Received: 2 June 2016 | Revised: 7 June 2018 | Accepted: 20 July 2018

DOI: 10.1002/sim.7941

RESEARCH ARTICLE

WILEY Statistics in Medicine

Goodness of fit tests for random effect models with binary responses

Antonia K. Korre | Vassilis G. S. Vasdekis

Department of Statistics, Athens University of Economics and Business, Athens, Greece

Correspondence
Antonia K. Korre, Department of Statistics. Athens University of Economics and Business, 76 Patission Str., GR10434 Athens, Greece.
Email: ant_korre@yahoo.com

Correlated binary responses are very common in longitudinal and repeated measures studies. Random effects models are often used to analyze such data and the h-likelihood estimating procedure provides an inferential tool. The objective of this study is to introduce goodness-of-fit score statistics for the fixed part of these models. The proposed statistics are based on processes that partition the observations into mutually exclusive groups. Weighted versions of these statistics are also introduced and are based on the correlation between an appropriately adjusted candidate covariate for entrance into the model and the model residuals. A simulation study indicates that the weighted statistics perform better than their unweighted counterparts, whereas the statistics that are based on the partitioning of the covariate space seem to perform slightly better compared with those based on other grouping procedures. The use of the proposed statistics is illustrated using a real data example.

KEYWORDS

binary correlated data, generalized linear mixed models, goodness-of-fit tests, h-likelihood, score statistics

1 | INTRODUCTION

In recent years, mixed-effects models have been proven to be particularly useful in the modeling of binary correlated responses. Random effects are unobserved random variables employed to capture associations and heterogeneity not captured by explanatory variables. Model checking techniques for the assessment of the goodness of fit of such models have been proposed by Evans and Hosmer[1] and Sturdivant and Hosmer,[2] merely as sums of squared residuals statistics, while Pan and Lin[3] proposed procedures based on cumulative sums of residuals. Another class of model checking techniques are statistics based upon partitioning of the covariate space or statistics using groups based on the ordered estimated probabilities. They were initially proposed in the ordinary logistic regression setting for sparse binary data cases and later, were extended into the Generalized Estimating Equations methodology by Barnhart and Williamson,[4] Horton et al,[5] or Evans and Li.[6] It appears that this type of statistics has not been applied in mixed effects models for binary data.

The aim of this article is to provide extensions of goodness-of-fit procedures proposed by Tsiatis,[7] Hosmer and Lemeshow,[8] and Pulkstenis and Robinson[9] for testing the fixed part of random effects models for binary correlated responses. Our interest is also to detect deviations towards specific alternatives of the fixed part of the mixed model. For that purpose, weighted extensions of these statistics are proposed in the spirit of Hosmer and Hjort.[10] The goodness-of-fit tests presented here are built within the h-likelihood estimating procedure framework. The procedure was proposed by Lee and Nelder[11-13] and is used here as an inferential tool to estimate the unknown model parameters and the unobservable random effects.

The article has the following structure. Section 2 presents the model setting along with the h-likelihood score equations used for the estimation of all unknown quantities. Section 3 suggests extensions of goodness-of-fit partitioning techniques to binary random effects models. It also introduces score statistics and suggests relevant weighted modifications for the assessment of the goodness of fit of the fixed part of a random effects model. Section 4 reports simulation results concerning the performance of the score tests, and finally, section 5 provides an application of these statistics using a real data set. Section 6 concludes the paper with a few remarks.

2 | MODEL AND SCORE EQUATIONS

We consider random effects models for binary correlated responses taken on a random sample of n subjects or clusters. Let Y_{ij} be the jth response of the ith subject or cluster, $i = 1, \ldots, n, j = 1, \ldots, n_i$ and let $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ be the $p \times 1$ and $q \times 1$ design vectors of fixed and random effects, respectively. We denote by $p_{ij} = E(Y_{ij} = 1|\boldsymbol{\upsilon})$ the conditional probability of success. The model that describes the dependence of Y_{ij} on the fixed and random effects has the form

$$\operatorname{logit}\left(p_{ij}\right) = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\top}\boldsymbol{\upsilon}, \tag{1}$$

where $\mathbf{x}_{ij}^{\top}\boldsymbol{\beta}$ is the fixed part, with $\boldsymbol{\beta}$ a p-dimensional fixed parameter vector and $\mathbf{z}_{ij}^{\top}\boldsymbol{\upsilon}$ the random part of the model with $\boldsymbol{\upsilon}$ the q-dimensional vector of random effects. Conditionally on the random effects $\boldsymbol{\upsilon}$, Y_{ij} are independent Bernoulli observations. The random effects $\boldsymbol{\upsilon}$ are assumed to be normally distributed with mean zero and covariance matrix Σ depending on the m-dimensional vector of parameters $\boldsymbol{\lambda}$. Each of these parameters is linked to a t-dimensional vector $\boldsymbol{\delta}$ through the model

$$\log(\lambda_r) = \boldsymbol{\psi}_r^{\top}\boldsymbol{\delta}, \qquad r = 1, \cdots, m, \tag{2}$$

where $\boldsymbol{\psi}_r$ is a t-dimensional design vector for the fixed effects $\boldsymbol{\delta}$. This is a generalized linear model for random components with log-link that may incorporate the effect of covariates as in Pourahmadi.[14] In the simple example of a random intercept model where all subjects are independent each other, $\boldsymbol{\upsilon} = (\upsilon_1, \cdots, \upsilon_n)^{\top}$, where $\operatorname{Var}(\upsilon_i) = \sigma_i^2$ for each i, and $\Sigma = \operatorname{diag}_{i=1,\cdots,n}\{\sigma_i^2\}$. In this case, $\lambda = \left(\sigma_1^2, \cdots, \sigma_n^2\right)^{\top}$ and $m = n$. Assuming further that random intercepts are homoscedastic with variance σ^2, $\boldsymbol{\psi}_r$, $r = 1, \ldots, n$ would be design scalars equal to one for the scalar parameter $\delta = \sigma^2$. Note that the setting described above is general enough to describe either positive or negative correlations between repeated binary measurements. The latter can be accomplished considering Σ to be a block diagonal matrix with $\Sigma = \operatorname{diag}_{i=1,\ldots,n}\{\Sigma_i\}$, where Σ_i is a $n_i \times n_i$ covariance matrix specific to subject i in an approach similar to Coull and Agresti.[15] Appropriate modeling of covariance parameters, assuming, for example, that the elements of Σ_i are the same for all i, using (2) is essential for estimability.

To make the following results more concrete, the design matrices $\boldsymbol{X}$, $\boldsymbol{Z}$, and Ψ are defined in such a way that their rows are given by the design vectors $\boldsymbol{x}_{ij}^{\top}$, $\boldsymbol{z}_{ij}^{\top}$, and $\boldsymbol{\psi}_r^{\top}$ and have dimensions $N \times p$, $N \times q$, and $m \times t$, respectively, where $N = \sum_{i=1}^{n} n_i$. Based on recommendations found in Yun and Lee,[16] we use the logarithm of the joint likelihood of the data and the random effects—called h-likelihood—for the estimation of $\boldsymbol{\upsilon}$ and approximations of the marginal and the restricted likelihoods denoted by $p_{\upsilon}(h)$ and $p_{\beta,\upsilon}(h)$, for the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, respectively. For the definition of these approximations and properties of the derived estimators, see other studies.[11,12,16,17]

It can be seen that, for the estimation of $\boldsymbol{\eta} = \left(\boldsymbol{\beta}^{\top}, \boldsymbol{\upsilon}^{\top}\right)^{\top}$, given $\hat{\boldsymbol{\delta}}$, maximization of the h-likelihood h and the approximate marginal likelihood $p_{\upsilon}(h)$ leads to an iterative weighted least squares equation of the form

$$\mathbf{T}^{\top}\Sigma_a^{-1}\mathbf{T}\hat{\boldsymbol{\eta}} = \mathbf{T}^{\top}\Sigma_a^{-1}\mathbf{z}_a, \tag{3}$$

with

$$\mathbf{T} = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0}_{q\times p} & \mathbf{I}_q \end{pmatrix} \quad , \quad \Sigma_a^{-1} = \Gamma_a^{-1}\mathbf{W}_a = \begin{pmatrix} \mathbf{I}_N & \mathbf{0}_{N\times q} \\ \mathbf{0}_{q\times N} & \Sigma^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{W} & \mathbf{0}_{N\times q} \\ \mathbf{0}_{q\times N} & \mathbf{I}_q \end{pmatrix},$$

$N = \sum_{i=1}^{n} n_i$ and $\mathbf{W} = \operatorname{diag}\left\{ \left(\frac{\partial p_{ij}}{\partial \operatorname{logit}(p_{ij})}\right)^2 \left(p_{ij}\left(1 - p_{ij}\right)\right)^{-1} \right\}$. For ease of notation below, we drop the double index i,j and use the single index $k = 1, \ldots, N$ to denote both the individual/cluster and the repeated observations within each cluster. The adjusted dependent variate $\mathbf{z}_a$ in (3) is written as $\left(\mathbf{z}_{a0}^{\top}, \ \mathbf{z}_{a1}^{\top}\right)^{\top}$, with the kth element of the $N \times 1$ vector $\mathbf{z}_{a0}$ being of the form $\mathbf{x}_k^\top \boldsymbol{\beta} + \mathbf{z}_k^\top \boldsymbol{\upsilon} + (y_k - s_k - p_k)/w_k$ and with $\mathbf{z}_{a1}$ being the $q \times 1$ vector of the form $\Sigma \mathbf{Z}^\top \mathbf{s}$. The kth element of the N-dimensional vector $\mathbf{s}$ has the form

$$s_k = \frac{w_k}{2}\left\{ P_{2k} w_k^2 \frac{\partial w_k}{\partial p_k}\frac{\partial p_k}{\partial g(p_k)} + \sum_{l=1}^{N} P_{2l} w_l^{-1} \frac{\partial w_l}{\partial p_l}\frac{\partial p_l}{\partial g(p_l)} c_{lk} \right\},$$

with P_{2k} being the kth diagonal element of $\mathbf{P}_2 = \mathbf{T}_2\left(\mathbf{T}_2^\top \Sigma_a^{-1} \mathbf{T}_2\right)^{-1} \mathbf{T}_2^\top \Sigma_a^{-1}$, $\mathbf{T}_2 = \left(\mathbf{Z}^\top\ \mathbf{I}_q\right)^\top$ and c_{lk} the l, k element of $\mathbf{Z}\left(\mathbf{T}_2^\top \mathbf{W}_a \mathbf{T}_2\right)^{-1} \mathbf{Z}^\top$. Some key steps for the derivation of (3) are the following. The estimation $\hat{\boldsymbol{\upsilon}}$ is taken from the score function based on h and after adopting a linearized version of y_k. The estimation for $\boldsymbol{\beta}$ is taken from the score function based on $p_\upsilon(h)$ after evaluating it at the current estimates $\hat{\boldsymbol{\upsilon}}$ and $\hat{\boldsymbol{\delta}}$. Noh and Lee[18] had shown how the estimators provided from these score functions can have the GLM estimating equations form (3).

The approximate covariance matrix of $\hat{\boldsymbol{\beta}}$ is equal to $\left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right)^{-1}$ with $\mathbf{V}^{-1} = \left(\mathbf{W}^{-1} + \mathbf{Z}\Sigma\mathbf{Z}^\top\right)^{-1}$ and takes into account the information that is lost when $\boldsymbol{\upsilon}$ has to be estimated.

The iterative weighted least squares equations for the estimation of $\boldsymbol{\delta}$ in model (2), have the form

$$\Psi_\delta^T \Sigma_\delta^{-1} \Psi_\delta = \Psi_\delta^T \Sigma_\delta^{-1} \mathbf{z}_{d1}, \tag{4}$$

with Σ_δ and $\mathbf{z}_{d1}$ defined as

$$\Sigma_\delta = \text{diag}_{r=1,\cdots,q}\left\{\frac{2}{1-q_{1r}}\right\} \quad \text{and} \quad \mathbf{z}_{d1,r} = \log \lambda_r + \frac{d_r^* - \lambda_r}{\lambda_r}.$$

d_r^* above is equal to $d_r/(1 - q_{1r})$ for the deviance components

$$d_r = 2\int_{\hat{\upsilon}_r}^{0} -\frac{s}{V_1(s)} ds,$$

and $V_1(s) = 1$ for the normal random effect. The quantities q_{1r} are defined according to whether the first or the second order Laplace approximation is used for the approximation of the restricted likelihood needed for the estimation of $\boldsymbol{\delta}$. Studies on the performance of the h-likelihood procedure, like those provided in Yun and Lee [16] and Noh and Lee,[18] have revealed that whenever the sample size within the levels of the random effects is small, the second order Laplace approximation should be used in order to avoid serious biases in the estimators. Noh and Lee[18] provided the explicit form of such an approximation for the type of models described in (1) and (2). Some key steps for the derivation of (3) and (4) can be found in Supporting Information. For a more detailed derivation, the interested reader may refer to Noh and Lee[18] or to the unpublished PhD thesis of the first author (2016).

3 | GOODNESS-OF-FIT STATISTICS

In this section, we provide descriptions of the goodness-of-fit score tests, which measure how well a specific model fits the data within regions defined using either covariate space partitioning or partitioning based on percentiles of estimated probabilities of success. These are extensions of score tests that have been proposed for logistic regression models in the case of independent observations, see other studies,[7-9] or dependent observations through GEE estimation,[6] to random effects models. We also introduce weighted versions of the proposed score statistics that have larger power than the corresponding unweighted ones when detecting specific alternatives.

One suggested method[7] for the assessment of goodness of fit of the fixed part of model (1) considers the model

$$\text{logit}(p_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{o}_{ij}^\top \gamma + \mathbf{z}_{ij}^\top \boldsymbol{\upsilon}. \tag{5}$$

This is an augmented form of model (1), which adopts the additional effect of a G-dimensional design vector $\mathbf{o}_{ij}$ whose entries indicate the group membership of the jth observation in the ith cluster into each one of G nonovelapping groups. Based on the groups defined, testing $\boldsymbol{\gamma} = 0$ is equivalent with testing that there are no significant deviations of the model fixed part to the direction defined by the G groups. Therefore, one could conclude that if $\boldsymbol{\gamma} = 0$, the observations within each group are adequately fitted by the fixed part of the model. However, the whole reasoning has two assumptions, which will hold throughout this study. The first is that the random structure is correctly specified. In the simulation section, the sensitivity of this assumption is studied. The second assumption is that the link function is correctly defined. We use the logit link, but the methods proposed here can be applied to other link functions as well. The assessment of whether the link function is correct or not is not straightforward, since any misspecification will result to a misspecified mean model that contains both fixed and random effects. Therefore, any assessment of the link function adequacy is more or less connected to an overall assessment of the goodness of fit of the mean model. To the best of our knowledge, there is no single test proposed to efficiently assess the goodness of fit of all parts in model (1). For example, tests proposed in Alonso et al,[19,20] which are based on the eigenvalues of the variance-covariance matrices of the fixed-effects parameters estimates, may perform well for random effects misspecification but show low power for detecting misspecification of the fixed part of the model.

A main element of the approaches suggested in this paper is the number of groups G on which the $\boldsymbol{\gamma}$ parameter will be based on. It is reasonable to adopt an automatic approach for their definition. Three main approaches have been suggested in the literature, and these are described below for the random effects modeling setting.

3.1 | Approaches through covariate partitioning

A first approach is to define the groups based on covariates partitions. As in Tsiatis[7] for the ordinary logistic regression setting and Barnhart and Williamson[4] for GEE models, the groups represented by $\mathbf{o}_{ij}$ in (5) are formed by considering G distinct regions in the pth dimensional space of $\mathbf{X}$. To that end, apart from the categorical variables in $\mathbf{X}$, groups are defined for each covariate, and the resulting partition is formed according to all the combinations of all the continuous covariates groups and all the categorical variables categories. In such a way, each observation is placed in the same group with other observations that have similar values in $\mathbf{X}$. Therefore, each row $\mathbf{o}_{ij}^{\top}$ of matrix $\mathbf{O}$ is written as $\mathbf{o}_{ij}^{\top} = (I_{ij,1}, \cdots, I_{ij,G})$, where $I_{ij,g}$ is an indicator variable that equals one if the jth observation in the ith cluster is in the gth region and zero otherwise.

3.2 | Approaches through estimated probabilities

The second approach is based on the expected probabilities of success, as in Hosmer and Lemeshow[8] for the ordinary logistic regression setting and Horton et al[5] for GEE models. Let us denote by $\hat{m}_{ij}$ the expected probability of success for observation j in cluster i based on model (5). The partitioning strategy, according to Hosmer and Lemeshow,[8] is to define cut points $\min\{\hat{m}_{ij}\} = r_0 \leq r_1 \leq \cdots \leq r_{G-1} \leq r_G = \max\{\hat{m}_{ij}\}$, with r_g being the g$\frac{100}{G}$th percentile of $\hat{m}_{ij}$. Then, the gth element of the row vector $\mathbf{o}_{ij}$ is defined as

$$o_{ij,g} = \begin{cases} 1 \ \text{if}\ r_{g-1} < \hat{m}_{ij} \leq r_g \\ 0 \ \text{otherwise} \end{cases} . \tag{6}$$

A complication arises when forming the groups, which accompanies the use of the expected model probabilities of success. Expected probabilities conditional on the value of the random effects represent cluster specific probabilities of success and do not refer to the population from which the cluster comes from. This means that the categorization of observations into groups is affected by the variability of the estimated conditional expected probabilities and as a consequence, observations with the same values of $\mathbf{x}_{ij}^{\top}\hat{\boldsymbol{\beta}}$ may be categorized in different groups. To disregard this issue, one must be certain that all clusters are quite similar reflecting small random effects variances. Unfortunately, this is not known in practice; therefore, in order to test the goodness of fit of the fixed part of the model, it is preferable to use the corresponding marginal expected probabilities of success as described by the null model. These are given by

$$\hat{m}_{ij} = (2\pi)^{-\frac{q}{2}} \int_{\upsilon} \frac{e^{\mathbf{x}_{ij}^{T}\boldsymbol{\beta}+\mathbf{z}_{ij}^{T}\boldsymbol{\upsilon}}}{1+e^{\mathbf{x}_{ij}^{T}\boldsymbol{\beta}+\mathbf{z}_{ij}^{T}\boldsymbol{\upsilon}}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{\upsilon}^{T}\Sigma^{-1}\boldsymbol{\upsilon}} d\upsilon \Big|_{\hat{\beta},\hat{\upsilon},\hat{\Sigma}} . \tag{7}$$

In practice, it is a difficult task to numerically evaluate those integrals; therefore, we suggest the use of an approximation given by Zeger et al[21] of the form

$$\hat{m}_{ij} \approx \hat{m}_{ij,a} = \frac{e^{\hat{a}_{ij}\mathbf{x}_{ij}^{\top}\hat{\boldsymbol{\beta}}}}{1+e^{\hat{a}_{ij}\mathbf{x}_{ij}^{\top}\hat{\boldsymbol{\beta}}}}, \tag{8}$$

where $\hat{a}_{ij} = \left|c^2\hat{\Sigma}\mathbf{z}_{ji}\mathbf{z}_{ij}^{\top} + \mathbf{I}_q\right|^{-\frac{1}{2}} = \left(1 + c^2\mathbf{z}_{ij}^{\top}\hat{\Sigma}\mathbf{z}_{ij}^{\top}\right)^{-\frac{1}{2}}$, with $c = (16\sqrt{3})/(15\pi)$, or the cruder approximation

$$\hat{m}_{ij,b} = \frac{e^{\mathbf{x}_{ji}^{\top}\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_{ij}^{\top}\hat{\boldsymbol{\beta}}}}, \tag{9}$$

valid in the zero limit of the dispersion parameters in Σ.[22]

3.3 | A combined approach

Pulkstenis and Robnison[9] and Evans and Li[6] defined a combined approach that uses both the covariate space and the estimated probabilities for the partitioning in the case of the ordinary logistic regression and the GEE setting, respectively. In the same way, we define the same partitioning approach for the random effects modeling setting. In case we denote by l_r the number of categories of categorical covariate r, then the total number of p_c choices of categories one from each covariate is $M = l_1 \times \cdots \times l_{p_c}$. Each of these combinations forms a group, b, to which the expected probabilities of success can belong. Partitioning, further, each group into two distinct regions according to the median of the expected probabilities of success, med_b, that falls within this group, creates $G = 2 \times M$ groups.

Then the elements of the row vector $\mathbf{o}_{ij}$, for $b = 1, \ldots, M$, are defined by

$$\begin{aligned} &o_{ij,2b-1} = \begin{cases} 1 \text{ if observation } j \text{ in cluster } i \text{ belongs to group b and } \hat{m}_{ij} \leq \text{med}_b \\ 0 \text{ otherwise} \end{cases} \\ &\text{and} \\ &o_{ij,2b} = \begin{cases} 1 \text{ if observation } j \text{ in cluster } i \text{ belongs to group b and } \hat{m}_{ij} > \text{med}_b \\ 0 \text{ otherwise.} \end{cases} \end{aligned} \tag{10}$$

Note that as in section 3.2, the expected values $\hat{m}_{ij}$ are approximated using (8) or (9).

3.4 | The score statistic

The score statistic for testing $\boldsymbol{\gamma} = 0$ in model (5) has the form

$$\mathbf{r}^{\top}\mathbf{D}^{-}\mathbf{r}, \tag{11}$$

where $\mathbf{r}$ is the score vector for the estimation of $\boldsymbol{\gamma}$ parameters evaluated at the null hypothesis $\boldsymbol{\gamma} = \mathbf{0}$, and $\mathbf{D}$ is its asymptotic covariance matrix, with $\mathbf{D}^{-}$ being a generalized inverse of $\mathbf{D}$. Both $\mathbf{r}$ and $\mathbf{D}$ are evaluated at $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\upsilon}}$, and $\hat{\boldsymbol{\delta}}$ from the null model.

The form of the $\mathbf{r}$ vector can be derived from the estimating Equation 3. In Appendix A, result R1 shows that $\mathbf{r}$ is a weighted sum of residuals of the form $\mathbf{r} = \mathbf{O}^{\top}(\mathbf{y}^{*} - \mathbf{p})$, where $\mathbf{y}^{*} = \mathbf{y} - \mathbf{s}$, and $\mathbf{O}$ a $N \times G$ design matrix for the group effects. The covariance matrix of the score vector is derived from the Hessian matrix of $p_{\upsilon}(h)$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Based on model (1), the expression for $\mathbf{D}$ takes the form (see Lee and Nelder[11,12] or the unpublished PhD thesis of the first author, 2016),

$$\mathbf{D} = \mathbf{O}^{\top}\mathbf{V}^{-1}\mathbf{O} - \mathbf{O}^{\top}\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{V}^{-1}\mathbf{O},$$

where $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{Z}\Sigma\mathbf{Z}^{\top}$ and $\mathbf{V}^{-1} = \mathbf{W} - \text{WZ}\left(\mathbf{Z}^{\top}\text{WZ} + \Sigma^{-1}\right)^{-1}\mathbf{Z}^{\top}\mathbf{W}$, assuming that the inverse matrix of $\mathbf{V}$ exists. Based on the properties of the h-likelihood estimators, under the null hypothesis $\boldsymbol{\gamma} = 0$, the score vector $\mathbf{r}$ is distributed as a multivariate normal variable with mean $\mathbf{0}$ and asymptotic covariance matrix $\mathbf{D}$. Therefore, under the null hypothesis, the score statistic (11) is asymptotically distributed as χ^2_{df} with df = rank ($\mathbf{D}^{-}$).

3.5 | A weighted score statistic

There are cases, however, where a particular partitioning method corresponds to a score test with low power. Studies that examined the performance of the most well-known goodness-of-fit statistics had pointed out specific alternatives, for example, an omitted interaction term, that cannot be easily detected, under situations where the sample size is small or the departures from the null model are not extreme. This is the case where an omitted covariate causes many residuals in (11) to be large. Hosmer and Hjort,[10] within the context of goodness-of-fit tests for ordinary logistic regression, noted that incorporating information about possible omitted covariates into goodness-of-fit tests may increase the power of the tests. In the same line of research, we provide a weighted version of the proposed score statistic in (11), which can be of great help in the detection of a particular missing covariate.

A closer look to the score $\mathbf{r}$ shows that its gth element is

$$\boldsymbol{r}_g = \sum_{i=1}^{n}\sum_{j=1}^{n_i} o_{ij,g}\left(y^{*}_{ij} - \hat{p}_{ij}\right) \qquad g = 1, \cdots G. \tag{12}$$

The weights are based on the notion of partial correlation between the dependent variable and the $N \times 1$ missing covariate $\mathbf{d}$ after taking into account the effect of covariates in $\mathbf{X}$. The residuals of the regression of $\mathbf{d}$ on the set of covariates $\mathbf{X}$ are $\mathbf{e} = (\mathbf{I}_N - \mathbf{H})\mathbf{d}$, where $\mathbf{H}$ is the weighted projection matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\hat{\mathbf{V}}^{-1}$. The correlation between these residuals and the residuals from model (1) is expected to be high, the higher the true effect of $\mathbf{d}$ is. It is reasonable then to define the weighted version of score (12) as

$$\mathbf{R}_{w,g} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} o_{ij,g}\left(d_{ij} - \mathbf{x}_{ij}^\top\left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{d}\right)\left(y_{ij}^* - \hat{p}_{ij}\right) \quad g = 1, \cdots G, \tag{13}$$

and in matrix form as $\mathbf{R}_w = \mathbf{K}^\top(\mathbf{y}^* - \hat{\mathbf{p}})$, where the gth column of matrix $\mathbf{K}$ has the form

$$K_{ij,g} = \begin{cases} o_{ij,g}\left(d_{ij} - \mathbf{x}_{ij}^\top\left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{d}\right) & \text{if observation } j \text{ in cluster } i \\ & \text{belongs to the } g\text{th partition} \\ 0 & \text{otherwise} \end{cases} .$$

The new score statistic is now

$$\mathbf{R}_w^\top\mathbf{D}_w^-\mathbf{R}_w, \tag{14}$$

where

$$\mathbf{D}_w = \mathbf{K}^\top\mathbf{V}^{-1}\mathbf{K} - \mathbf{K}^\top\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{K}.$$

Based on the same arguments as those provided for the asymptotic distribution of the statistic (11), the asymptotic distribution of the statistic (14) under the null hypothesis that the model (1) is correct is χ^2_{df} with df = rank $\left(\mathbf{D}_w^-\right)$.

If the structure of the random effects in (1) is correctly defined, the statistic (14) could be used for testing whether any particular lack of fit of the null model (1) could be attributed to an omission of the fixed effect of the covariate $\mathbf{d}$. The intuition behind the test is that if the omitted covariate is important then rejection of the null hypothesis could be attributed to that omitted covariate. The form of $\mathbf{d}$ could be, for example, a quadratic term of a continuous covariate, an omitted interaction term between a continuous and a categorical/continuous covariate, or could also be a continuous covariate that is not included in the null model. If we consider the quantities $y_{ij}^* - \hat{p}_{ij}$ in $\mathbf{R}$ as the residuals for the adjusted responses y_{ij}^*, then, under the hypothesis that the random structure in (1) is correct, possible large values of the statistic (14) could be associated with the presence of a sufficiently large number of observations with large residuals under model (1), and at the same time, the values of the omitted covariate for these observations are not well approximated by the covariates $\mathbf{X}$ in the null model.

4 | SIMULATION STUDY

We conducted two simulation studies. The first aimed at exploring the effect of using the ordered values of estimated expected probabilities $\hat{m}_{ij,a}$ or $\hat{m}_{ij,b}$ in the formation of groups, instead of using the estimated expected values $\hat{m}_{ij}$. The second simulation study aimed at evaluating the performance of the goodness-of-fit tests proposed in the previous section.

4.1 | Estimated expected probabilities performance

For the comparison between using (8) or (9) instead of (7) in the evaluation of approaches through estimated probabilities, we considered three models for data generation and five repeated measures for each subject. The first model (model 1a) was

$$\text{logit}\left(p_{ij}\right) = 0.5 - 0.8X_{i,1} + 0.8X_{i,2} + \upsilon_{i,1} + \upsilon_{i,2}X_{i,1},$$

where $X_{i,1} \sim \text{Uniform}(-1, 1)$, $X_{i,2} \sim \text{Bernoulli}(0.5)$, and for two different numbers of subjects giving $i = 1, \ldots, 300$, or $i = 1, \ldots, 50$.

$$\begin{pmatrix}\upsilon_{i,1} \\ \upsilon_{i,2}\end{pmatrix} \sim N\left(0, \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right).$$

The second model (model 1b) was

$$\text{logit}\left(p_{kij}\right) = 0.5 - 0.8X_{kij,1} + 0.8X_{ki,2} + \upsilon_{k,1} + \upsilon_{ki,2},$$

with $X_{kij,1} \sim \text{Uniform}(-1, 1)$, $X_{kij,2} \sim \text{Bernoulli}(0.5)$, $k = 1, \dots, 6$, $i = 1, \dots, 50$, and the random effects $\upsilon_{k,1} \sim N(0, 1)$, $\upsilon_{ki,2} \sim N(0, 2)$ as independent random variables. The third model was

$$\text{logit}\left(p_{ij}\right) = 0.5 + 0.8X_{ij,1} - 0.8X_{i,2} + \upsilon_i,$$

with $X_{ij,1} \sim \text{Uniform}(-1, 1)$, $X_{i,2} \sim \text{Bernoulli}(0.5)$, again for two different numbers of subjects, giving $i = 1, \dots, 400$ or $i = 1, \dots, 50$ and $\upsilon_i \sim N(0, 2)$. The first model is an example of a model with between subjects covariates and two correlated random effects. The second is an example of a multilevel model with three levels, two covariates, one in the first level and one in the second level, and two random effects, one in the second level and one in the third. The third model is an example of a random intercept model with two covariates, one between subjects and one within.

We simulated 100 data sets using each of the models defined above. For each data set, the model used for data generation was also used for data analysis. We applied the partitioning strategy through estimated probabilities with $G = 10$, using (7), (8), and (9). Two 10 × 10 association tables were generated for each data set and each model. The first table cross classified each observation according to the group that was assigned when its expected probability of success was calculated using $\hat{m}_{ij,a}$ and $\hat{m}_{ij}$. The second table shows the same information for $\hat{m}_{ij,b}$ and $\hat{m}_{ij}$. If the approximations $\hat{m}_{ij,a}$ and $\hat{m}_{ij,b}$ were equivalent to $\hat{m}_{ij}$, then nonzero frequencies would appear only in the main diagonal. The frequencies in the first subdiagonal below the main one represent observations with expected probability of success calculated using $\hat{m}_{ij,a}$, which were placed in the previous group from which they should have been placed had the expected probabilities of success $\hat{m}_{ij}$ been used. Similar interpretations for the other subdiagonals show that any deviation from the main diagonal corresponds to deviations between the use of the marginal expected probabilities $\hat{m}_{ij}$ and the use of $\hat{m}_{ij,a}$ or $\hat{m}_{ij,b}$. The proportion of observations falling in the diagonal and each of the subdiagonals of these two association tables was computed for each of 100 simulations. The proportion of observations that fall in the subdiagonals of each of these two association tables would be a measure of misplacement caused by the use of $\hat{m}_{ij,a}$ or $\hat{m}_{ij,b}$, instead of using $\hat{m}_{ij}$ for the grouping of the observations. Table 1 displays the mean proportions over the 100 simulations for $\hat{m}_{ij,a}$ and $\hat{m}_{ij,b}$ for each model. The 0 columns represent classifications in the main diagonal. Negative numbers represent classifications in the subdiagonals below the main one, while positive numbers represent classifications to the opposite direction. It can be seen that almost all cases fall in the main diagonal for models 1b and 1c, even for a small sample size. For the complex model 1a, however, there is some loss of information with the use of $\hat{m}_{ij,a}$ and $\hat{m}_{ij,b}$, which becomes more serious with a small sample size, reaching around 30% of observations misplaced, when $\hat{m}_{ij,b}$ is used.

4.2 | Score tests performance

In the second study, the performance of the proposed score tests was examined. The aim was to study the effect of the number of clusters and the type of omitted covariates on the performance of the score tests. We examined situations where the random fitted part was correctly specified, and we compared the performance of the proposed tests with the cumulative sums tests proposed in Pan and Lin.[3] An important part of this study was the behavior of the weighted tests under an incorrect guess for the fixed part specification. Finally, we considered situations when an incorrect random part was fitted to the model, and we examined the impact of the number of groups formed, on the performance of the proposed tests.

TABLE 1 Mean proportions of placement into groups, over 100 simulations, showing the degree of deviation from grouping using $\hat{m}_{ij}$ with the grouping using $\hat{m}_{ij,a}$ or $\hat{m}_{ij,b}$

		Deviation from Grouping Using $\hat{m}_{ij}$									
		$\hat{m}_{ij,a}$					$\hat{m}_{ij,b}$				
	n	**≤ −2**	**−1**	**0**	**1**	**≥ 2**	**≤ −2**	**−1**	**0**	**1**	**≥ 2**
Model 1a	50	7.0	3.7	83.0	1.8	4.5	9.1	7.5	71.1	5.5	6.7
	300	4.0	1.8	91.2	1.02	2.0	4.0	7.5	79.5	6.9	2.0
Model 1b	50	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
Model 1c	50	0.02	0.01	99.93	0.04	0.0	0.02	0.01	99.93	0.04	0.0
	400	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0

4.2.1 | Performance under a correctly specified random part

The tests were designed so as to detect misspecifications of the fixed part in model (1) and assume a correctly defined random part. Under this assumption, to assess the performance of the tests, two different numbers of clusters ($n = 100$ and $n = 400$) were used for data generation using a random intercept logit normal model with $\upsilon_i \sim N\left(0, \sigma^2\right)$. Five different models were considered for data generation, where subscript $j = 1, \ldots, 5$, denotes the observations within each cluster:

- Model 2a: $\text{logit}\left(p_{ij}\right) = 0.5 - 0.8X_{i,1} + 0.8X_{i,2} + \beta_3 X_{i,1}X_{i,2} + \upsilon_i$, where $X_{i,1}$ was generated as Bernoulli(0.5) and $X_{i,2}$ as Uniform $(-1,1)$.
- Model 2b: $\text{logit}\left(p_{ij}\right) = 0.5 - 0.8X_{i,1} + 0.8X_{ij,2} + \beta_3 X_{i,1}X_{ij,2} + \upsilon_i$, where both $X_{i,1}$ and $X_{ij,2}$ were generated as Uniform$(-1,1)$.
- Model 2c: $\text{logit}\left(p_{ij}\right) = 0.5 - 0.8X_{i,1} + \beta_2 X_{i,1}^2 + \upsilon_i$, where $X_{i,1}$ was generated as Uniform$(-1,1)$.
- Model 2d: $\text{logit}\left(p_{ij}\right) = 0.5 - 0.8X_{i,1} + \beta_2 X_{i,2} + \upsilon_i$, where $X_{i,1}$ and $X_{i,2}$ were generated as Normal(0,1) and Uniform$(-1,1)$, respectively.
- Model 2e: $\text{logit}\left(p_{ij}\right) = 0.5 - 0.8X_{i,1} + 0.8\log\left(X_{ij,2}\right) + \upsilon_i$, where $X_{i,1}$ was generated as Bernoulli(0.5) and $X_{ij,2}$ as lognormal(0,1).

Model 2a was considered for checking the performance of score tests in testing a dichotomous-continuous interaction, while Model 2b was considered for checking continuous-continuous interaction. Model 2d was considered for checking the linear effect of an additional covariate. Finally, models 2c and 2e were used for testing the functional form of a covariate. Data were generated from models 2a and 2b, using $\beta_3 = 0, 0.8, 1.6$ and were fitted using $\beta_3 = 0$. Data were generated from models 2c and 2d, using $\beta_2 = 0, 0.8, 1.6$ and were fitted using $\beta_2 = 0$. The data generated from model 2e were fitted using $\text{logit}\left(p_{ij}\right) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{ij,2} + \upsilon_i$. Five hundred data sets were generated for each model and each simulation scheme. Estimation of the unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\upsilon}$, and $\boldsymbol{\delta}$ was based on the iterative weighted least squares Equations 3 and 4.

The results are shown in Tables 2 and 3. The score tests were denoted using a subscript that discriminates between different partitioning approaches. A star superscript denotes the weighted version of the score. Therefore, the score tests S_T and S_T^* are the unweighted and the weighted score statistics defined in Section 3 and are based on the following partitions of the covariate space. For model 2a, four groups were formed according to the possible combinations of the indicators $I(x_{i,1} = 1)$ and $I(x_{i,2} > 0)$. For model 2b, four groups were formed according to all possible combinations of the indicators $I(x_{i,1} > 0)$ and $I(x_{ij,2} > 0)$. For model 2c, four groups were formed according to the $x_{i,1}$ partition set values $\{-\infty, -0.5, 0, 0.5, \infty\}$. For model 2d, four groups were formed according to the $x_{i,1}$ partition set values $\{-\infty, -0.6, 0, 0.6, \infty\}$. Finally, for model 2e, six groups were formed according to the possible combinations of the indicator $I(x_{i,1} = 1)$ and the $x_{ij,2}$ partition set values $\{-\infty, q_1, q_2, \infty\}$, with $q_1$andq_2 being the first and the third quartiles of x_2, respectively. The score tests S_{HL} and S_{HL}^* are defined in section 3.2, and the elements of matrix $\mathbf{O}$ are computed as described in (6) with $G = 5$ and $\hat{r}_1$ being the 20% percentile of (9), $\hat{r}_2$ is the 40% percentile of (9), and so on.

The score statistics S_{PR} and S_{PR}^* are based on the grouping strategy described in section 3.3 and are provided only for models 2a and 2e, which include both categorical and continuous covariates in the fixed part of their linear predictors. For these tests, the elements of matrix $\mathbf{O}$ were computed as described in (10), where the observations were initially partitioned into two groups according to the indicator x_1 and, at a second step, were partitioned into two further groups according to the median of (9). The $\mathbf{d}$s for the computation of the score statistics S_{HL}^*, S_T^*, and S_{PR}^* were $d_{ij} = x_{i,1} * x_{i,2}$ for model 2a, $d_{ij} = x_{i,1} * x_{ij,2}$ for model 2b, $d_{ij} = x_{i,1}^2$ for model 2c, $d_{ij} = x_{i,2}$ for model 2d, and $d_{ij} = \log\left(x_{ij,2}\right)$ for model 2e, that is, whenever a guess had to be made for the evaluation of a weighted score, this was a correct guess.

Table 2 presents empirical sizes and powers of the tests based on 500 replications of each simulated scenario for models 2a to 2d. Table 3 provides the empirical power of the tests for scenario 2e. The results indicate that the weighted statistics have better performance compared with their unweighted counterparts for all the models considered and each size of n. Furthermore, it seems that the statistics S_T and S_T^* perform better than S_{HL} and S_{HL}^*, respectively. Moreover, the statistics S_{PR} and S_{PR}^* have an equivalent performance with that of the statistics S_T and S_T^* for model 2a, whereas for model 2e, S_{PR}^* performs better than S_T^*, and S_T performs better than S_{PR}.

Nonreported results (unpublished PhD thesis of the first author, 2016) indicate that as n and/or m get larger, the power of the tests increases, whereas when σ^2 gets larger—given n and m—the power of the tests decreases. Lastly, as regards

TABLE 2 Empirical type I error rates and empirical powers (using $\alpha = 0.05$) of the score tests, the weighted score tests, and the supremum statistics for models 2a to 2d

	n	Test	$\beta_3 = 0$	$\beta_3 = 0.8$	$\beta_3 = 1.6$
Model 2a	100	$S_{HL}(S^*_{HL})$	0.064 (0.056)	0.110 (0.134)	0.168 (0.396)
		$S_T(S^*_T)$	0.052 (0.040)	0.206 (0.182)	0.404 (0.504)
		S_{PR} (S^*_{PR})	0.048 (0.040)	0.172 (0.174)	0.420 (0.504)
		S_x	0.082	0.128	0.318
		S_g	0.062	0.110	0.252
	400	$S_{HL}(S^*_{HL})$	0.052 (0.050)	0.246 (0.452)	0.634 (0.978)
		$S_T(S^*_T)$	0.054 (0.054)	0.476 (0.548)	0.956 (0.992)
		S_{PR} (S^*_{PR})	0.040 (0.054)	0.478 (0.552)	0.962 (0.992)
		S_x	0.060	0.420	0.922
		S_g	0.058	0.348	0.724
Model 2b	100	$S_{HL}(S^*_{HL})$	0.042 (0.028)	0.148 (0.304)	0.550 (0.930)
		$S_T(S^*_T)$	0.038 (0.032)	0.238 (0.346)	0.728 (0.944)
		S_x	0.044	0.104	0.328
		S_g	0.052	0.154	0.474
	400	$S_{HL}(S^*_{HL})$	0.018 (0.038)	0.718 (0.952)	1.0 (1.0)
		$S_T(S^*_T)$	0.026 (0.038)	0.828 (0.970)	1.0 (1.0)
		S_x	0.060	0.402	0.966
		S_g	0.070	0.614	0.994
			$\beta_2 = 0$	$\beta_2 = 1.8$	$\beta_2 = 1.6$
Model 2c	100	$S_{HL}(S^*_{HL})$	0.040 (0.050)	0.120 (0.100)	0.440 (0.460)
		$S_T(S^*_T)$	0.052 (0.050)	0.106 (0.114)	0.470 (0.530)
		S_x	0.040	0.200	0.612
		S_g	0.048	0.174	0.522
	400	$S_{HL}(S^*_{HL})$	0.040 (0.054)	0.412 (0.454)	0.970 (0.986)
		$S_T(S^*_T)$	0.058 (0.064)	0.404 (0.492)	0.966 (0.990)
		S_x	0.064	0.678	0.998
		S_g	0.054	0.622	0.998
Model 2d	100	$S_{HL}(S^*_{HL})$	0.034 (0.066)	0.042 (0.456)	0.028 (0.982)
		$S_T(S^*_T)$	0.040 (0.072)	0.054 (0.468)	0.034 (0.986)
		S_x	0.056	0.038	0.052
		S_g	0.044	0.050	0.030
	400	$S_{HL}(S^*_{HL})$	0.038 (0.058)	0.028 (0.976)	0.102 (1.0)
		$S_T(S^*_T)$	0.054 (0.054)	0.028 (0.986)	0.058 (1.0)
		S_x	0.076	0.064	0.142
		S_g	0.066	0.046	0.098

TABLE 3 Empirical powers (using $\alpha = 0.05$) of the score tests, the weighted score tests, and the supremum statistics for the model 2e

Test	$n = 100$	$n = 400$
$S_{HL}(S^*_{HL})$	0.484 (0.780)	0.992 (1.0)
S_T (S^*_T)	0.586 (0.760)	0.998 (1.0)
S_{PR} (S^*_{PR})	0.478 (0.822)	0.990 (1.0)
S_x	0.626	1.0
S_g	0.328	0.932

the performance of the tests according to the type of model misspecification, lower powers were detected when the continuous-dichotomous interaction was omitted and further, the unweighted tests in model 2d cannot detect the omitted continuous main effect.

Tables 2 and 3 also report the results of the supremum statistics defined in Pan and Lin[3] for the same simulation schemes. Specifically, $S_x = \sup_{\mathbf{x}} |W(\mathbf{x})|$, $S_g = \sup_r |W_g(r)|$ for the processes

$$W(\mathbf{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n_i} I\left(\mathbf{x_{ij}} \leq \mathbf{x}\right) \hat{\mathrm{e}}_{ij}, \quad \mathbf{x} \in \mathbb{R}^p$$

$$W_g(r) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n_i} I\left(\hat{m}_{ij} \leq r\right) \hat{\mathrm{e}}_{ij}, \quad r \in \mathbb{R}$$

with $\hat{\mathrm{e}}_{ij} = y_{ij} - \hat{m}_{ij}$, $\hat{m}_{ij}$ as defined in (7) and

$$I\left(\mathbf{x}_{ij} \leq \mathbf{x}\right) = \begin{cases} 1 \text{ if } x_{ij,1} \leq x_1, \cdots, x_{ij,p} \leq x_p \\ 0 \quad \text{otherwise} \end{cases},$$

for the first process and analogously defined for the second one.

The statistic S_x in the tables refers to the process that is formed based on the successfully larger partitions of the joint space of covariates in the null design matrix, whereas S_g refers to the same statistic when using the expected values of the model fitted. Comparing these results with those given for the proposed unweighted score tests, we can notice the better performance of the tests S_T, S_{PR} for the missing interaction schemes (models 2a and 2b), and the better performance of S_x for the misspecified functional form schemes (models 2c and 2e). The weighted score tests have better performance as compared with their unweighted analogous, most of the times. In general, they displayed better performance as compared with the supremum statistics except from the missing quadratic terms settings, where the supremum statistics performed the best.

4.2.2 | Performance under an incorrect guess of a covariate

Unweighted score tests or tests based on supremum statistics provide information about the model's goodness of fit. In contrast, weighted score tests attempt to provide information about the fit of a possible model effect. Therefore, weighted score tests are based on a guess about the possible model specification. Note that the results of Tables 2 and 3 are based on the assumption that a correct guess was made. To identify what is the effect of an incorrect proposed specification, we incorrectly assumed a missing quadratic effect, for models 2b and 2d when these were fitted to the data assuming $\beta_3 = 0$ and $\beta_2 = 0$, respectively. In these situations, the weighted tests were formed with $d_{ij} = x_{ij,2}^2$ for model 2b and $d_{ij} = x_{i,1}^2$ for model 2d. Additionally, we considered the following scheme:

- Model 2f: logit $\left(p_{ij}\right) = 0.5 - 0.8X_{i,1} + 0.8X_{i,2} + \beta_3 X_{ij,3} + \upsilon_i$, where $X_{i,1}$ was generated as Normal(0,1), $X_{i,2}$ as Bernoulli(0.5), and $X_{ij,3}$ as Uniform(0,1),

where $\upsilon_i \sim N\left(0, \sigma^2\right)$. For model 2f, groups were formed according to the possible combinations of $I(x_{i,2} = 1)$ and $I(x_{i,1} > 0)$. The data generated were fitted using $\beta_3 = 0$. For this model, we incorrectly assumed $d_{ij} = x_{i,1}x_{i,2}$.

The results for the evaluation of the weighted score tests are given in Table 4. Note that S_{PR}^* can be evaluated for model 2f only. In general, the power of the tests is larger when the coefficient of the true missing effect is larger. However, all the estimated powers are small apart that of model 2b for a large cluster size. One can think of this as a reasonable result since in principle, the weighted tests should not direct to alternatives other than the true ones. For the case of model 2d and for the large sample size, the large power can be explained by the fact that the weights ($d_{ij} = x_{ij,2}^2$) are formed by a covariate, which is included in the true missing interaction term.

4.2.3 | Performance under an incorrectly specified random part

In this subsection, the behavior of the proposed score tests under a misspecification of the model random part is studied. The following models encountering different types of random structure were used for data generation:

- Model 3a: logit $\left(p_{ij}\right) = 0.5 - 0.8\,(j-3) + 0.8X_{i,1} + \beta_3 X_{ij,2} + \upsilon_{i,1} + \upsilon_{i,2}\,(j-3)$ with $X_{i,1}$ a Bernoulli(0.5), $X_{ij,2} = (j-3)^2$ or $X_{ij,2} = X_{i,1}\,(j-3)$ and $\left(\upsilon_{i,1}, \upsilon_{i,2}\right) \sim N\left(\mathbf{0}, \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$, $i = 1, \ldots, 300$, $j = 1, \ldots, 5$
- Model 3b: logit $\left(p_{ijk}\right) = 0.5 - 0.8X_{ki,1} + 0.8X_{ki,2} + \beta_3 X_{ki,3} + \upsilon_{k,1} + \upsilon_{ki,2}$ with $X_{ki,1}$ Bernoulli(0.5), $X_{ki,2}$ Uniform(−1,1), $X_{ki,3} = X_{ki,2}^2$ or $X_{ki,3} = X_{ki,1}X_{ki,2}$ and $\upsilon_k \sim N(0,1)$, $\upsilon_{ki} \sim N(0,1.5)$, $k = 1,2,3$, $i = 1, \ldots, 100$, $j = 1, \ldots, 5$.

The first model is an example with one random intercept and a (correlated) random slope, while the second is an example of a multilevel model. We examined data generation using two different specifications of the fixed part in both models. The first used the quadratic effect $X_{ij,2} = (j - 3)^2$ (called quadratic effect version of model 3a), and the second

TABLE 4 Empirical powers (using $\alpha = 0.05$) of the weighted score tests for models 2b, 2d, and 2f with an incorrect guess for model specification

	n	Test	$\beta_3 = 0.8$	$\beta_3 = 1.6$
Model 2b	100	S^*_{HL}	0.078	0.204
		S^*_T	0.086	0.194
	400	S^*_{HL}	0.358	0.918
		S^*_T	0.260	0.794
Model 2d	100	S^*_{HL}	0.052	0.011
		S^*_T	0.038	0.022
	400	S^*_{HL}	0.036	0.104
		S^*_T	0.046	0.058
Model 2f	100	S^*_{HL}	0.050	0.046
		S^*_T	0.050	0.040
		S^*_{PR}	0.046	0.036
	400	S^*_{HL}	0.080	0.084
		S^*_T	0.058	0.074
		S^*_{PR}	0.058	0.070

TABLE 5 Empirical type I error rates and powers (using $\alpha = 0.05$) of the score and the supremum tests for each of the four versions of models 3a and 3b

	Test	Quadratic Effect Version[a]		Interaction Effect Version[b]	
β_3		0	0.8	0	0.8
Model 3a	S_{HL}	0.148	1.0	0.148	0.910
	S^*_{HL}	0.134	1.0	0.114	0.974
	S_T	0.188	1.0	0.188	0.964
	S^*_T	0.122	1.0	0.074	0.994
	S_{PR}	0.094	1.0	0.094	0.984
	S^*_{PR}	0.138	1.0	0.102	0.994
	S_x	0.102	1.0	0.102	0.984
	S_g	0.114	1.0	0.114	0.966
Model 3b	S_{HL}	0.06	0.140	0.06	0.096
	S^*_{HL}	0.074	0.174	0.064	0.162
	S_T	0.076	0.154	0.076	0.154
	S^*_T	0.072	0.194	0.064	0.154
	S_{PR}	0.042	0.050	0.042	0.342
	S^*_{PR}	0.088	0.266	0.100	0.220
	S_x	0.130	0.074	0.130	0.070
	S_g	0.062	0.078	0.062	0.164

[a] Quadratic effect version: Data generated using $(j-3)^2$ and $X^2_{ki,2}$ for models 3a and 3b, respectively.

[b] Interaction effect version: Data generated using $X_{i,1}\,(j-3)$ and $X_{ki,1}X_{ki,2}$ for models 3a and 3b, respectively.

used the interaction effect $X_{ij,2} = X_{i,1}(j-3)$ (called interaction effect version of model 3a) . A similar specification was adopted for $X_{ki,3}$ in model 3b. Two different values for β_3 ($\beta_3 = 0$ and $\beta_3 = 0.8$), finally formed four versions for each of models 3a and 3b for data generation. Random intercept models with a single random effect v_i, distributed as $N\left(0, \sigma^2\right)$, were fitted to all the data sets generated, assuming that $\beta_3 = 0$.

Table 5 presents the results concerning the performance of the unweighted and the weighted score tests. For the latter, it was assumed that a correct guess was made about the missing effect. For both models, S_{HL} and S^*_{HL} were computed using $G = 10$. As regards statistics S_T and S^*_T, they were based on the partitions provided by combinations of the indicator $I(x_{i,1} = 1)$ and the $j-3$ partition set $\{-\infty, -1, 0, 1, \infty\}$ for model 3a and for model 3b, combinations of the indicator $I(x_{ki,1} = 0)$ and the $x_{ki,2}$ partition set $\{-\infty, -0.5, 0, 0.5, \infty\}$.

Table 5 provides the empirical type I error rates (in columns which correspond to $\beta_3 = 0$) and powers (in columns which correspond to $\beta_3 = 0.8$) for the proposed tests, as well as for the supremum tests S_x, S_g, based on 500 replications of each simulated scenario. We observed inflated empirical type I error rates for all the tests for model 3a and some of the tests for model 3b. Because of these biases, we empirically adjusted the powers of all tests, and the results are shown in columns labeled $\beta_3 = 0.8$. The idea was first to estimate, using linear interpolation, the nominal significance level α for which the simulated type I error rate of each test is equal to 0.05. For that purpose, we obtained the empirical type I error rates, through simulation, when $\beta_3 = 0$ at the 5% and 2% nominal significance levels. Let us denote these numbers by p_1 and p_2. We obtained the required nominal significance level α for which the empirical type I error rate would be 5% from $(\alpha - 0.02)/(0.05 - p_1) = (0.05 - 0.02)/(p_2 - p_1)$. We then used that nominal significance level to compute the empirical powers when $\beta_3 = 0.8$.

As a general comment, we note that all tests perform equally well for model 3a. For model 3b, we can, in general, observe slightly larger empirical power values when detecting an omitted quadratic effect as compared with those when detecting an interaction effect. The weighted statistics seem to perform better than their unweighted counterparts or the supremum tests. Nevertheless, the inflated type I error rates for the tests in the majority of the situations indicate that a test for the correct specification of the random part may be useful before proceeding to the examination of the fixed part of the model. Such tests are proposed in Alonso et al.[19,20]

4.2.4 | Sensitivity based on the number of groups

The number of groups G, specified in the calculation of the score tests S_T, S_T^*, S_{HL}, and S_{HL}^*, may have a critical role in the acceptance or rejection of a null hypothesis. S_{PR} and S_{PR}^* tests use always a specific number of groups as this is defined by the number of categorical or categorized covariates in $\mathbf{X}$. Therefore, these tests are not affected by any relevant subjective choice. To investigate the effect of changing the number of groups in the performance of the score tests, we performed a small simulation study evaluating the score tests S_T, S_T^*, S_{HL}, and S_{HL}^* using different number of groups. We denote by $S_{.,G}$ or $S_{.,G}^*$ a specific score test, which is based on G groups. $S_{T,4}^*$ and $S_{HL,5}^*$ were both based on partitions described in section 4.2.1 and are provided for comparison. For model 2a, $S_{T,8}^*$ was formed according to the combinations of the indicator $I(x_{i,1} = 1)$ and the partition set values for $x_{i,2}$ equal to $\{-\infty, -0.5, 0, 0.5, \infty\}$. For model 2b, $S_{T,9}^*$ was formed according to the combinations of the $x_{i,1}$ partition set values $\{-\infty, -0.3, 0.3, \infty\}$ and the $x_{ij,2}$ partition set values $\{-\infty, -0.3, 0.3, \infty\}$. For model 2d, $S_{T,6}^*$ was formed according to the $x_{i,1}$ partition set values $\{-\infty, -1, -0.5, 0, 0.5, 1, \infty\}$.

Table 6 presents the results for three sample sizes (50, 100, and 400) based on 500 simulations. Type I error rates do not seem to be affected. As regards the power, increasing the number of groups lowers the empirical power. However, differences are not acute. These differences are a bit larger for the weighted score tests. One possible explanation for this phenomenon is that when the number of groups grow, there are not enough observations in each group corresponding to success and failure. Therefore, the validity of the chi-square distribution of the score tests is questionable and recommendations as in Barnhart and Williamson[4] might be in order. These include changing the number of partition regions so that each region should contain at least 10 individuals, only 25% of the total number of regions should have less than 25 individuals, and finally, all of the regions should have greater than zero frequency for both binary responses.

5 | EXAMPLE

In this section, the proposed statistics are illustrated with the use of a real data set. This is the Indonesian study on respiratory infection.[23] Goodness-of-fit tests to fitted random effect models for this data set were implemented in Pan and Lin.[3] We adopt the same random effect modeling as in Pan and Lin,[3] in order to check the performance of the proposed, in the previous sections, goodness-of-fit tests.

According to the description of the study, 275 preschool children were examined up to six consecutive quarters for the presence of respiratory infection (presence = 1, absence = 0). The information on respiratory infection is not complete for all the 275 children. A total number of 1200 observations is available. Here, we make the assumption that missing data are all Missing Completely at Random (MCAR). The covariates used for the present analysis include x_1, which is the gender indicator (male = 0, female = 1); x_2, which is the height of the child as a percentage of the National Center of Health Statistics Standards height according to his/her age (centered at 90%) and indicates longer term nutritional status; x_3, which indicates the presence (=1) or absence (=0) of xerophthalmia, an ocular manifestation of chronic vitamim A deficiency; x_4, the age of the child in months (centered at 36 months); and x_5 and x_6, the cosin and sin terms of the annual cycle in order to adjust for seasonality.

TABLE 6 Empirical type I error rates and power of the tests (using $\alpha = 0.05$) under different number of groups for models 2a, 2b, and 2d

	n	Test	$\beta_3 = 0$	$\beta_3 = 0.8$	$\beta_3 = 1.6$
Model 2a	50	$S_{HL,10}(S^*_{HL,10})$	0.016 (0.04)	0.042 (0.058)	0.056 (0.096)
		$S_{HL,5}(S^*_{HL,5})$	0.044 (0.038)	0.06 (0.056)	0.05 (0.18)
		$S_{T,8}(S^*_{T,8})$	0.044 (0.034)	0.06 (0.066)	0.144 (0.136)
		$S_{T,4}(S^*_{T,4})$	0.044 (0.040)	0.078 (0.084)	0.172 (0.232)
	100	$S_{HL,10}(S^*_{HL,10})$	0.050 (0.058)	0.082 (0.092)	0.158 (0.274)
		$S_{HL,5}(S^*_{HL,5})$	0.064 (0.056)	0.110 (0.134)	0.168 (0.396)
		$S_{T,8}(S^*_{T,8})$	0.040 (0.046)	0.140 (0.142)	0.358 (0.348)
		$S_{T,4}(S^*_{T,4})$	0.052 (0.040)	0.206 (0.182)	0.404 (0.504)
	400	$S_{HL,10}(S^*_{HL,10})$	0.066 (0.060)	0.230 (0.360)	0.644 (0.936)
		$S_{HL,5}(S^*_{HL,5})$	0.052 (0.050)	0.246 (0.452)	0.634 (0.978)
		$S_{T,8}(S^*_{T,8})$	0.058 (0.036)	0.400 (0.418)	0.944 (0.946)
		$S_{T,4}(S^*_{T,4})$	0.054 (0.054)	0.476 (0.548)	0.956 (0.992)
Model 2b	50	$S_{HL,10}(S^*_{HL,10})$	0.032 (0.022)	0.072 (0.088)	0.298 (0.5)
		$S_{HL,5}(S^*_{HL,5})$	0.036 (0.028)	0.084 (0.164)	0.352 (0.684)
		$S_{T,9}(S^*_{T,9})$	0.034 (0.018)	0.114 (0.12)	0.444 (0.544)
		$S_{T,4}(S^*_{T,4})$	0.038 (0.03)	0.156 (0.17)	0.548 (0.734)
	100	$S_{HL,10}(S^*_{HL,10})$	0.040 (0.026)	0.154 (0.208)	0.542 (0.814)
		$S_{HL,5}(S^*_{HL,5})$	0.042 (0.028)	0.148 (0.304)	0.550 (0.930)
		$S_{T,9}(S^*_{T,9})$	0.028 (0.030)	0.214 (0.234)	0.760 (0.854)
		$S_{T,4}(S^*_{T,4})$	0.038 (0.032)	0.238 (0.346)	0.728 (0.944)
	400	$S_{HL,10}(S^*_{HL,10})$	0.016 (0.020)	0.696 (0.890)	1.0 (1.0)
		$S_{HL,5}(S^*_{HL,5})$	0.018 (0.038)	0.718 (0.952)	1.0 (1.0)
		$S_{T,9}(S^*_{T,9})$	0.028 (0.036)	0.844 (0.906)	1.0 (1.0)
		$S_{T,4}(S^*_{T,4})$	0.026 (0.038)	0.828 (0.970)	1.0 (1.0)
			$\beta_2 = 0$	$\beta_2 = 0.8$	$\beta_2 = 1.6$
Model 2d	50	$S^*_{HL,10}$	0.028	0.086	0.626
		$S^*_{HL,5}$	0.042	0.158	0.86
		$S^*_{T,6}$	0.040	0.132	0.814
		$S^*_{T,4}$	0.042	0.188	0.884
	100	$S^*_{HL,10}$	0.062	0.338	0.960
		$S^*_{HL,5}$	0.066	0.456	0.982
		$S^*_{T,6}$	0.050	0.404	0.978
		$S^*_{T,4}$	0.072	0.468	0.986
	400	$S^*_{HL,10}$	0.054	0.942	1.0
		$S^*_{HL,5}$	0.058	0.976	1.0
		$S^*_{T,6}$	0.070	0.978	1.0
		$S^*_{T,4}$	0.054	0.986	1.0

The models that were fitted in Pan and Lin[3] were, also, considered here. Specifically, we examined the goodness of fit of the models

1. $$\operatorname{logit}\left(p_{ij}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{ij,2} + \beta_3 x_{ij,3} + \beta_4 x_{ij,4} + \beta_5 x_{ij,5} + \beta_6 x_{ij,6} + \upsilon_i,$$
2. $$\operatorname{logit}\left(p_{ij}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{ij,2} + \beta_3 x_{ij,3} + \beta_4 \frac{x_{ij,4}}{10} + \beta_5 \left(\frac{x_{ij,4}}{10}\right)^2 + \beta_6 x_{ij,7} + \beta_7 x_{ij,8} + \beta_8 x_{ij,9} + \upsilon_i,$$
3. $$\operatorname{logit}\left(p_{ij}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{ij,2} + \beta_3 x_{ij,3} + \beta_4 \frac{x_{ij,4}}{10} + \beta_5 \left(\frac{x_{ij,4}}{10}\right)^2 + \beta_6 \left(\frac{x_{ij,4}}{10}\right)^3 + \beta_7 x_{ij,7} + \beta_8 x_{ij,8} + \beta_9 x_{ij,9} + \upsilon_i,$$

TABLE 7 Model parameters estimates for the respiratory infection data

Covariate	First model			Second model			Third model		
	$\hat{\beta}$	SE	*P* value	$\hat{\beta}$	SE	*P* value	$\hat{\beta}$	SE	*P* value
Constant	−2.21	0.22	< 0.0001	−2.26	0.36	< 0.0001	−2.21	0.359	< 0.0001
x_1	−0.50	0.25	0.046	−0.52	0.26	0.044	−0.54	0.253	0.033
x_2	−0.04	0.022	0.049	−0.045	0.022	0.042	−0.042	0.022	0.056
x_5	−0.61	0.18	0.001	…	…	…	…	…	…
x_6	−0.17	0.18	0.351	…	…	…	…	…	…
x_3	0.56	0.48	0.243	0.54	0.49	0.268	0.547	0.483	0.26
x_4	−0.03	0.001	< 0.0001	…	…	…	…	…	…
x_4^2	−0.001	0.0004	0.01	…	…	…	…	…	…
$\frac{x_4}{10}$	…	…	…	−0.301	0.082	< 0.0001	−0.52	0.136	< 0.0001
$\left(\frac{x_4}{10}\right)^2$	…	…	…	−0.11	0.042	0.009	−0.14	0.043	0.001
$\left(\frac{x_4}{10}\right)^3$	…	…	…	…	…	…	0.034	0.017	0.04
x_7	…	…	…	−0.71	0.52	0.17	−0.734	0.52	0.16
x_8	…	…	…	0.78	0.349	0.026	0.79	0.35	0.023
x_9	…	…	…	0.036	0.38	0.924	0.022	0.375	0.953
$log\left(\sigma^2\right)$	−0.897	0.25	< 0.0001	−0.773	0.241	0.001	−0.898	0.254	< 0.0001

Abbreviation: SE, standard error.

where the $\boldsymbol{\upsilon}$s are assumed to be iid zero mean normal variates with variance σ^2, and x_7, x_8, and x_9 are indicator variates for the first, second, and third seasons, respectively. The estimate of the unknown parameters $\boldsymbol{\beta}, \boldsymbol{\upsilon}$ was based on the iterative weighted least square Equation 3, while the estimation of the scalar parameter δ, for the dispersion model $\log\left(\sigma^2\right) = \delta$, was based on the iterative weighted least square Equation 4 emerged from the adjusted profile likelihood that is based on the second order Laplace approximation to the marginal. The estimates of these parameters for the three models are given in Table 7. The estimates provided in Table 7 are almost the same as the ones provided in Pan and Lin.[3]

To assess the goodness of fit of models 1 to 3 using S_{HL} and S_{HL}^*, the expected values of the observations were calculated using (9) and $G = 10$. The partitioning of the observations for the evaluation of the score tests S_T and S_T^* was carried out as follows: For model 1, the observations were partitioned according to all possible combinations of the indicators $I(x_1 = 1)$, $I(x_2 > 0)$, $I(x_3 = 1)$, $I(x_4 > 0)$, $I(x_5 > 0)$, and $I(x_6 > 0)$. For models 2 and 3, the groups were formed according to all possible combinations of the indicators $I(x_1 = 1)$, $I(x_2 > 0)$, $I(x_3 = 1)$, $I(x_4 > 0)$, and the four groups according to the season that the measurements were recorded. The weighted statistics S_{HL}^*, S_T^*, and S_{PR}^* were calculated for assessing whether any misspecification of the fixed part could be attributed to the omission of a cubic term of x_4 for models 1 and 2.

Table 8 provides the score tests that assess the goodness of fit of models 1 to 3. The table includes the results from the evaluation of the supremum tests that correspond to cumulative processes of Pan and Lin.[3] We can observe that only S_T and S_{x_6} clearly indicate a lack of fit of the first model. The *P* value for S_T^* suggests that a possible reason for the lack of fit for both models 1 and 2 could be the wrongly defined functional form of x_4. Model 2 lack of fit is indicated by S_{HL} and S_T. Finally, all tests signify the adequacy of model 3.

Based on the good performance of the weighted score statistics, it is important to see how the notion of partial correlation between the adjusted dependent variable and a possible missing covariate given all covariates in X is exploited for the calculation of the weights. Figure 1 illustrates how the weights contribute to the value of a weighted score statistic for two different possible specifications, that of $d_{ij} = x_{ij,4}^3$ and $d_{ij} = x_{ij,2}^2$ in model 1. The weights are the residuals of the regression of each d_{ij} against the covariates already in $\boldsymbol{X}$. These are plotted against the residuals $\boldsymbol{y}^* - \hat{\boldsymbol{p}}$. The lower part of the figure corresponds to a situation where there is no correlation between these two quantities. This is reflected to the high *P* value of S_T^* $(P = 0.578)$, and therefore, the effect of $x_{ij,2}^2$ is not significant. In the upper part of the figure, large residual values $\boldsymbol{y}^* - \hat{\boldsymbol{p}}$ in the range between 0 and 0.4 are associated with large weights based on $d_{ij} = x_{ij,4}^3$. Note that the *P* value of S_T^* $(P = 0.0023)$ rejects model adequacy.

TABLE 8 Goodness-of-fit tests for the three models fitted to the respiratory infection data

Statistic	First model			Second model			Third model		
	Value	df	*P* value	Value	df	*P* value	Value	df	*P* value
S_{HL}	14.73	9	0.099	22.93	9	0.006	7.32	9	0.610
S^*_{HL}	13.22	10	0.210	10.86	10	0.370	…	…	…
S_T	58.2	38	0.019	66.24	50	0.062	64.4	50	0.090
S^*_T	71.4	41	0.0023	73.94	55	0.045	…	…	…
S_{PR}	5.9	10	0.820	17.1	22	0.760	15.81	22	0.830
S^*_{PR}	2.27	10	0.994	25.33	24	0.390	…	…	…
S_x	…	…	0.117	…	…	0.314	…	…	0.547
S_g	…	…	0.158	…	…	0.274	…	…	0.534
S_{x_5,x_6}	…	…	0.073	…	…	…	…	…	…
S_{x_6}	…	…	0.038	…	…	…	…	…	…
S_{x_7,x_8,x_9}	…	…	…	…	…	0.973	…	…	…
S_{x_4}	…	…	…	…	…	0.097	…	…	0.671

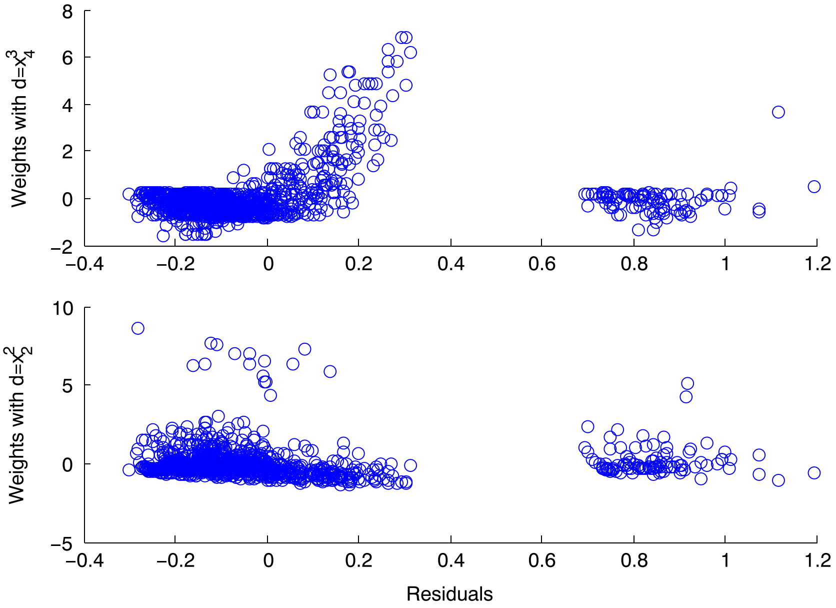

FIGURE 1 Score tests weights vs residuals $y^* - \hat{p}$ under two proposed specifications for model entrance, x_4^3 and x_2^2, in model 1 [Colour figure can be viewed at wileyonlinelibrary.com]

6 | DISCUSSION

In this article, we suggested score statistics based on partitioning processes for the assessment of the goodness of fit of the fixed part of a random effects model with binary responses. We focused our study on models with normal random effects; however, analogous statistics could be provided for distributions other than normal. The computation of the score and the weighted score tests is not complex unless the second order Laplace approximation is used for the estimation of the dispersion parameters. Some statistical packages, like Genstat or R, have incorporated these h-likelihood methods for inference in random effect models.

Based on the simulation results, we can conclude that the weighted versions of the statistics perform better than their unweighted analogs provided that an idea about the type of model misspecification is available. Further, all weighted score tests present equivalent performance at least for a large cluster size. Among the unweighted tests, the partitioning of the observations according to the covariate space seems to provide statistics with better performance in the majority of scenarios considered.

One could raise a subjectiveness issue for the statistics based on partitioning processes due to the subjective choice of the number of groups and their formation. According to our simulation study, the differences in inference because of the different number of groups exist but are not acute. Other grouping procedures could also be encountered such as those using clustering methods, as proposed in Xie et al[24] for the ordinary logistic regression setting. The issue needs further investigation, and further improvements of the proposed tests could be introduced incorporating the optimal choice for the number of groups.

The use of the statistics suggested in this article is based on the assumption that the structure of the random effects is correctly defined in the mean model (1). In cases where the random effects structure is not correctly defined, the results of the simulation study indicate that the behavior of the score statistics depends on the type of misspecification in the random effects structure. To the best of our knowledge, there is not an overall test proposed in the literature that examines the goodness of fit of the whole mean model (1). As a complement to the approaches proposed here, one can examine the structure of the random part by using the tests proposed in Alonso et al.[19,20] Future research could be dedicated to find extensions of the proposed tests so as to be able to examine the goodness of fit of the whole mean model.

ACKNOWLEDGMENT

We would like to thank the referees and the associate editor for their constructive comments and suggestions, which led to a significant improvement of the paper.

ORCID

Antonia K. Korre http://orcid.org/0000-0002-1546-3724

REFERENCES

1. Evans SR, Hosmer DW. Goodness of fit tests in mixed effects logistic models characterized by clustering. *Commun Stat-Theory and Methods*. 2004;33(5):1139-1155.
2. Sturdivant RX, Hosmer DW. A smoothed residual based goodness of fit statistic for logistic hierarchical regression models. *Comput Stat Data Anal*. 2007;51(8):3898-3912.
3. Pan Z, Lin YD. Goodness-of-fit methods for generalized linear mixed models. *Biometrics*. 2005;61:1000-1009.
4. Barnhart XH, Williamson MJ. Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics*. 1998;54:720-729.
5. Horton NJ, Bebchuk JD, Jones CL, et al. Goodness-of-fit for GEE: an example with mental health service utilization. *Stat Med*. 1999;18:213-222.
6. Evans SR, Li L. A comparison of goodness of fit tests for the logistic GEE model. *Stat Med*. 2005;24:1245-1261.
7. Tsiatis A. A note on a goodness-of-fit test for the logistic regression model. *Biometrika*. 1980;67:250-251.
8. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat-Theory and Methods*. 1980;9(10):1043-1069.
9. Pulkstenis E, Robinson JT. Two goodness-of-fit tests for logistic regression models with continuous covariates. *Stat Med*. 2002;21:79-93.
10. Hosmer DW, Hjort LN. Goodness of fit processes for logistic regression: simulation results. *Stat Med*. 2002;21:2723-2738.
11. Lee Y, Nelder JA. Hierarchical generalized linear models (with discussion). *J R Stat Soc B*. 1996;58:619-656.
12. Lee Y, Nelder JA. Hierarchical generalized linear models: a synthesis of generalized linear models, random effect model and structured dispersion. *Biometrika*. 2001;88:987-1006.
13. Lee Y, Nelder JA. Double hierarchical generalized linear models (with discussion). *Appl Stat*. 2006;55:139-185.
14. Pourahmadi M. *Biometrika*. 2000;87:425-435.
15. Coull BA, Agresti A. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics*. 2000;56:73-80.
16. Yun S, Lee Y. Comparison of hierarchical and marginal likelihood estimators for binary outcomes. *Comput Stat Data Anal*. 2004;45:639-650.
17. Lee Y, Nelder JA, Pawitan Y. *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*. London: Chapman and Hall; 2007.
18. Noh M, Lee Y. REML Estimation for binary data in GLMMs. *J Multivar Anal*. 2007;98:896-915.
19. Alonso A, Litiere S, Molenberghs G. A family of tests to detect misspecifications in random effects structure of generalized linear mixed models. *Comput Statist Data Anal*. 2008;52(9):4474-4486.
20. Alonso A, Litiere S, Molenberghs G. Testing for misspecification in generalized linear mixed models. *Biostatistics*. 2010;11(4):771-786.
21. Zeger SL, Liang KY, Albert P. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988;44:1049-1060.
22. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88(421):9-25.

23. Sommer A, Katz J, Tarwotjo I. Increased risk of respiratory infection and diarrhea in children with pre-existing mild vitamin A deficiency. *Am J Clin Nutr*. 1984;40:1090-1095.

24. Xie XJ, Pendergast J, Clarke W. Increasing the power: a practical approach to goodness-of-fit test for logistic regression models with continuous predictors. *Comput Statist Data Anal*. 2008;52(5):2703-2713.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Korre AK, Vasdekis VGS. Goodness of fit tests for random effect models with binary responses. *Statistics in Medicine*. 2018;37:4571–4587. https://doi.org/10.1002/sim.7941

APPENDIX A

R1: (*Form of the score equation*). For model (5), matrix Σ_α^{-1} is as defined in (3). The form of $\mathbf{T}$ and $\mathbf{z}_\alpha$ is as follows:

$$\mathbf{T} = \begin{pmatrix} \mathbf{X} & \mathbf{O} & \mathbf{Z} \\ \mathbf{0}_{q\times p} & \mathbf{0}_{q\times G} & \mathbf{I}_q \end{pmatrix}, \qquad \mathbf{z}_\alpha = \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} + \mathbf{O}\boldsymbol{\gamma} + \mathbf{Z}\boldsymbol{\upsilon} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{s} - \mathbf{p}) \\ \Sigma\mathbf{Z}^T\mathbf{s} \end{pmatrix}.$$

Substituting these values into (3), the set of equations becomes

$$\begin{pmatrix} \mathbf{X}^T\mathrm{WX} & \mathbf{X}^T\mathrm{WO} & \mathbf{X}^T\mathrm{WZ} \\ \mathbf{O}^T\mathrm{WX} & \mathbf{O}^T\mathrm{WO} & \mathbf{O}^T\mathrm{WZ} \\ \mathbf{Z}^T\mathrm{WX} & \mathbf{Z}^T\mathrm{WO} & \mathbf{Z}^T\mathrm{WZ} + \Sigma^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \\ \boldsymbol{\upsilon} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{W}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{O}\boldsymbol{\gamma} + \mathbf{Z}\boldsymbol{\upsilon} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{s} - \mathbf{p})\right) \\ \mathbf{O}^T\mathbf{W}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{O}\boldsymbol{\gamma} + \mathbf{Z}\boldsymbol{\upsilon} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{s} - \mathbf{p})\right) \\ \mathbf{Z}^T\mathbf{W}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{O}\boldsymbol{\gamma} + \mathbf{Z}\boldsymbol{\upsilon} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{s} - \mathbf{p})\right) + \mathbf{Z}^T\mathbf{s} \end{pmatrix}.$$

Therefore, the score equations for the fixed effects $\boldsymbol{\beta}, \boldsymbol{\gamma}$ become

$$\mathbf{X}^T(\mathbf{y} - \mathbf{s} - \mathbf{p}) = \mathbf{0},$$
$$\mathbf{O}^T(\mathbf{y} - \mathbf{s} - \mathbf{p}) = \mathbf{0},$$

from which one can see that the score for $\boldsymbol{\gamma}$ is $\mathbf{O}^T(\mathbf{y} - \mathbf{s} - \mathbf{p})$.