure R^2_{marg} shares the same

interpretation and intuitive appeal as R^2 and reduces to R^2 for $T = 1$. It equals 1, its upper bound, when there is perfect prediction. It equals 0 when there is no association between the response and the predictors. It assumes a negative value when the variation is greater under the model of interest than under the null model, indicating poor prediction. As $n \to \infty$, by Taylor's theorem it approaches the population value $1 - \frac{E\{\text{var}(Y|X)\}}{\text{var}(Y)}$, which has a range of (0,1).

3.2. Concordance correlation coefficient r_c and concordance index c

We consider two approaches to generalizing r_c and c to a marginal model. The first one summarizes the agreement between the response and the fitted value simultaneously across all the time points and the second one summarizes the cross-sectional agreement. We illustrate with r_c and the same ideas apply to c.

Let $\bar{\hat{Y}} = \frac{1}{Tn}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{Y}_{it}$. Under the first approach, the concordance correlation coefficient for a marginal model is defined as

$$r_c = \frac{2\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it} - \bar{Y})(\hat{Y}_{it} - \bar{\hat{Y}})}{\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it} - \bar{Y})^2 + \sum_{i=1}^{n}\sum_{t=1}^{T}(\hat{Y}_{it} - \bar{\hat{Y}})^2}$$

It has a range of $(-1, 1)$, with similar interpretation as its GLM counterpart. When the mean model is correctly specified, it approaches the population value $\rho_c = \frac{2\text{cov}(Y,\mu)}{\text{var}(Y)+\text{var}(\mu)}$ as $n \to \infty$.

Let $\bar{Y}_t = \frac{1}{n}\sum_{i=1}^{n} Y_{it}$ and $\bar{\hat{Y}}_t = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_{it}$. Under the second approach, the summary measure takes the form $r_c' = \frac{1}{T}\sum_{t=1}^{T}\frac{2\sum_{i=1}^{n}(Y_{it}-\bar{Y}_t)(\hat{Y}_{it}-\bar{\hat{Y}}_t)}{\sum_{i=1}^{n}(Y_{it}-\bar{Y}_t)^2+\sum_{i=1}^{n}(\hat{Y}_{it}-\bar{\hat{Y}}_t)^2}$. It is the average of T cross-sectional concordance correlation coefficients. Its population value is $\frac{1}{T}\sum_{t=1}^{T}\frac{2\text{cov}(Y_t,\mu_t)}{\text{var}(Y_t)+\text{var}(\mu_t)}$, which is the multivariate version of ρ_c (Section 2.4) applied to $(\mathbf{Y}, \boldsymbol{\mu})$ under the assumption that $\text{var}\binom{\mathbf{Y}}{\boldsymbol{\mu}}$ and w_i are invariant across subjects. The measure r_c' can be misleading due to its conditional approach. For instance, in a marginal model that uses time as a single predictor, $\hat{Y}_t$ is a constant cross-sectionally and $r_c' = 0$ regardless of the fit of the model. Owing to this deficiency, we do not recommend using this measure.

3.3. Discussion

The covariance matrix under a marginal model is missing from the above discussion. In our opinion, goodness of fit is concerned with the agreement between the response and the prediction. The covariance matrix is only relevant to the point that it affects the fitted value through the parameter estimates, but is not of interest by itself. A measure that incorporates the covariance matrix may assume a low value because of poor fit and/or inappropriate correlation structure, and therefore confounds goodness of fit with correlation modelling. Such a measure does not serve the unique purpose of summarizing the goodness of fit and therefore is not considered [20].

4. GOODNESS-OF-FIT MEASURES OF A RANDOM EFFECTS MODEL

A random effects model assumes that given the predictors and random effects, the repeated responses for a subject are mutually independent and follow a GLM [21]. Let b denote a vector of random effects, which is often assumed to follow a multivariate normal distribution with mean

0 and covariance matrix D. Let Z denote a design matrix for the random effects. The prediction is $\mu = E(Y|X, b)$, the conditional expectation of the response given the predictors and the random effects. It depends on the random effects through $g(\mu) = X\beta + Zb$, with β being the parameters for the fixed effects X.

For estimation, the joint marginal likelihood function of the repeated responses generally does not have a closed form. The estimates used here maximize an approximation to the likelihood function, called penalized quasi-likelihood (PQL) function [22]. The fitted value is obtained by replacing β and b in μ with their PQL estimates $\hat{\beta}$ and $\hat{b}$. It can be used to forecast for an observed subject. In the normal theory mixed linear model, it is called best linear unbiased prediction (BLUP) and is applied to animal breeding and various other fields [23–26].

To summarize the performance of a random effects model, we propose three measures: proportional reduction in entropy, proportional reduction in deviance, and proportional reduction in PQL. The concordance measures r_c and c can be similarly applied as in a marginal model.

4.1. Proportional reduction in entropy H_{rand}

For T repeated measurements on a categorical response with K categories, there are a total number of K^T response patterns. Let $\delta_o = P(\mathbf{Y} = o|X, b)$ denote the probability of response pattern o, conditioning on the predictors and random effects. It can be obtained by the conditional independence assumption. To illustrate, suppose a binary response is measured three times with probabilities of a positive response denoted by p_1, p_2 and p_3. Then $P(\mathbf{Y} = (1, 1, 0)|X, b) = p_1 p_2(1 - p_3)$. Let $\hat{\delta}_o$ be an estimate of δ_o. Let $p_o = P(\mathbf{Y} = o)$ and let $\hat{p}_o$ be its estimate. The entropy associated with the joint distribution of $\mathbf{Y}$ is defined as $\text{En} = -\sum_{o=1}^{K^T} \delta_o \log(\delta_o)$.

The summary measure H_{rand} is defined as

$$H_{\text{rand}} = 1 - \frac{\sum_{i=1}^{n} \sum_{o=1}^{K^T} \hat{\delta}_{io} \log(\hat{\delta}_{io})}{n \sum_{o=1}^{K^T} \hat{p}_o \log(\hat{p}_o)}$$

It can be interpreted as the proportional reduction in entropy due to the model of interest. It equals 1, its upper bound, when a response pattern occurs with estimated probability 1, which implies perfect prediction if the fitted model is correct. It equals 0 when the model of interest has the same degree of uncertainty as the null model and it assumes a negative value when the uncertainty is greater under the model of interest than under the null model. Similar to H_{marg}, a modified measure uses $nT \log(K)$ as the denominator and achieves a lower bound of 0.

4.2. Proportional reduction in deviance D_{rand}

Let $L(\boldsymbol{\mu}, \phi; \mathbf{Y})$ denote the log of the joint likelihood function given the predictors and random effects. Let $d_{\text{rand}}(\mathbf{Y}, \boldsymbol{\mu})$ denote the unit deviance associated with $\mathbf{Y}$, and define $d_{\text{rand}}(\mathbf{Y}, \boldsymbol{\mu})/\phi = -2(L(\boldsymbol{\mu}, \phi; \mathbf{Y}) - L(\mathbf{Y}, \phi; \mathbf{Y})) = \sum_{t=1}^{T} d(Y_t, \mu_t)/\phi$. The last equality holds by the conditional independence assumption. Define the deviance under the model of interest as $\sum_{i=1}^{n} \sum_{t=1}^{T} d_{it}(Y_{it}, \mu_{it})$. The negative of the PQL is defined as [22]

$$-\text{PQL} = \frac{1}{2\phi} \sum_{i=1}^{n} \sum_{t=1}^{T} d_{it}(Y_{it}, \mu_{it}) + \frac{1}{2} b^t D^{-1} b$$

It is the sum of the scaled deviance and a penalty term for the random effects. The PQL estimates $\hat{\beta}$ and $\hat{b}$ maximize PQL, or equivalently minimize $-$PQL. Let $-\text{PQL}_\text{M}$ and $-\text{PQL}_\text{N}$ denote the respective minimized $-$PQL under the model of interest and the null model. Then the following relationship holds:

$$0 < \frac{1}{2\phi}\sum_{i=1}^{n}\sum_{t=1}^{T} d_{it}(Y_{it}, \hat{Y}_{it}) < -\text{PQL}_\text{M} < -\text{PQL}_\text{N} = \frac{1}{2\phi}\sum_{i=1}^{n}\sum_{t=1}^{T} d_{it}(Y_{it}, \bar{Y})$$

The summary measure is defined as

$$D_\text{rand} = 1 - \frac{\sum_{i=1}^{n}\sum_{t=1}^{T} d_{it}(Y_{it}, \hat{Y}_{it})}{\sum_{i=1}^{n}\sum_{t=1}^{T} d_{it}(Y_{it}, \bar{Y})}$$

It can be interpreted as the proportional reduction in deviance due to the model of interest, with a large value indicating good fit. It lies between 0 and 1 by the previous relationship. It equals 0 when the model of interest provides no improvement in prediction over the null model and it equals 1 when there is perfect prediction. In the normal theory mixed linear model case, $D_\text{rand} = 1 - \frac{\sum_{i=1}^{n}\sum_{t=1}^{n}(Y_{it}-\hat{Y}_{it})^2}{\sum_{i=1}^{n}\sum_{t=1}^{n}(Y_{it}-\bar{Y})^2}$, which is similar in form to R^2 in linear regression.

4.3. Proportional reduction in PQL P_rand

The third measure is defined as

$$P_\text{rand} = 1 - \frac{-\text{PQL}_\text{M}}{-\text{PQL}_\text{N}} = 1 - \frac{\sum_{i=1}^{n}\sum_{t=1}^{T} d_{it}(Y_{it}, \hat{Y}_{it})/(2\phi) + \hat{b}^t\hat{D}^{-1}\hat{b}/2}{\sum_{i=1}^{n}\sum_{t=1}^{T} d_{it}(Y_{it}, \bar{Y})/(2\phi)}$$

It gives the proportional reduction in PQL due to the model of interest. A large value indicates both good prediction and small random effects while a small value suggests lack of fit and/or large random effects. It has a range of $(0, 1)$. It equals 1 when there is perfect prediction and all the random effects are 0. It equals 0 when prediction is poor and/or random effects are large. It is monotone increasing in the complexity of the model.

The penalty for the random effects in P_rand is advantageous. For a random effects model, one can potentially improve its prediction by incorporating more and more random effects. The measure P_rand serves as a check on such overfitting by balancing the reduction in the deviance with a possible increase in the penalty term. Since its value is affected by both the quality of prediction and the contribution of the random effects, it is mostly useful as a measure of the overall performance of a model. Similar measures that also incorporate penalty for overfitting are Akaike's information criterion (AIC) [27] and Schwarz's Bayesian information criterion (BIC) [28], which are essentially the maximized log-likelihood penalized by the number of predictors. They are used in PROC LOGISTIC in SAS for stepwise model selection and PROC MIXED in SAS for selecting the variance structure for normal theory mixed linear models. The measure P_rand plays a similar role for a random effects models that AIC and BIC play for the other two models.

5. EXAMPLES

In this section, we present two examples to illustrate the measures and their use in model selection. Data for the first example [29] consist of growth measurements for 11 girls and 16 boys at ages 8,

Table I. Summary measures for models of children's growth data.

	Random intercept					Random intercept and slope		
Measure	1	G	A	$A+G$	$A \times G$	A	A+G	$A \times G$
D_{rand}	0.53	0.52	0.81	0.81	0.83	0.86	0.86	0.86
P_{rand}	0.51	0.50	0.79	0.79	0.80	0.82	0.83	0.83
r_c	0.64	0.64	0.89	0.89	0.90	0.92	0.92	0.92
c	0.76	0.76	0.90	0.90	0.90	0.91	0.92	0.92

10, 12 and 14. Potential predictors are age (A), gender (G), and their interaction. We considered two types of models, those that include a random intercept and those that also include a random slope for age. All models used the identity link function.

Table I displays the summary measures D_{rand}, P_{rand}, r_c and c for the models considered. The models are represented by their highest order predictor. For instance, '1' represents a model with only a random intercept, $A+G$ represents the model with predictors age and gender, and $A \times G$ represents the model also including their interaction. The first five models include a random intercept, and the last three also include a random slope for age.

Adding gender to model 1 brings little change in the measures, but adding age to model 1 brings a considerable increase. The measures remain almost constant as more predictors are introduced into model A. Thus among models with a random intercept, model A seems to fit as well as more complex ones. Adding a random slope to model A leads to a slight increase in the measures, which remain almost constant for additional predictors. Therefore the two models with age as a predictor provide reasonable choices. The predictor gender and its interaction with age contribute little to prediction according to the proposed measures. However, they might be important for a subset of the subjects. This possibility is ruled out in this example because the addition of gender and its interaction with age to the two models A brings little change in the fitted values (less than 5 per cent in magnitude). In addition, the fitted values correlate about 0.99 (by concordance correlation coefficient) between model A and model $A \times G$ with a random slope. This evidence, combined with the proposed measures, shows that gender and its interaction with age is of minimal practical importance, although the interaction has statistical significance ($p = 0.024$). This demonstrates that statistical significance is different from practical importance and the proposed measures can help us make a more informative choice.

Data for the second example came from a clinical trial [30] comparing two treatments for a respiratory illness. A total number of 111 subjects from two centres were randomly assigned to a treatment and placebo group. For each subject, a binary outcome on the respiratory status was determined at four visits. Potential explanatory variables were centre (C), sex (S), age (A) at time of study, visit (V), treatment (T), and a binary baseline (B) respiratory status. We considered marginal models with the logit link and a unstructured covariance matrix. For model selection, we first chose the best single predictor based on the summary measures H_{marg}, R^2_{marg}, r_c and c. We next 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.

Table II displays the summary measures for a subset of the models considered. As expected, the baseline respiratory status is the best single predictor. Adding treatment to the model B leads to a considerable increase in all the measures. Adding age to the model $B+T$ increases the magnitude

Table II. Summary measures for models of respiratory disease data.

Measure	B	$B+T$	$B+T+A$	$B+T+C+A$	$B+T+C+A+S+V$
H_{marg}	0.14	0.23	0.23	0.25	0.25
R^2_{marg}	0.17	0.19	0.19	0.20	0.20
r_c	0.30	0.38	0.38	0.40	0.41
c	0.49	0.67	0.78	0.79	0.79

of the concordance index c, but not of other measures, showing that these other measures fail to capture the improvement in prediction. The summary measures remain almost constant for more complex models. Therefore the marginal models with baseline and treatment or baseline, treatment and age as covariates provide reasonable choices. Other predictors, including centre, sex and visit, contribute little to prediction. They also lack statistical significance given baseline respiratory status and treatment (both significant) in the model. Thus in this example, we reach similar conclusions whether considering the summary measures or statistical significance tests.

6. DISCUSSION

In this paper, we proposed natural and reasonable generalizations of four goodness-of-fit measures of a GLM to marginal and random effects models for longitudinal data. The four measures are the proportional reduction in entropy measure, the proportional reduction in deviance measure, the concordance correlation coefficient and the concordance index. The proposed measures have a simple structure, a familiar interpretation, a range of $(0, 1)$ or $(-1, 1)$, with the exception of R^2_{marg}. They result in their GLM counterparts when $T = 1$. In the case of a marginal model, their population parameters are analogous to those of their GLM counterparts. In addition to extending the four GLM measures, we also proposed a measure that penalizes overfitting in a random effects model.

The proposed measures apply to data with missing observations as well. Except for H_{rand}, the summation over subjects in their definitions is replaced by the summation over the varying number of subjects with the response and predictors available at a particular time point. For H_{rand}, the summation is over subjects with complete observations on the response and the predictors across all the time points. The interpretation, range and population values of these measures remain unchanged.

Standard errors and confidence intervals for these measures may be obtained using the bootstrap method. For future research, the GLM measures may be extended to provide detailed diagnostic information regarding goodness of fit at each time point or for each subject. They also hold great potential for other types of models, such as models for survival data.

REFERENCES

1. Mittlböck M, Schemper M. Explained variation for logistic regression. *Statistics in Medicine* 1996; **15**:1987–1997.
2. Ash A, Shwartz M. R^2: a useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine* 1999; **18**:375–384.
3. Estrella A. A new measure of fit for equations with dichotomous dependent variables. *Journal of Business & Economic Statistics* 1998; **16**:198–205.

4. Theil H. On the estimation of relationships involving qualitative variables. *American Journal of Sociology* 1970; **76**:103–154.
5. Haberman SJ. Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association* 1982; **77**:568–580.
6. Shannon CE. A mathematical theory of communication. *Bell System Technical Journal* 1948; **27**:379–423, 623–656.
7. McCullagh P, Nelder JA. *Generalized Linear Models*, 2nd edn. Chapman & Hall: New York, 1989; 33.
8. Jørgensen B. *The Theory of Dispersion Models.* Chapman & Hall: New York, 1997; 4.
9. 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.
10. McFadden D. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics.* Academic Press: New York, 1974; 105–142.
11. 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.
12. Schemper M. The explained variation in proportional hazards regression (correction in 1994; **81**:631). *Biometrika* 1990; **77**:216–218.
13. Schemper M. Further results on the explained variation in proportional hazards regression. *Biometrika* 1992; **79**: 204–202.
14. 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.
15. Hanley JA, McNeil BJ. The meaning of and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
16. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**:255–268.
17. Chinchilli VM, Martel JK, Kumanyika S, Lloyd T. A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics* 1996; **52**:341–353.
18. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
19. Efron B. Regression and anova with zero-one data: Measures of residual variation. *Journal of the American Statistical Association* 1978; **73**:113–121.
20. Goodman LA, Kruskal WH. Measures of association for cross classifications. *Journal of the American Statistical Association* 1954; **49**:732–764.
21. Diggle PJ. Liang KY, Zeger SL. *Analysis of Longitudinal Data.* Oxford University Press: Oxford, 1994.
22. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
23. Henderson CR. *Applications of Linear Models in Animal Breeding.* University of Guelph: Guelph, Canada, 1984.
24. Robinson GK. That blup is a good thing: the estimation of random effects. *Statistical Science* 1991; **6**:15–51.
25. Mclean RA, Sanders WL, Stroup WW. A unified approach to mixed linear model. *American Statistician* 1991; **45**: 54–64.
26. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for mixed Models.* SAS Institute Inc.: Cary, NC, 1996.
27. Akaike H. A new look at the statistical model identification. *IEEE Transaction on Automatic Control* 1974; **AC-19**:716–723.
28. Schwarz G. Estimating the dimension of a model. *Annal of Statistics* 1978; **6**:461–464.
29. Pothoff RF, Roy SN. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 1964; **51**:313–326.
30. Stokes ME, CS Davis, GG Koch. *Categorical Data Analysis using the SAS System*. SAS Institute: Cary, NC, USA, 1995.

