TEST (2019) 28:827–843
https://doi.org/10.1007/s11749-018-0602-6

ORIGINAL PAPER

Link misspecification in generalized linear mixed models with a random intercept for binary responses

Shun Yu[1] · Xianzheng Huang[2]

Received: 14 August 2017 / Accepted: 18 July 2018 / Published online: 3 August 2018

Abstract
We present in this paper tests for link misspecification in generalized linear mixed models with a random intercept for binary responses. To facilitate model diagnosis, we consider two types of grouped responses induced from the original responses and also stochastically create reclassified binary responses. Maximum likelihood estimators based on the original observed data and the counterpart estimators based on different induced data sets are investigated in the presence of link misspecification. Results from this investigation motivate four diagnostic tests for assessing the adequacy of an assumed link, which can provide information on how the true link differs from an assumed symmetric link when the proposed tests reject the assumed link. The performance of these tests is illustrated via simulation studies in comparison with an existing method for checking link assumptions. These tests are applied to a data set from a longitudinal study that was analyzed in the existing literature using logistic regression.

Keywords Grouped response · Logistic regression · Reclassified response

Mathematics Subject Classification 62J12 · 62F03

1 Introduction

Since the seminal paper of Nelder and Wedderburn (1972), the class of generalized linear models (GLM) has received wide acceptance in a host of applications (McCullagh

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11749-018-0602-6) contains supplementary material, which is available to authorized users.

✉ Xianzheng Huang
huang@stat.sc.edu

[1] Wells Fargo & Company, Charlotte, NC 28282, USA

[2] Department of Statistics, University of South Carolina, Columbia, SC 29208, USA

and Nelder 1989). It provides a practically interpretable and mathematically flexible platform for studying the association between a non-normal response and covariates of interest. When the response associated with an experimental unit is a vector of correlated components, as encountered in longitudinal studies for instance, random effects are included in the models to account for the correlation between responses from the same experimental unit. This modification of GLM leads the class of generalized linear mixed models (GLMM, Molenberghs and Verbeke 2005). This study focuses on GLMM with a random intercept for clustered binary responses, which is an important subclass of GLMM that has received lots of attention from researchers (Neuhaus et al. 1992; Wang and Louis 2003, 2004; Tchetgen and Coull 2006; Caffo et al. 2007, etc.).

The widely entertained GLMM for binary responses often assume one of the popular links such as logit, probit, and complementary log–log, mostly due to ease of interpretation and convenient implementation using standard software. However, from a practical standpoint, a symmetry link, such as logit and probit, may not be reasonable in many applications (Jiang et al. 2013), and, although asymmetric, the complementary log–log link only allows a fixed negative skewness. From a theoretical standpoint, Li and Duan (1989) studied regression analysis under a misspecified link function in general regression settings; Czado and Santner (1992) zoomed in on GLM for a binary response to study the effects of link misspecification on regression analysis. These authors provided theoretical and empirical evidence of the adverse effects of a misspecified link in GLM on likelihood-based inference. They showed that the maximum likelihood estimators (MLE) of regression coefficients obtained under an inappropriate link can be inconsistent and inefficient. These adverse effects carry over to likelihood-based inference in the GLMM framework with an inadequate link (Samejima 2000; Pan and Lin 2005).

There are two ways to avoid an inadequate assumed link function. The more actively explored way is to formulate a flexible class of link functions (Aranda-Ordaz 1981; Guerrero and Johnson 1982; Morgan 1983; Whittemore 1983; Stukel 1988; Kim et al. 2008; Jiang et al. 2013). Alternatively, one keeps the simple and popular choice of link, such as the logit link, and assume a more flexible functional form through which covariates enter the conditional mean model of the response. This approach can be unattractive to practitioners when a specific (often simple) form of the linear predictor in GLMM is desirable for meaningful interpretations of the fixed or random effects. This is the case in, for example, models in the item response theory as discussed in Samejima (2000). There, the author showed that maximum likelihood estimation based on a logistic regression produces results that contradict with the psychological reality. As a remedy, she proposed a family of models with asymmetric links while keeping the functional form of the linear predictor unchanged.

Before employing a potentially complicated flexible link or revising the functional form of covariates in the linear predictor, data analysts may wish to validate an assumed link using some diagnostic tools. If one finds sufficient evidence that a simpler (and easier to interpret) link function is adequate for a given application, sticking to it can potentially improve efficiency of follow-up statistical inference compared to when one adopts a more complicated flexible link right off the bat. Many researchers have developed diagnostic methods for assessing the adequacy of an assumed GLMM. Most of the existing methods are goodness-of-fit (GOF) tests in nature that are designed to assess the overall adequacy of a GLMM rather than a specific assumption of the model (e.g.,White 1981; Ritz 2004; Pan and Lin 2005). There is also a sizable collection of methods for testing assumptions on random effects in GLMM (Waagepetersen 2006; Tchetgen and Coull 2006; Alonso et al. 2008; Huang 2009; Verbeke and Molenberghs 2013; Yu and Huang 2017), but there are much fewer effective diagnostics tools for checking link assumptions, especially in the framework of GLMM. In this study, we assume correct modeling on the random effect and focus on developing tests to detect link misspecification. The proposed diagnostic tools are designed for one source of model misspecification at a time.

One common theme running through our proposed methods is to make use of different types of clustered binary responses induced from the original responses. The intention of creating these induced responses is to draw likelihood-based inference based on the induced data and then to compare with the counterpart inference based on the raw data, or simply compare with the truth when the truth is known to data analysts. With an adequately assumed link in the GLMM for the original data, all inference should be consistent under regularity conditions regardless of which induced data set is used, so is that from the raw data. With an inadequate link, these inferences can be biased in different ways. Although this common theme was exploited in Huang (2009) and Yu and Huang (2017) to develop diagnostic tests for random-effects assumptions in GLMM, they only used one particular type of induced responses considered in our current study. Here, we explore more versatile ways to create induced data, leading to different diagnostic tools that can be effective in detecting link misspecification in different contexts. In addition, the introduction of a user-specified extraneous parameter arising in one of the three types of considered induced response leads to a testing procedure that only requires one round of maximum likelihood estimation as opposed to two rounds as needed in the two aforementioned existing works, which also adds to the novelty of our current proposals. Section 2 is devoted to describing these three types of induced responses and presenting the corresponding likelihood functions. In Sect. 3, we investigate MLEs in the presence of link misspecification resulting from different induced data, as well as those based on the raw data. Patterns of these estimators motivate the tests for link specification presented in Sect. 4. In Sect. 5, simulation experiments are conducted to demonstrate the operating characteristics of the proposed tests. Using the proposed tests, we assess the adequacy of the logit link in the logistic GLMM used to analyze a data set from a longitudinal study in Sect. 6. We summarize the contributions of our study and discuss future research in Sect. 7.

2 Model and induced responses

Denote by $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^t$ the vector of observed binary responses from cluster i, and by $\mathbf{X}_i$ the $n_i \times p$ matrix of covariates, with the jth row being $\mathbf{X}_{ij}$, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Suppose one posits a conditional mean of Y_{ij} given the covariates and the random effect as $E(Y_{ij}|\mathbf{X}_{ij}, b_{i0}; \boldsymbol{\beta}) = h(\mathbf{X}_{ij}\boldsymbol{\beta} + b_{i0})$, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effects, b_{i0} is the random intercept, $h(\cdot)$ is a differentiable non-decreasing link function, and

$\mathbf{X}_{ij}\boldsymbol{\beta} + b_{i0}$ is referred to as the linear predictor. Suppose the distribution of b_{i0} is (correctly) specified by the probability density function (pdf), $f_b(b_{i0}; \tau)$, where τ is the parameter associated with this distribution. Then the contribution to the observed data likelihood from cluster i is $f_{\mathbf{Y}}(\mathbf{Y}_i|\mathbf{X}_i; \boldsymbol{\beta}, \tau) = \int f_b(b_{i0}; \tau) \prod_{j=1}^{n_i} h(\mathbf{X}_{ij}\boldsymbol{\beta} + b_{i0})^{Y_{ij}} \left\{1 - h(\mathbf{X}_{ij}\boldsymbol{\beta} + b_{i0})\right\}^{1-Y_{ij}} \mathrm{d}b_{i0}$, for $i = 1, \ldots, m$. For the purpose of developing diagnostic tools, we next construct three types of responses induced from the observed responses, followed by the likelihood function of each induced data set.

The first type of induced responses is motivated by group testing (Dorfman 1943). To create such responses, one partitions cluster i into g_i groups; then, one defines grouped responses $Y_{ig}^* = \max_{j \in I_{ig}} Y_{ij}$, where I_{ig} is the index set such that $j \in I_{ig}$ indicates that observation j is in group g within cluster i, for $i = 1, \ldots, m$, $g = 1, \ldots, g_i$. Letting $\mathbf{Y}_i^* = (Y_{i1}^*, \ldots, Y_{ig_i}^*)^t$, we refer to the so-constructed responses, $\{\mathbf{Y}_i^*\}_{i=1}^m$, the unbalanced grouped responses. The name for this type of responses is to reflect that 0 and 1 are not treated equally in the definition of the grouped response because Y_{ig}^* takes value 1 if at least one of the original responses in $\{Y_{ij},\ j \in I_{ig}\}$ is 1. Following this definition, one has the likelihood of the unbalanced grouped responses from cluster i as, for $i = 1, \ldots, m$,

$$f_{\mathbf{Y}^*}(\mathbf{Y}_i^*|\mathbf{X}_i; \boldsymbol{\beta}, \tau) = \int f_b(b_{i0}; \tau) \prod_{g=1}^{g_i} \left[1 - \prod_{j \in I_{ig}} \left\{1 - h(\mathbf{X}_{ij}\boldsymbol{\beta} + b_{i0})\right\}\right]^{Y_{ig}^*}$$
$$\times \left[\prod_{j \in I_{ig}} \left\{1 - h(\mathbf{X}_{ij}\boldsymbol{\beta} + b_{i0})\right\}\right]^{1-Y_{ig}^*} \mathrm{d}b_{i0}.$$

The second type of induced responses is similar to the unbalanced grouped responses in that it also requires dividing cluster i into g_i groups, for $i = 1, \ldots, m$. The difference is that, for group g in cluster i, one defines a grouped response as $Y_{ig}^{**} = I(\sum_{j \in I_{ig}} Y_{ij} \geq n_{ig}/2)$, where $I(\cdot)$ is the indicator function, and n_{ig} is size of group g in cluster i, for $i = 1, \ldots, m$, $g = 1, \ldots, g_i$. Denoting by $\mathbf{Y}_i^{**} = (Y_{i1}^{**}, \ldots, Y_{ig_i}^{**})^t$, we call the so-constructed responses, $\{\mathbf{Y}_i^{**}\}_{i=1}^m$, the balanced grouped responses. This term is motivated by the fact that the definition of Y_{ig}^{**} treats 0 and 1 more fairly in the sense that Y_{ig}^{**} takes value 1 when at least half of the original responses in that group, $\{Y_{ij},\ j \in I_{ig}\}$, are equal to 1. With this definition, the contribution of cluster i to the balanced grouped data likelihood is, for $i = 1, \ldots, m$,

$$f_{\mathbf{Y}^{**}}(\mathbf{Y}_i^{**}|\mathbf{X}_i; \boldsymbol{\beta}, \tau) = \int f_b(b_{i0}; \tau) \prod_{g=1}^{g_i} \left\{P(Y_{ig}^{**} = 1|\mathbf{X}_i, b_{i0}; \boldsymbol{\beta}, \tau)\right\}^{Y_{ig}^{**}}$$
$$\times \left\{P(Y_{ig}^{**} = 0|\mathbf{X}_i, b_{i0}; \boldsymbol{\beta}, \tau)\right]^{1-Y_{ig}^{**}} \mathrm{d}b_{i0},$$

where $P(Y_{ig}^{**} = 0|\mathbf{X}_i, b_{i0}; \boldsymbol{\beta}, \tau) = P(\sum_{j \in I_{ig}} Y_{ij} < n_{ig}/2|\mathbf{X}_i, b_{i0}; \boldsymbol{\beta}, \tau)$. Because conditioning on $\mathbf{X}_i$ and b_{i0}, $\sum_{j \in I_{ig}} Y_{ij}$ is the sum of n_{ig} independent but not identically distributed Bernoulli random variables, one can see that $\sum_{j \in I_{ig}} Y_{ij}$ follows a Poisson binomial distribution (Wang 1993). For illustration purposes, we provide an example of how to derive $P(Y_{ig}^{**} = 0|\mathbf{X}_i, b_{i0}; \boldsymbol{\beta}, \tau)$ in Appendix A in supplementary material.

The third type of induced responses is referred to as the reclassified responses. Here, one replaces the value of Y_{ij} by $1 - Y_{ij}$ with certain probability, producing a new response $\tilde{Y}_{ij}$, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. For simplicity, we use the reclassification mechanism, $P(\tilde{Y}_{ij} = 1 - Y_{ij}|Y_{ij}) = \rho$, as the stochastic model according to which $\tilde{Y}_{ij}$ is generated, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where $\rho \in (0, 0.5)$ is a user-specified constant. Denoting by $\tilde{\mathbf{Y}}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{in_i})^t$ the vector of reclassified responses for cluster i, one can show that the likelihood of $\tilde{\mathbf{Y}}_i$ is, for $i = 1, \ldots, m$,

$$f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{Y}}_i|\mathbf{X}_i; \boldsymbol{\beta}, \tau) = \int f_b(b_{i0}; \tau) \prod_{j=1}^{n_i} \{1 - \rho - (1 - 2\rho) h(\mathbf{X}_{ij}\boldsymbol{\beta} + b_{i0})\}^{1-\tilde{Y}_{ij}}$$
$$\times \{\rho + (1 - 2\rho) h(\mathbf{X}_{ij}\boldsymbol{\beta} + b_{i0})\}^{\tilde{Y}_{ij}} \mathrm{d}b_{i0}.$$

3 Maximum likelihood estimators

With three types of induced responses constructed and the corresponding likelihood functions derived, we now can study the MLEs of relevant parameters resulting from each type of induced responses, as well as the counterpart MLEs based on the original responses. The goal of the investigation in this section is to understand how the consistency of the MLE of a parameter is compromised by link misspecification when each of the four types of responses is used for maximum likelihood estimation.

To gain insight on the asymptotic mean of these MLEs, we generate a sample with $m = 10^5$ clusters, each cluster of size $n_i = 6$, from the true GLMM given by $E(Y_{ij}|\mathbf{X}_i, b_{i0}; \boldsymbol{\beta}, \tau) = h_0(\beta_0 + \beta_1 X_{ij,1} + \beta_2 X_{ij,2} + \beta_3 X_{ij,1} X_{ij,2} + b_{i0})$, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where $h_0(\cdot)$ is the true link given by the cumulative distribution function (cdf) of a skew normal (SN) distribution (Azzalini 1985). More specifically, this link is given by

$$h_0(s) = \Phi(s) - 2T(s, \alpha), \tag{1}$$

for $-\infty < s < \infty$, where $\Phi(\cdot)$ is the standard normal cdf, $T(s, \alpha)$ is the Owen's T function (Owen 1956), and α is a shape parameter. When $\alpha = 0$, $h_0(\cdot)$ is the probit link; when $\alpha < 0\,(> 0)$, the corresponding skew normal distribution is left (right) skewed. Viewing a link function as a cdf in general, we say a link function is symmetric or left/right skewed if the cdf corresponds to a symmetric or left-/right-skewed pdf. In the true GLMM, we set $X_{ij,1} = x_i$, with x_i equal to 0 for $m/2$ clusters, and equal to 1 for the remaining $m/2$ clusters, and we set $X_{ij,2} = (j - 1)/(n_i - 1)$, for $j = 1, \ldots, n_i$. The random intercept, $\{b_{i0}\}_{i=1}^m$, is generated from $N(0, \sigma^2)$, where $\sigma = 1$ is the only parameter, i.e., τ, in the pdf $f_b(b_{i0}; \tau)$. Finally, the values of the fixed effects, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^t$, are $(-2, 1, 0.5, -0.25)^t$. Given a simulated raw data set, $\{(\mathbf{Y}_i, \mathbf{X}_i)\}_{i=1}^m$, we then generate three types of induced data, $\{(\mathbf{Y}_i^*, \mathbf{X}_i)\}_{i=1}^m$,

$\{(\mathbf{Y}_i^{**}, \mathbf{X}_i)\}_{i=1}^m$, and $\{(\tilde{\mathbf{Y}}_i, \mathbf{X}_i)\}_{i=1}^m$, respectively, following their definitions in Sect. 2. To create the (un)balanced grouped responses, we split each cluster into $g_i = 2$ groups with $I_{i1} = \{1, 2, 3\}$ and $I_{i2} = \{4, 5, 6\}$, for $i = 1, \ldots, m$. To create different sets of reclassified responses, we vary ρ over the range $(0, 0.5)$ at increments of 0.05.

Denote by $\boldsymbol{\Omega} = (\boldsymbol{\beta}^t, \sigma)^t$ the vector of all parameters in the GLMM, and by $\hat{\boldsymbol{\Omega}}$ the raw-data MLE of $\boldsymbol{\Omega}$, by $\hat{\boldsymbol{\Omega}}^*$, $\hat{\boldsymbol{\Omega}}^{**}$, and $\tilde{\boldsymbol{\Omega}}$ the counterpart estimators based on the unbalanced grouped data, the balanced grouped data, and the reclassified data, respectively. This notational convention carries over to MLEs of θ, where θ denotes a generic parameter in $\boldsymbol{\Omega}$. For parameter estimation using the reclassified data, we also estimate ρ along with $\boldsymbol{\Omega}$ and denote the MLE of ρ by $\tilde{\rho}$.

When computing the MLEs from the grouped responses, $\hat{\boldsymbol{\Omega}}^*$ and $\hat{\boldsymbol{\Omega}}^{**}$, and compare them with the counterpart raw-data MLE, $\hat{\boldsymbol{\Omega}}$, we assume the raw data come from a probit GLMM, that is, the assumed link is $h(s) = \Phi(s)$. To vary the severity of link misspecification, we vary α in (1) from -5 to 5 at increments of 0.5, with $\alpha = 0$ giving rise to the scenario of correct link specification. Figure 1 depicts $\hat{\boldsymbol{\Omega}}$, $\hat{\boldsymbol{\Omega}}^*$, and $\hat{\boldsymbol{\Omega}}^{**}$ as α varies. As one would expect, except for when $\alpha = 0$, all parameter estimates exhibit bias, with the amount increases in absolute value as α deviates further from zero. For $\hat{\boldsymbol{\Omega}}^*$, the bias tends to be more substantial when $h_0(s)$ is right skewed than when it is left skewed. In contrast, the bias associated with $\hat{\boldsymbol{\Omega}}^{**}$ is more comparable in absolute value when the sign of α switches from negative to positive. We believe that the even treatment of 0 and 1 in the construction of the balanced grouped responses contributes to this symmetric pattern of bias in $\hat{\boldsymbol{\Omega}}^{**}$. And the asymmetric pattern of $\hat{\boldsymbol{\Omega}}^*$ is due to the uneven treatment of 0 and 1 in the definition of the unbalanced grouped responses. While acknowledging that results in Fig. 1 are estimates based on finite samples, and thus are subject to sampling error and numerical inaccuracy in, for instance, computing the integrals that define the likelihood functions in Sect. 2, we believe that these estimates based on samples of size $m = 10^5$ preserve key features of the corresponding limiting MLEs as $m \to \infty$. Although the depicted MLEs can be wiggly (especially $\hat{\boldsymbol{\Omega}}^{**}$) as α varies, it is not unreasonable to expect that the limiting (non-random) MLEs are smooth functions of α that exhibit the patterns highlighted in Fig. 1.

A more important phenomenon implied in Fig. 1 that directly motivates the proposed diagnostic methods described in Sect. 4 is that $\hat{\theta}$, $\hat{\theta}^*$, and $\hat{\theta}^{**}$ do not coincide when $\alpha \neq 0$. Figure 2 shows the difference between each pair of the estimates for each parameter based on two different data sets. Among these differences, $\hat{\theta}^* - \hat{\theta}^{**}$ deviates from zero more substantially for the parameters β_0, β_1, and σ, especially when $\alpha > 0$. This phenomenon also has an indication in the power of the diagnostic tests provided in Sect. 4.

When the reclassified data are used for developing diagnostic methods, we only make use of the MLE of ρ, $\tilde{\rho}$. Because ρ is a user-specified parameter in the reclassification model, its true value is known to data analysts, which makes assessing the bias in $\tilde{\rho}$ practically possible. In this numerical study, we assume a logit GLMM, that is, $h(s) = 1/(1 + e^{-s})$, when computing $\tilde{\rho}$ (along with $\tilde{\boldsymbol{\Omega}}$). The raw data with $m = 10^5$ clusters are generated according to a GLMM with the true link being the logit, probit,

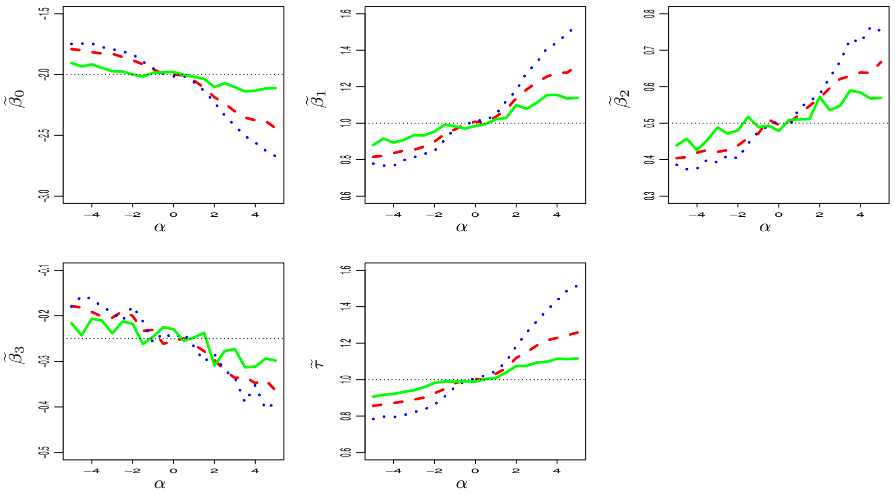

Fig. 1 Maximum likelihood estimates from different data sets with $m = 10^5$ versus the shape parameter, α, in the true link function $h_0(s)$. Raw-data MLEs, $\hat{\mathbf{\Omega}}$: dashed lines; MLEs based on the unbalanced grouped data, $\hat{\mathbf{\Omega}}^*$: dotted lines; MLEs based on the balanced grouped data, $\hat{\mathbf{\Omega}}^{**}$: solid lines. The horizontal dotted reference line in each panel refers to the true parameter value

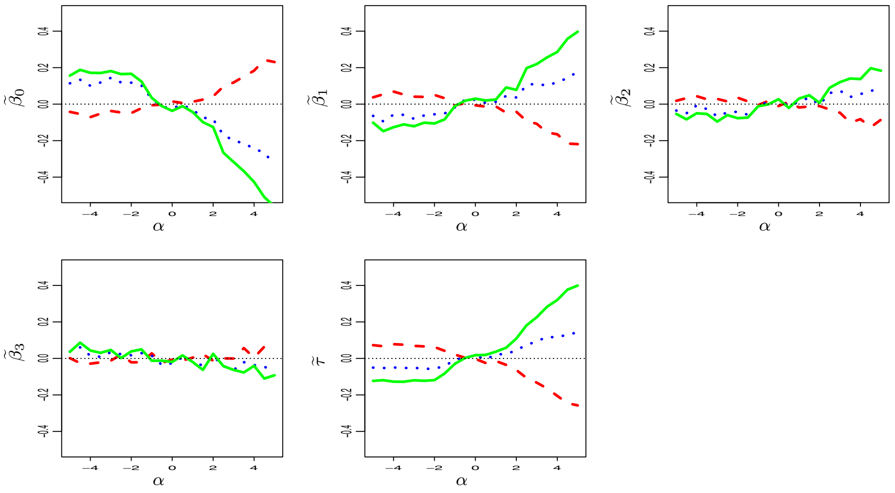

Fig. 2 Differences between two MLEs from two different data sets with $m = 10^5$ versus the shape parameter, α, in the true link function $h_0(s)$. $\hat{\mathbf{\Omega}} - \hat{\mathbf{\Omega}}^*$: dashed lines; $\hat{\mathbf{\Omega}} - \hat{\mathbf{\Omega}}^{**}$: dotted lines; $\hat{\mathbf{\Omega}}^* - \hat{\mathbf{\Omega}}^{**}$: solid lines. The horizontal dotted reference lines signify the value zero

and $h_0(s)$ in (1) with $\alpha = -5,\ 5$. The first choice of the true link yields a case without link misspecification; the choice of the probit link leads to link misspecification that is nearly negligible because the logit link and the probit link are virtually indistinguishable in most inference contexts (Chambers and Cox 1967). These four links are shown in the left panel of Fig. 3. The values of $\tilde{\rho}$ as ρ varies within (0, 0.5) under

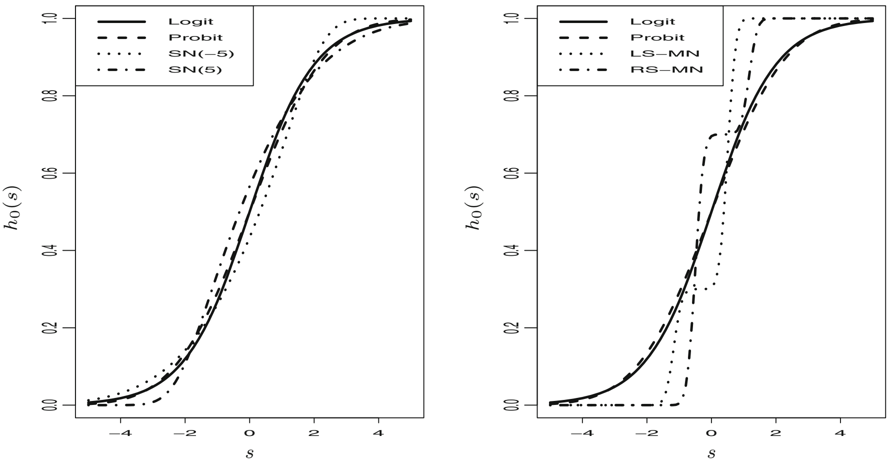

Fig. 3 The left panel shows four link functions: logit (solid line), probit link (dashed line), the skew normal cdf $h_0(s)$ with $\alpha = -5$ (dotted line), and $h_0(s)$ with $\alpha = 5$ (dotted-dashed line). The right panel shows the following four link functions: logit (solid line), probit (dashed line), the cdf of a left-skewed mixture normal distribution (dotted line), and the cdf of a right-skewed mixture normal distribution (dotted-dashed line)

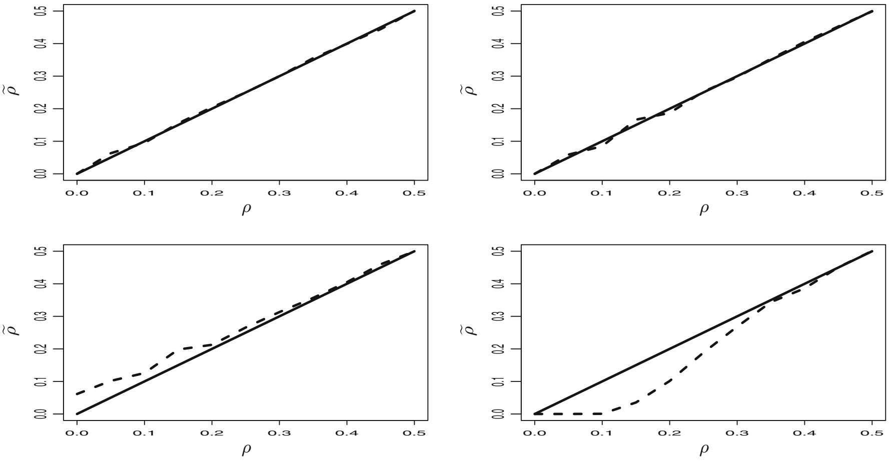

Fig. 4 MLE of ρ, $\tilde{\rho}$ (dashed lines), versus ρ when the true link is logit (top left), probit (top right), SN (− 5) (bottom left), and SN (5) (bottom right). The solid line in each panel is the 45° reference line

these four combinations of assumed link (logit) and the true link are shown in Fig. 4. In the absence of link misspecification or with practically negligible misspecification, $\tilde{\rho}$ matches ρ closely. In the presence of noticeable link misspecification, $\tilde{\rho}$ can deviate from the truth significantly. More interestingly, $\tilde{\rho}$ tends to overestimate ρ when the true link is left skewed, and it underestimates the truth when the true link is right skewed. We will exploit this phenomenon in a proposed diagnostic test next to not only detect link misspecification but also gain information on the skewness of the true link.

4 Tests for link misspecification

Even though the adverse effects of link misspecification on the likelihood-based inference have been well studied and acknowledged, such knowledge cannot be straightforwardly used for model diagnosis because one does not get to see the bias in an MLE of a parameter unless one knows the true value of that parameter. Using the findings in Sect. 3, we can make use of such bias in two ways to detect link misspecification. One is to use the discrepancy between two MLEs resulting from two data sets explored above as an indicator of link misspecification. The key to formulating a useful indicator here is to have two related data sets based on which two MLEs of a parameter have different bias in the presence of link misspecification. Once this is accomplished, even without knowing the true parameter value, the discrepancy (in large-sample sense) between the two MLEs essentially indicates the existence of bias in at least one of the estimators, which in turn suggests the presence of link misspecification. This theme idea has been used in Huang (2009, 2011, 2013) and Yu and Huang (2017) to assess random-effects assumptions in GLMM and linear mixed models (LMM). Another novel way is to introduce an extraneous parameter involved in creating an induced data set from the raw data, such as ρ in the model used to generate the reclassified responses. Using the induced data set to infer ρ allows one to directly assess the bias in the MLE of ρ, $\tilde{\rho}$, since one knows the truth about ρ. The key to the success of this approach is to bring in an extraneous parameter that can be identified along with $\mathbf{\Omega}$ based on the induced data, which is fortunately the case for ρ in the reclassification model used in Sect. 3 (Neuhaus 2002).

Following these two theme ideas, we define four test statistics below,

$$t_{1,\theta} = (\hat{\theta} - \hat{\theta}^{*})/\hat{\nu}_{1,\theta}, \quad t_{2,\theta} = (\hat{\theta} - \hat{\theta}^{**})/\hat{\nu}_{2,\theta},$$
$$t_{3,\theta} = (\hat{\theta}^{*} - \hat{\theta}^{**})/\hat{\nu}_{3,\theta}, \quad t_{4,\rho} = (\tilde{\rho} - \rho)/\hat{\nu}_{\rho},$$

where the first three test statistics can be defined for an arbitrary parameter in $\mathbf{\Omega}$; $\hat{\nu}_{1,\theta}$, $\hat{\nu}_{2,\theta}$, and $\hat{\nu}_{3,\theta}$ are estimators of the standard errors of the corresponding differences on the numerator of the first three test statistics; and $\hat{\nu}_{\rho}$ is the estimator of the standard error of $\tilde{\rho}$. More specifically, $\hat{\nu}_{\rho}$ is the square root of the $[p+2,\ p+2]$ entry in the sandwich variance estimator (Boos and Stefanski 2013) for the $(p+2)$-dimensional MLE of $(\mathbf{\Omega}^{t},\ \rho)^{t}$. The decision rule for all four tests is to reject the null (of lack of link misspecification) when the value of the test statistic deviates significantly from zero. By the asymptotic properties of MLE, in the absence of link misspecification, $t_{4,\rho}$ follows a t distribution with $m-p-2$ degrees of freedom asymptotically. In Appendix B in supplementary material, we elaborate the construction of $\hat{\nu}_{1,\theta}$ and show that the null distribution of $t_{1,\theta}$ is a t distribution with $m-p-1$ degrees of freedom asymptotically. Similarly, it can be shown that $t_{2,\theta}$ and $t_{3,\theta}$ also have the same null distribution as that of $t_{1,\theta}$. Besides theoretical justification, Appendix B also provides empirical evidence of these claimed null distributions using QQ plots of realizations of the test statistics from simulation study.

The test based on $t_{4,\rho}$ has the advantage over the first three tests in that, under mild regularity conditions, a consistent $\tilde{\rho}$, which tends to yield an insignificant value of $t_{4,\rho}$, typically implies a correct model for the raw data in most practical scenarios. Here, the regularity conditions are classical regularity conditions for maximum likelihood estimation (Cox and Hinkley 1974, page 281). Under these conditions, ρ is identifiable based on the induced reclassified data, along with other parameters in the model. Thus, under the correct model, a consistent MLE of ρ is expected, and failing to reject the null by $t_{4,\rho}$ can be interpreted as lack of data evidence against the null. In contrast, using the other three tests, when one fails to reject the null due to an insignificant value of the test statistic, one can only conclude that the two MLEs are similar, or tend to the same limit as $m \to \infty$, not necessarily that the assumed link is adequate. Indeed, two inconsistent MLEs can have the same asymptotic mean under a particular wrong model. Instead of validating the assumed link, failing to reject the null by the first three tests can be interpreted as robustness in the MLEs of θ in the presence of (potential) model misspecification, which is a desirable feature for an estimator. In addition, although beyond the scope of the current study, the first three tests can be more informative than the fourth when random slopes are included in a GLMM and other sources of model misspecification besides a link misspecification are of interest (Huang 2009).

In the next section, we present simulation studies to demonstrate the operating characteristics of the four proposed tests in the presence of different link misspecification scenarios. In the simulation experiment, we use the residual-based method of testing link misspecification in GLMM proposed by Pan and Lin (2005) as the competing method.

5 Empirical evidence

Unlike the numerical study presented in Sect. 3, which aims to approximate the limiting MLEs of relevant parameters, here we set $m = 600$ for each simulation setting, each cluster of size 6, and we create 1000 Monte Carlo (MC) replicates under each setting to monitor the rejection rate of each test. In the raw-data-generating process, the true GLMM has $b_{i0} \sim N(0, 1)$ and the same covariates setting and values of $\boldsymbol{\beta}$ as those in Sect. 3. We consider four true link functions when generating clustered binary responses: logit, probit, a link defined as the cdf of the left-skewed normal given by $(7/10)N(1/2, 0.2^2) + (3/10)N(-7/6, 0.2^2)$, and a link defined as the cdf of the right-skewed normal given by $(7/10)N(-1/2, 0.2^2) + (3/10)N(7/6, 0.2^2)$. These four link functions are displayed in the right panel of Fig. 3. When computing the MLEs needed for the tests, we assume a logit GLMM. Finally, for the test $t_{4,\rho}$, we experiment on two levels of ρ, 0.05 and 0.1. Using a significance level of 0.05, we record under each true model setting the rejection rates across 1000 MC replicates. These rates are given in Table 1. Under the first two true link settings, all tests have rejection rates around 0.05, suggesting that all tests preserve the right size.

For the three proposed tests based on (un)balanced grouped responses, where there are multiple parameters for which each test statistic can be computed, we recommend to only use the test statistic associated with σ to avoid the issue of multiple testing. Alternatively, one can consider a quadratic-form test statistic based on the discrepancy

Table 1 Rejection rates of the four proposed tests and GOF test in Pan and Lin (2005) across 1000 Monte Carlo replicates when the true link is logit, probit, the cdf of a left-skewed mixture normal, and the cdf of a right-skewed mixture normal. The two entries in parentheses under $t_{4,\rho}$ for each true link setting are the rejection rates when $\rho = 0.05$ and 0.1, respectively

θ	$t_{1,\theta}$	$t_{2,\theta}$	$t_{3,\theta}$	$t_{4,\rho}$	GOF	$t_{1,\theta}$	$t_{2,\theta}$	$t_{3,\theta}$	$t_{4,\rho}$	GOF
	The true link is logit					The true link is probit				
σ	0.05	0.06	0.04	(0.05, 0.04)	0.04	0.05	0.05	0.06	(0.06, 0.05)	0.04
β_0	0.05	0.06	0.05			0.05	0.06	0.07		
β_1	0.04	0.04	0.05			0.05	0.07	0.06		
β_2	0.05	0.05	0.05			0.05	0.05	0.06		
β_3	0.05	0.04	0.06			0.05	0.05	0.05		
	The true link is left skewed					The true link is right skewed				
σ	0.08	0.67	0.57	(0.92, 0.78)	0.09	1.00	0.67	1.00	(0.61, 0.39)	1.00
β_0	0.10	0.61	0.59			1.00	0.84	1.00		
β_1	0.08	0.51	0.40			0.94	0.56	0.96		
β_2	0.07	0.13	0.13			0.49	0.26	0.61		
β_3	0.06	0.08	0.08			0.08	0.13	0.14		

vector, say, $\hat{\boldsymbol{\Omega}} - \hat{\boldsymbol{\Omega}}^*$, to combine evidence from individual $t_{1,\theta}$'s as a way to avoid multiple testing as done in Huang (2009). For a cleaner comparison between the first three tests and the fourth, we focus on the above t test statistics in this study. The following features of the operating characteristics of $t_{1,\sigma}$, $t_{2,\sigma}$, and $t_{3,\sigma}$ provide helpful guidance for practical implementation of these tests to check the adequacy of an assumed link.

The test based on $t_{1,\sigma}$ possesses high power to detect link misspecification when the true link is right skewed, whereas it shows little power when the true link is left skewed. Therefore, when one rejects the null due to a significant test based on $t_{1,\sigma}$, one may conclude not only that the assumed logit (or a symmetric) link is inadequate for the current data, but also that the true link is very likely to be right skewed. On the other hand, an insignificant value of $t_{1,\sigma}$ may not be enough to support an assumed symmetric link. In this case, one may follow up with a test based on $t_{2,\sigma}$, which shows in the simulation moderate or high power under both asymmetric true link settings. The asymmetric pattern in the power associated with $t_{1,\theta}$ can be explained by the phenomenon observed in Sect. 3 that the bias of $\hat{\theta}^*$ tends to be more substantial when the true link is right skewed than when it is left skewed. The more comparable power associated with $t_{2,\theta}$ when the direction of skewness of the true link switches relates to the symmetric pattern of $\hat{\theta}^{**}$ for asymmetric true links of opposite skewness directions. Lastly, the test based on $t_{3,\theta}$ has similar power as that of $t_{2,\theta}$ when the true link is left skewed, and is very comparable with $t_{1,\theta}$ when the true link is right skewed. In other words, $t_{3,\theta}$ is similar to whichever the "winner" is between $t_{1,\theta}$ and $t_{2,\theta}$ given any asymmetric true link setting. This is not surprising considering the most substantial discrepancy $\hat{\theta}^* - \hat{\theta}^{**}$ compared with the other discrepancies under different asymmetric true links we already observe in the numerical study in Sect. 3.

Table 2 Averages of the MLE of ρ, $\tilde{\rho}$, across 1000 MC replicates when the true link is logit, probit, a left-skewed mixture normal cdf (LS-MN), and a right-skewed mixture normal cdf (RS-MN), respectively. The empirical standard error associated with each of these averages is below 0.003

ρ	Logit	Probit	LS-MN	RS-MN
0.05	0.05	0.05	0.13	0.03
0.1	0.10	0.09	0.17	0.08

Because of this feature of $t_{3,\theta}$, if one is only interested in revealing the existence of some kind of link misspecification, one only needs to implement the test using $t_{3,\sigma}$. But if one also wishes to correct an assumed link, we recommend use of a sequential testing strategy, where one computes $t_{2,\sigma}$ if one fails to reject the null using $t_{1,\sigma}$.

The test based on $t_{4,\rho}$ shows promising power in both link misspecification settings, especially when the true link is left skewed, which is much higher than the power of the competing GOF test. Moreover, the sign of $t_{4,\rho}$ directly relates to the skewness of the true link because $\tilde{\rho}$ overestimates ρ when the true link is left skewed, resulting in a positive $t_{4,\rho}$, and it underestimates the truth when the true link is right skewed, resulting in a negative $t_{4,\rho}$. Table 2 provides the MC average of $\tilde{\rho}$ under each true link setting. To this end, one may use $t_{4,\rho}$ as an informative test for link misspecification, of which a significant positive/negative value provides evidence of a true link being left/right skewed. When implementing this test, we recommend a conservative choice of ρ so that the misclassification rate is low (such as 0.05 or 0.1) when generating the reclassified responses. With too high (i.e., too close to 0.5 from below) of a misclassification rate, the resulting reclassified responses may suffer too much information loss in the original data, which can compromise the power of $t_{4,\rho}$.

To this end, we have the assumed GLMM correctly specified except for the link function in the empirical study. This design of the numerical study is dictated by our focus in this article, which is link misspecification. In practice, one shall bear in mind that other assumptions involved in a GLMM may be inadequate as well, such as those on the random intercept and the linear predictor. For the fourth test with test statistic $t_{4,\rho}$, any source of model misspecification that leads to an inconsistent estimator for ρ can trigger a significant test result. As for the first three proposed tests, which are based on the comparison between two MLEs resulting from two related data sets, they can also return significant values in the presence of other forms of model misspecification whenever the two considered MLEs are affected by the model misspecification(s) differently in limit. In fact, Yu and Huang (2017) used $t_{1,\theta}$ to assess distributional assumptions on the random intercept in a GLMM. There the authors showed that, besides promising power of the tests when the true random intercept distribution is skewed as opposed to an assumed normal, $t_{1,\sigma}$ can also reveal the direction of skewness of the true distribution. Like our study here, their study also focused on one source of misspecification at a time. We conjecture that, like $t_{1,\theta}$, $t_{2,\theta}$, $t_{3,\theta}$, and $t_{4,\rho}$ are sensitive to certain types of misspecification on the random intercept as well. When it is only the form of the linear predictor that is misspecified, we repeat part of the simulation studies with the raw data generated from a GLMM of which the conditional mean model given by $E(Y_{ij}|\mathbf{X}_i, b_{i0}; \boldsymbol{\beta}, \tau) = h_0(\beta_0 + \beta_1 X_{ij,1} + \beta_2 X_{ij,2}^2 + \beta_3 X_{ij,1} X_{ij,2} + b_{i0})$, and the true link being logit. Then we compute MLEs based on the raw data and induced data assuming a GLMM with logit link and $E(Y_{ij}|\mathbf{X}_i, b_{i0}; \boldsymbol{\beta}, \tau) = h_0(\beta_0 + \beta_1 X_{ij,1} + \beta_2 X_{ij,2} + \beta_3 X_{ij,1} X_{ij,2} + b_{i0})$. In the presence of this (mild) misspecification on the linear predictor, only the fourth test $t_{1,\rho}$ exhibits moderate power to conclude rejection of the null, with a rejection rate of 43% across 1000 MC replicates at the significance level of 0.05. The rejection rates associated with the other three tests, $t_{1,\theta}$, $t_{2,\theta}$, $t_{3,\theta}$, all remain around the nominal level of 0.05 in this case when θ takes any one of the five parameters in $\boldsymbol{\Omega}$.

Systematic investigations on the operating characteristics of the four proposed tests in the presence of other sources of model misspecification are beyond the scope of the current study. But, based on our preliminary empirical studies (partly described above), we conjecture that $t_{4,\rho}$ can be more responsive than the other three tests whenever there exists some form(s) of model misspecification, whereas the first three tests enjoy certain level of robustness to linear predictor misspecification because of the similarity of the two MLEs under comparison within a test statistic when this is the only source of misspecification.

6 Real data application

We now revisit the longitudinal data from the Indonesian children's health study on respiratory infection (Sommer et al. 1984) analyzed in Pan and Lin (2005). In Pan and Lin (2005), the authors applied their GOF test to search for some improvement in the functional form of the linear predictor over the initial assumed logit GLMM given by

$$E(Y_{ij}|\mathbf{X}_i, b_{i0}; \boldsymbol{\beta}) = \{1 + \exp(-\beta_0 - \beta_1 X_{ij,1} - \beta_2 X_{ij,2} - \beta_3 X_{ij,3} - \beta_4 X_{ij,4} - \beta_5 X_{ij,5} - \beta_6 X_{ij,6} - \beta_7 X_{ij,6}^2 - b_{i0}\}^{-1}, \tag{2}$$

where Y_{ij} is the indicator of child i suffering from respiratory infection at the jth occasion when this child was examined, $X_{ij,1}$ is the only between-cluster covariate in the model, which refers to the gender of child i, $X_{ij,2}$ is height for age as a percentage of the United States National Center for Health Statistics standards, $X_{ij,3}$ and $X_{ij,4}$ are annual cosine and sine variables to adjust for seasonality, $X_{ij,5}$ is an indicator taking value 1 if child i is suffering from xerophthalmia at the jth occasion, and $X_{ij,6}$ is the centered age. Like our study throughout this article, they assumed $b_{i0} \sim N(0, \sigma^2)$. Unlike our analyses presented in this section, they focused on testing the adequacy of the functional form in which the covariates enter the linear predictor.

For illustration purposes, we use a subset of the data analyzed in Pan and Lin (2005) with 122 children who had six records. Preserving the functional form of the linear predictor in (2), we apply the following four proposed tests to assess the adequacy of the logit link, $t_{1,\sigma}$, $t_{2,\sigma}$, $t_{3,\sigma}$, and $t_{4,\rho}$. When implementing the first three tests, we partition the six records within each child into two groups of equal size after these six records are sorted by $X_{ij,2}$ and $X_{ij,6}$. This sorting before partitioning each cluster yields higher across-groups variability associated with the within-cluster covariates, which increases the efficiency of the MLEs and typically leads to more powerful tests compared to when the groups are formed randomly within a cluster. For the test based on $t_{4,\rho}$, we repeat it twice, one with $\rho = 0.05$ and the other with $\rho = 0.1$. The p values associated with the first three tests are 0.30, 0.02, and 0.06, respectively. Hence, the test based on $t_{1,\sigma}$ is insignificant, whereas the test using $t_{2,\sigma}$ is clearly much more significant than the first test and so is the test using $t_{3,\sigma}$. This is the pattern of how these three tests compared observed in Sect. 5 when the true link is left skewed. Additionally, the two tests based on $t_{4,\rho}$ yield highly significant positive values of the test statistic, with p values both less than 10^{-4}. This is yet another indication that the logit link is inadequate for the current data set, and a left-skewed link is likely to yield a better fit for the data. Although in Pan and Lin (2005) the authors did not consider testing the link function, they did find strong evidence from their GOF test that the functional form of the linear predictor in (2) is problematic. In conclusion, our tests, as well as their GOF test, suggest lack of fit of the initial logit GLMM model, and the next step is to improve this model, either by using a flexible link function that allows negative skewness, such as a generalized logit link (Stukel 1988), or by attempting different functional forms for the linear predictor as pursued in Pan and Lin (2005). We do not sidetrack to modeling using a flexible link in this analysis.

7 Discussion

In this study, we propose four diagnostic tests for link misspecification following two theme ideas, both of which involve creating induced responses from the original responses. The first theme idea, which leads to three proposed tests, is the same as that in Huang (2009). In both works, the discrepancy between two MLEs of a parameter resulting from two related data sets serves as an indicator of model misspecification. The contribution of this work is that we use this idea to achieve an informative diagnostic test for link misspecification, whereas Huang (2009) concerns random-effects assumptions in GLMM. More importantly, besides the unbalanced grouped responses, which is the only type of induced responses considered in Huang (2009) in developing diagnostic methods, we also exploit the balanced grouped responses. This new addition leads to a test, $t_{2,\theta}$, that has impressive power to detect link misspecification when the true link deviates from the assumed symmetric link in either direction; it also motivates a third test, $t_{3,\theta}$, that has competitive power when the true link deviates from an assumed symmetric link from either direction. Another advantage of using balanced group responses is that, compared to the unbalanced group responses, it is less likely to have the induced grouped responses to be all one's (or all zero's), a situation making sensible maximum likelihood estimation infeasible. The nature of this first theme idea shares some similarity with that in Agresti and Caffo (2002), where the authors constructed descriptive measures of relative model fit based on the comparison between two geometric means or means of the likelihood functions associated with two models. The second theme idea leads to a test that only requires one MLE of an extraneous parameter whose truth is known to data analysts. Using the difference between the MLE of this extraneous parameter and its truth as an indicator of link misspecification not only allows direct validation of an assumed link, but also provides information on the skewness direction of the true link when the assumed symmetric link is rejected.

We compare our tests with the residual-based GOF test proposed by Pan and Lin (2005) and observe higher or comparable power from at least two of our proposed tests compared to the power of their test in all misspecification scenarios considered. In terms of computation, our methods only require routine maximum likelihood estimation, with test statistics follow some t distributions asymptotically under the null, whereas the null distribution of their test statistic is far more complex, and consequently, time-consuming bootstrap procedures are needed to approximate the corresponding p value. For a data set of the size as that in the simulation studies in Sect. 5, that is, $m = 600$ and $n_i = 6$, it takes 190.3 s to implement the GOF test, whereas it takes 45.6, 99.4, 112.7, and 59.8 s, respectively, to obtain test results from $t_{1,\theta}$, $t_{2,\theta}$, $t_{3,\theta}$, and $t_{4,\rho}$ when implemented separately using the R codes created by the first author on a Dell XPS 934 with Core i7 processor and 2.40 GHz CPU.

When implementing the tests based on grouped responses, we recommend sort the observations within each cluster according to the within-cluster covariate(s) before dividing each cluster into groups in order to produce less variable parameter estimates, which can in turn yield more efficient tests. Throughout the simulation study and data analyses we create two groups per cluster for simplicity. In the presence of link misspecification, the two MLEs under comparison in the first three test statistics tend to differ more (in limit for large samples) when the number of groups in a cluster, $g_i (\geq 2)$, is much smaller than the cluster size, n_i. But the grouped data MLEs are more variable with a smaller g_i, causing a larger denominator of the test statistics. Due to this trade-off between the numerator and denominator of the first three test statistics, the choice of g_i is secondary and should not substantially affect the power of the tests. When implementing the test based on reclassified responses, we suggest one use a misclassification probability ρ much lower than 0.5 to guard against too much information loss in the reclassified responses compared to the original responses. One may also consider experimenting a grid of ρ's over the range (0, 0.5) and defines the supremum of $|t_{4,\rho}|$ as the test statistic to achieve a more powerful test, although this test does not have a null distribution as simple as the one we use here.

We consider a symmetric link under the null hypothesis throughout the study. Under such null formulation, we are able to extract information from the proposed test statistics regarding the skewness direction of an underlying link when the null is rejected. If one already assumes an asymmetric link under the null, some of the tests should still have power to detect an inadequate assumed link as long as the misspecification leads to discrepancy between $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Omega}}^*$ or $\hat{\boldsymbol{\Omega}}^{**}$, or the misspecification results in an inconsistent MLE of ρ. But, when the null is rejected in this case, what information one can extract from these test statistics relating to how the asymmetric link is misspecified is less clear. Lastly, we assume in this study correct modeling for the random intercept. The operating characteristics of the proposed tests in the presence of random-effects misspecification in addition to link misspecification are unclear. We do not think the tests based on grouped responses can distinguish these two sources of model misspecification, although we are hopeful that the test based on reclassified response generated according to a more strategically designed misclassification model can disentangle different sources of model misspecification. These are within the scope of our follow-up research topics.

References

Agresti A, Caffo B (2002) Measures of relative model fit. Comput Stat Data Anal 39:127–136

Alonso A, Litière S, Molenberghs G (2008) A family of tests to detect misspecifications in random-effects structure of generalized linear mixed models. Comput Stat Data Anal 52:4474–4486

Aranda-Ordaz FJ (1981) On two families of transformations to additivity for binary response data. Biometrika 68:357–363

Azzalini A (1985) A class of distributions which includes the normal ones. Scand J Stat 12:171–178

Boos DD, Stefanski LA (2013) Essential statistical inference, theory and methods. Springer, New York

Caffo B, Ming-Wen A, Rohde C (2007) Flexible random intercept models for binary outcomes using mixtures of normals. Comput Stat Data Anal 51:5220–5235

Chambers E, Cox D (1967) Discrimination between alternative binary response models. Biometrika 67:250–251

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Czado C, Santner TJ (1992) The effect of link misspecification on binary regression inference. J Stat Plan Inference 33:213–231

Dorfman R (1943) The detection of defective members of large populations. Ann Math Stat 14:436–440

Guerrero VM, Johnson RA (1982) Use of the Box–Cox transformation with binary response models. Biometrika 69:309–314

Huang X (2009) Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. Biometrics 65:361–368

Huang X (2011) Detecting random-effects model misspecification via coarsened data. Comput Stat Data Anal 55:703–714

Huang X (2013) Tests for random effects in linear mixed models using missing data. Stat Sinica 23:1043–1070

Jiang X, Dey D, Prunier R, Wilson A, Holsinger K (2013) A new class of flexible link functions with application to species co-occurrence in cape floristic region. Ann Appl Stat 7:2180–2204

Kim S, Chen M, Dey D (2008) Flexible generalized t-link models for binary response data. Biometrika 95:93–106

Li K, Duan N (1989) Regression analysis under link violation. Ann Stat 17:1009–1052

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall/CRC, London

Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. Springer series in statistics. Springer, New York

Morgan B (1983) Observations on quantitative analysis. Biometrics 39:879–886

Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A 135:370–384

Neuhaus JM (2002) Analysis of clustered and longitudinal binary data subject to response misclassification. Biometrics 58:675–683

Neuhaus JM, Hauck WW, Kalbfleisch JD (1992) The effects of mixture distribution specification when fitting mixed-effects logistic models. Biometrics 79:755–762

Owen DB (1956) Tables for computing bivariate normal probabilities. Ann Math Stat 27:1075–1090

Pan Z, Lin DY (2005) Goodness-of-fit methods for generalized linear mixed models. Biometrics 61:1000–1009

Ritz C (2004) Goodness-of-fit tests for mixed models. Scand J Stat 31:443–458

Samejima F (2000) Logistic positive exponent family of models: virtue of asymmetric item characteristic curves. Psychometrika 65:319–335

Sommer A, Katz J, Tarwotjo I (1984) Increased risk of respiratory infection and diarrhea in children with preexisting mild vitamin A deficiency. Am J Clin Nutr 40:1090–1095

Stukel T (1988) Generalized logistic models. J Am Stat Assoc 83:426–431

Tchetgen EJ, Coull BA (2006) A diagnostic test for the mixing distribution in a generalised linear mixed model. Biometrika 93:1003–1010

Verbeke G, Molenberghs G (2013) The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. Biostatistics 14:477–490

Waagepetersen R (2006) A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. Scand J Stat 33:721–731

Wang YH (1993) On the number of successes in independent trials. Stat Sinica 3:295–312

Wang Z, Louis TA (2003) Matching conditional and marginal shapes in binary mixed-effects models using a bridge distribution function. Biometrika 90:765–775

Wang Z, Louis TA (2004) Marginalized binary mixed-effects with covariate-dependent random effects and likelihood inference. Biometrics 60:884–891

White H (1981) Consequence and detection of misspecified nonlinear regression models. J Am Stat Assoc 76:419–433

Whittemore A (1983) Transformations to linearity in binary regression. J Appl Math 43:703–710

Yu S, Huang X (2017) Random-intercept misspecification in generalized linear mixed models for binary responses. Stat Methods Appl 26:333–359