Improved wrong-model inference for generalized linear models for binary responses in the presence of link misspecification

Xianzheng Huang

Statistical Methods & Applications
Journal of the Italian Statistical Society

ISSN 1618-2510

Stat Methods Appl
DOI 10.1007/s10260-020-00529-3

ONLINE FIRST

Springer

Statistical Methods & Applications
https://doi.org/10.1007/s10260-020-00529-3

ORIGINAL PAPER

Improved wrong-model inference for generalized linear models for binary responses in the presence of link misspecification

Xianzheng Huang[1]

Accepted: 18 May 2020

Abstract

In the framework of generalized linear models for binary responses, we develop parametric methods that yield estimators for regression coefficients less compromised by an inadequate posited link function. The improved inference are obtained without correcting a misspecified model, and thus are referred to as wrong-model inference. A byproduct of the proposed methods is a simple test for link misspecification in this class of models. Impressive bias reduction in estimators for the regression coefficients from the proposed methods and promising power of the proposed test to detect link misspecification are demonstrated in simulation studies. We also apply these methods to a classic data example frequently analyzed in the existing literature concerning this class of models.

Keywords Bias · Binary response · Logistic regression · Model misspecification · Reclassification

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 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. In this study we focus on GLM for a binary response. The most popular GLM for binary responses assume a logit link, or probit,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10260-020-00529-3) contains supplementary material, which is available to authorized users.

✉ Xianzheng Huang
huang@stat.sc.edu

[1] University of South Carolina, Columbia, SC, USA

Published online: 06 June 2020

complementary log-log, etc., mainly due to ease of interpretation and convenient implementation using standard software. However, it has come to practitioners' attention that a symmetry link, such as logit and probit, may not be reasonable in many applications; and the asymmetric complementary log-log link only allows a fixed negative skewness, making it too rigid for many scenarios. Many theoreticians concur with this concern regarding routine use of these popular links. For instance, Czado and Santner (1992) considered GLM for binary responses and showed that the maximum likelihood estimators (MLE) of regression coefficients obtained under an inappropriate link can be biased and inefficient.

There are two ways to avoid an inadequate link function. The more explored way is employing a flexible class of link functions (Aranda-Ordaz 1981; Guerrero and Johnson 1982; Morgan 1983; Whittemore 1983; Stukel 1988; Kim et al. 2007; Jiang et al. 2013). With a nonstandard link involved, inference results from methods along this line are usually harder to interpret than those from methods that use a standard link function. But, when a standard link is inadequate, methods adopting flexible links can better preserve certain integrity of covariate effects. This is especially important when the question of interest is whether or not there exists a significant covariate effect. Another way is to stick to one of the routinely used links, 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 simple form of the linear predictor in GLM is desirable for meaningful interpretations of a covariate effect. This is the case in, for example, models in the item response theory as discussed in Samejima (2000). There, the author showed that the MLE for regression coefficients 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 without revising the functional form of the linear predictor.

In this study, we develop parametric methods to achieve estimators for regression coefficients less compromised by link misspecification without correcting the assumed link or adopting a nonlinear predictor. Similar to methods that employ flexible link functions, the main benefit of the proposed methods is to avoid distorting inference for covariate effects due to a misspecified link, even though one sacrifices simple interpretation for the estimated covariate effects. To gain insight on the impact of link misspecification on regression coefficients estimation, we investigate asymptotic bias in the MLE for regression coefficients in the presence of link misspecification in Sect. 2. Results from the bias analysis motivate the first proposed bias reduction method we present in Sect. 3, followed by a second proposed method that also leads to substantial bias reduction in the MLE for regression coefficients in the presence of link misspecification. Section 4 reports simulation studies designed to illustrate the implementation and performance of the proposed methods. In Sect. 5, a simple yet powerful test for link misspecification is developed using byproducts of the proposed estimation methods, which unifies parameter estimation and model verification in one round of analysis based on maximum likelihood. The bias reduction methods and the new test for link misspecification are applied to a classic real-life data example in Sect. 6. Finally, we address some practical considerations, refinement of the proposed methods, and future research agenda in Sect. 7.

2 Effects of local link misspecification

2.1 Models and data

Suppose that one has a random sample consisting of n realizations of the response-covariate pair, (Y, X), and the mean of the binary response Y given X is $E(Y_i|X_i) = G(\beta_0 + \beta_1 X_i)$, for $i = 1, \ldots, n$, where the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ are of central interest, and $G(t)$ is a non-decreasing differentiable link function. Denote by $H(t)$ the link one assumes for the mean model $E(Y_i|X_i)$ that differs from $G(t)$. For notational simplicity, we assume a scalar covariate in this section. Generalization to multivariate regression models will be addressed in the next two sections.

Denoted by $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \tilde{\beta}_1)^{\mathrm{T}}$ the naive MLE for $\boldsymbol{\beta}$ resulting from the misspecified model. To obtain an estimator less biased than $\tilde{\boldsymbol{\beta}}$, we propose two strategies making use of a reclassified response Y^* generated according to

$$P(Y^* = Y|Y, X) = \pi. \qquad (1)$$

In principle, one may let π depend on (Y, X). For simplicity, a constant π is used in the sequel. Combining (1) and the assumed GLM, one has the assumed model for Y^* given X specified by $E(Y_i^*|X_i) = (2\pi - 1)H(\beta_0 + \beta_1 X_i) + 1 - \pi$, for $i = 1, \ldots, n$. It has been shown that π and $\boldsymbol{\beta}$ are identifiable using the reclassified data $\{(Y_i^*, X_i)\}_{i=1}^n$ when $\pi \neq 0.5$ (Carroll et al. 2006, Section 15.3). Denote by $\hat{\pi}$ and $\hat{\boldsymbol{\beta}}(\pi)$ the resulting MLEs for π and $\boldsymbol{\beta}$ based on the assumed model, respectively.

Since π is a known constant in the user-designed reclassification model, one can literally see the finite sample bias in $\hat{\pi}$. Moreover, $\hat{\pi}$ and $\hat{\boldsymbol{\beta}}(\pi)$ are entwined in the sense that the bias in one estimator correlates with the bias in the other estimator. This connection is the gateway to an estimator for $\boldsymbol{\beta}$ that is less biased than $\tilde{\boldsymbol{\beta}}$. We develop the first bias reduction strategy by exploiting this connection explicitly. Our second proposed strategy makes use of this connection implicitly. In particular, the first strategy is developed based on the following bias analysis of $\hat{\boldsymbol{\beta}}(\pi)$ under mild link misspecification.

2.2 Asymptotic bias

Denote by p and $\boldsymbol{b} = (b_0, b_1)^{\mathrm{T}}$ the limiting MLEs for π and $\boldsymbol{\beta}$, respectively, under the assumed model based on the reclassified data as $n \to \infty$. According to the theories of MLEs resulting from misspecified models (White 1982), under regularity conditions, $(p, \boldsymbol{b})$ is the point in the parameter space associated with $(\pi, \boldsymbol{\beta})$ where the Kullback-Leibler distance between the true model likelihood and the assumed model likelihood for the reclassified data is minimized. Equivalently, $(p, \boldsymbol{b})$ solves the following normal score equations, which we derive in the “Appendix”,

$$E\left(\frac{\mu_0-\mu}{\mu(1-\mu)}\begin{bmatrix}2H(\eta)-1\\H'(\eta)\\H'(\eta)X\end{bmatrix}\right)=\boldsymbol{0}, \tag{2}$$

where the expectation is with respect to the distribution of X, $\eta=b_0+b_1X$, $H'(\eta)=(d/d\eta)H(\eta)$,

$$\mu=(2p-1)H(\eta)+1-p, \tag{3}$$

is the mean of Y^* given X under the assumed model evaluated at the limiting MLEs, (p,b_0,b_1), and

$$\mu_0=(2\pi-1)G(\eta_0)+1-\pi \tag{4}$$

is the mean of Y^* given X under the correct model evaluated at the true parameter values, (π,β_0,β_1), in which $\eta_0=\beta_0+\beta_1X$.

To gain insight on the properties of $\boldsymbol{b}$, we first consider a local model misspecification where the link misspecification is mild. More specifically, suppose that the true link relates to the assumed link via

$$G(t)=(1-\epsilon)H(t)+\epsilon G_c(t), \tag{5}$$

for some small $\epsilon\in[0,1]$, where $G_c(t)$ can be interpreted as the contamination link. In the presence of outliers in data, (5) can be interpreted as that, had there been no outliers (corresponding to $\epsilon=0$), $H(t)$ would be an adequate link function in a GLM characterizing the data; with more extreme outliers (corresponding to a larger ϵ), one has to modify one's favorite link function, such as the logit link, to construct a less popular link $G(t)$ in order to better capture the data with outliers as a whole. Following this interpretation, one essentially alleviates influence of outliers on regression coefficient estimation when one reduces bias in $\tilde{\boldsymbol{\beta}}$, and the resultant bias-reduced covariate effect estimators can be interpreted in the same way as if the logit link, or other assumed $H(t)$ one chooses, were the true link for outlier-free data.

Under the formulation of (5), $G_c(t)$ and ϵ together control the discrepancy between the true link and the assumed link. To signify the dependence of p and $\boldsymbol{b}$ on the severity of link misspecification and the level of data coarsening, one may view these limiting MLEs as functions of ϵ and π, denoted by $p(\epsilon,\pi)$ and $\boldsymbol{b}(\epsilon,\pi)$, respectively. Because setting $\epsilon=0$ in (5) gives $G(t)=H(t)$, i.e., the case without link misspecification, one has $p(0,\pi)=\pi$ and $\boldsymbol{b}(0,\pi)=\boldsymbol{\beta}$. With a small ϵ in the presence of a mild link misspecification, we consider a first order Taylor expansion of the two elements in $\boldsymbol{b}(\epsilon,\pi)$, $b_0(\epsilon,\pi)$ and $b_1(\epsilon,\pi)$, around $\epsilon=0$,

$$\begin{aligned}b_0(\epsilon,\pi)&=\beta_0+b_0'(0,\pi)\epsilon+o(\epsilon),\\b_1(\epsilon,\pi)&=\beta_1+b_1'(0,\pi)\epsilon+o(\epsilon),\end{aligned} \tag{6}$$

where $b_0'(0,\pi)$ is equal to $(\partial/\partial\epsilon)b_0(\epsilon,\pi)$ evaluated at $\epsilon=0$, which can be interpreted as a first order bias factor associated with b_0, and thus an asymptotic first order bias factor associated with $\hat{\beta}_0(\pi)$; $b_1'(0,\pi)$ is similarly defined and has a similar interpretation relating to b_1 and $\hat{\beta}_1(\pi)$. We next derive these two bias factors by exploring an approximated solution to (2).

Solving (2) for $(p, \boldsymbol{b})$ cannot be done explicitly in general. To simplify the equations to be solved, we assume that $(p, \boldsymbol{b})$ is a point in the parameter space such that $\mu_0 - \mu = 0$ with probability one over the support of X. Such point may not exist in the parameter space except in some special model settings. This assumption is made for the sole purpose of envisioning an approximated solution to (2) that allows one to obtain some approximated first order bias factors in (6), which can shed some light on the effects of link misspecification. Following this assumption and defining $\delta = p - \pi$, one has

$$\begin{aligned}0 &= \mu_0 - \mu \\ &= (2\pi - 1)G(\eta_0) + 1 - \pi - \{(2p-1)H(\eta) + 1 - p\}, \text{ by (3) and (4)}, \\ &= (2\pi - 1)\{(1-\epsilon)H(\eta_0) + \epsilon G_c(\eta_0) - H(\eta)\} - 2\delta H(\eta) + \delta, \text{ by (5)}, \\ &= (2\pi - 1)[\epsilon\{G_c(\eta_0) - H(\eta_0)\} + H(\eta_0) - H(\eta)] + \{1 - 2H(\eta)\}\delta.\end{aligned} \tag{7}$$

Bearing in mind the dependence of $(p, \boldsymbol{b})$ on (ϵ, π), we differentiate (7) with respect to ϵ to yield

$$\begin{aligned}0 = (2\pi - 1)\{G_c(\eta_0) - H(\eta_0)\} + \{1 - 2H(\eta)\}\frac{\partial p(\epsilon, \pi)}{\partial \epsilon} + H'(\eta)\bigg\{\frac{\partial b_0(\epsilon, \pi)}{\partial \epsilon} \\ + \frac{\partial b_1(\epsilon, \pi)}{\partial \epsilon}X\bigg\}(1 - 2p).\end{aligned} \tag{8}$$

Setting $\epsilon = 0$ in (8) gives

$$\begin{aligned}0 = (2\pi - 1)\{G_c(\eta_0) - H(\eta_0)\} + \{1 - 2H(\eta_0)\}p'(0, \pi) \\ + H'(\eta_0)\{b_0'(0, \pi) + b_1'(0, \pi)X\}(1 - 2\pi),\end{aligned} \tag{9}$$

where $p'(0, \pi)$ is equal to $(\partial/\partial\epsilon)p(\epsilon, \pi)$ evaluated at $\epsilon = 0$. Now that it is assumed that (9) holds with probability one over the support of X, one may evaluate X at any value in the support in this equation. Suppose that the support contains zero and one. By first setting $X = 0$ and then setting $X = 1$ in (9), one obtains the two bias factors in (6) given by

$$b_0'(0, \pi) = \frac{G_c(\beta_0) - H(\beta_0)}{H'(\beta_0)} + \frac{\{1 - 2H(\beta_0)\}p'(0, \pi)}{(2\pi - 1)H'(\beta_0)}, \tag{10}$$

$$b_1'(0, \pi) = \frac{G_c(\beta_0 + \beta_1) - H(\beta_0 + \beta_1)}{H'(\beta_0 + \beta_1)} + \frac{\{1 - 2H(\beta_0 + \beta_1)\}p'(0, \pi)}{(2\pi - 1)H'(\beta_0 + \beta_1)} - b_0'(0, \pi). \tag{11}$$

These first order bias factors can reveal some effects of link misspecification on the MLE for $\boldsymbol{\beta}$. For instance, if $\beta_0 = 0$ and $H(t)$ is a symmetric link, then (10) reduces to $b_0'(0, \pi) = \{G_c(0) - 0.5\}/H'(0)$, for all π. This simple result of $b_0'(0, \pi)$ suggests that, if $G_c(t)$ is left-skewed (right-skewed), then $b_0'(0, \pi) < 0 (> 0)$, and thus b_0

tends to be smaller (bigger) than the truth. If $G(t)$ is also symmetric, which means that $G_c(t)$ has to be symmetric unless $\epsilon = 0$, then $b_0'(0,\pi) = 0$, suggesting that the asymptotic bias in $\hat{\beta}_0(\pi)$ is of order $o(\epsilon)$ for all π. These patterns of b_0 are indeed observed for $\hat{\beta}_0(\pi)$, as well as $\tilde{\beta}_0$, in our simulation study. Moreover, according to (11), $\beta_1 = 0$ implies $b_1'(0,\pi) = 0$ for all π. This is in line with the well established fact that $\beta_1 = 0$ implies $b_1 = 0$ even in the presence of link misspecification. Besides these implications on the direction of bias in $\hat{\boldsymbol{\beta}}(\pi)$ and $\tilde{\boldsymbol{\beta}}$, these bias factors along with (6) reveal a bias correction method we elaborate next.

3 Bias reduction using reclassified data

3.1 Explicit bias reduction

Evaluating (10) at two different values of π, π_1 and π_2, and forming the difference between the two resultant equations yields

$$b_0'(0,\pi_1) - b_0'(0,\pi_2) = \frac{\{1-2H(\beta_0)\}}{H'(\beta_0)} \left\{ \frac{p'(0,\pi_1)}{2\pi_1 - 1} - \frac{p'(0,\pi_2)}{2\pi_2 - 1} \right\},$$

hence, by (6),

$$\begin{aligned} b_0(\epsilon,\pi_1) - b_0(\epsilon,\pi_2) \approx & \frac{\{1-2H(\beta_0)\}}{H'(\beta_0)} \left\{ \frac{p'(0,\pi_1)}{2\pi_1 - 1} - \frac{p'(0,\pi_2)}{2\pi_2 - 1} \right\}\epsilon \\ \approx & R(\beta_0)\left(\frac{p_1 - \pi_1}{2\pi_1 - 1} - \frac{p_2 - \pi_2}{2\pi_2 - 1} \right), \end{aligned} \tag{12}$$

where $R(\beta_0) = \{1 - 2H(\beta_0)\}/H'(\beta_0)$, $p_k = p(\epsilon, \pi_k)$, for $k = 1, 2$, and the substitution leading to the last equation is based on a first order Taylor expansion of $p(\epsilon,\pi)$ around $\epsilon = 0$, $p(\epsilon,\pi) = \pi + p'(0,\pi)\epsilon + o(\epsilon)$. Inspired by (12), we propose the following estimator for β_0,

$$\hat{\beta}_0^{(1)} = R^{-1}\left[\left\{ \hat{\beta}_0(\pi_1) - \hat{\beta}_0(\pi_2) \right\} \left(\frac{\hat{\pi}_1 - \pi_1}{2\pi_1 - 1} - \frac{\hat{\pi}_2 - \pi_2}{2\pi_2 - 1} \right)^{-1} \right], \tag{13}$$

where $\hat{\beta}_0(\pi_k)$ is the MLE for β_0 under the assumed model based on the reclassified data with $\pi = \pi_k$, for $k = 1, 2$, and $R^{-1}(\cdot)$ is the inverse function of $R(\beta_0)$. For instance, if the assumed link $H(t)$ is the logit function, then $R^{-1}(s) = \log 2 - \log(\sqrt{s^2 + 4} + s)$.

Using (11) and following similar derivations that inspire $\hat{\beta}_0^{(1)}$, we construct the following estimator for β_1,

$$\hat{\beta}_1^{(1)} = R^{-1}\left[\left\{\hat{\beta}_0(\pi_1) - \hat{\beta}_0(\pi_2) + \hat{\beta}_1(\pi_1) - \hat{\beta}_1(\pi_2)\right\}\left(\frac{\hat{\pi}_1 - \pi_1}{2\pi_1 - 1} - \frac{\hat{\pi}_2 - \pi_2}{2\pi_2 - 1}\right)^{-1}\right] - \hat{\beta}_0^{(1)}, \tag{14}$$

where $\hat{\beta}_1(\pi_k)$ is the MLE for β_1 under the assumed model based on the reclassified data with $\pi = \pi_k$, for $k = 1, 2$.

How effectively the first proposed estimator $\hat{\boldsymbol{\beta}}^{(1)} = (\hat{\beta}_0^{(1)}, \hat{\beta}_1^{(1)})^{\mathrm{T}}$ reduces bias in $\tilde{\boldsymbol{\beta}}$ depends on how severe the link misspecification is, since the construction of $\hat{\boldsymbol{\beta}}^{(1)}$ originates from the Taylor expansion around $\epsilon = 0$ in (6). In addition, $\hat{\boldsymbol{\beta}}^{(1)}$ is derived based on the assumption that the solution to (2) solves a much simpler equation, $\mu_0 - \mu = 0$. This assumption allows us to derive the approximated bias factors in (10) and (11) without directly finding or approximating the solution to (2). If the so-obtained bias factors are misleading representations of the direction or magnitude of the true bias, $\hat{\boldsymbol{\beta}}^{(1)}$ can be more biased than $\tilde{\boldsymbol{\beta}}$. This can happen, for example, when X is a vector covariate, making the assumption that $\mu_0 - \mu = 0$ with probability one over the support of X further from reality. However, when X is a scalar whose support contains zero and one, $\hat{\boldsymbol{\beta}}^{(1)}$ can substantially improve over $\tilde{\boldsymbol{\beta}}$ as evidenced in the simulation study in Sect. 4.

3.2 Implicit bias reduction

The connection between $\hat{\boldsymbol{\beta}}(\pi)$ and $\hat{\pi}$ is only marginally exploited in the first proposed estimator because $\hat{\boldsymbol{\beta}}^{(1)}$ only uses two levels of reclassification, π_1 and π_2. More substantial bias reduction can be achieved by more fully exploiting the relationship between $\hat{\boldsymbol{\beta}}(\pi)$ and $\hat{\pi}$, or, equivalently, the connection between $\hat{\boldsymbol{\beta}}(\pi)$ and $d = \hat{\pi} - \pi$. Instead of viewing $\hat{\boldsymbol{\beta}}$ as a function of π, now it is more helpful to view it as a function of d, writing it as $\hat{\boldsymbol{\beta}}(d)$. Since π is a user-specified parameter in the reclassification model, one can empirically explore the connection between $\hat{\boldsymbol{\beta}}(d)$ and d by setting π at a sequence of K values over (0.5, 1), denoted by $\{\pi_k\}_{k=1}^K$, and computing $d_k = \hat{\pi_k} - \pi_k$ and $\hat{\boldsymbol{\beta}}(d_k)$ for each $k \in \{1, \ldots, K\}$. Using the sequence, $\{\hat{\boldsymbol{\beta}}(d_k), d_k\}_{k=1}^K$, one may apply an extrapolant on $\hat{\boldsymbol{\beta}}(d)$ to extrapolate to $\hat{\boldsymbol{\beta}}(0)$. This extrapolation is intuitively sensible if one believes that $\hat{\pi}$ is inconsistent for π unless the model for Y^* given X is correctly specified, in which case $\hat{\boldsymbol{\beta}}(d)$ is also consistent for $\boldsymbol{\beta}$. In what follows, we summarize this proposed method in an algorithm that leads to our second proposed estimator for $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}^{(2)} = (\hat{\beta}_0^{(2)}, \hat{\beta}_1^{(2)})^{\mathrm{T}}$.

RC-1 For each $(j, k) \in \{1, \ldots, J\} \times \{1, \ldots, K\}$, generate reclassified responses $\{Y_{ijk}^*, i = 1, \ldots, n\}_{j=1}^J$ according to (1) with $\pi = \pi_k$.

RC-2 For each $(j, k) \in \{1, \ldots, J\} \times \{1, \ldots, K\}$, compute the MLEs for π and $\boldsymbol{\beta}$ based on data $\{(Y_{ijk}^*, X_i)\}_{i=1}^n$, resulting in estimates denoted by $\hat{\pi}_{k,j}$ and $\hat{\boldsymbol{\beta}}_{k,j}$.

Compute $\hat{\pi}_k = J^{-1} \sum_{j=1}^{J} \hat{\pi}_{k,j}$, $d_k = \hat{\pi}_k - \pi_k$, and $\hat{\boldsymbol{\beta}}(d_k) = J^{-1} \sum_{j=1}^{J} \hat{\boldsymbol{\beta}}_{k,j}$, for $k = 1, \ldots, K$.

RC-3 View $\{\hat{\boldsymbol{\beta}}(d_k),\, d_k\}_{k=1}^{K}$ as K realizations of the response-predictor pair $(\hat{\boldsymbol{\beta}}(d), d)$. Use these realizations to carry out regression analysis assuming a user-specified regression function.

RC-4 Use the regression results from RC-3 to extrapolate the response $\hat{\boldsymbol{\beta}}(d)$ at $d = 0$, leading to the proposed estimate for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}^{(2)}$.

Figure 1 gives a pictorial illustration of the rationale behind this implicit bias reduction method. To produce plots in the upper panels, we fix the true model, from which a data set of size $n = 600$ is generated, as a GLM with the link $G(t) = 1/[1 + \exp\{-h(t)\}]$, in which $h(t) = 10\{\exp(\alpha_1 t) - 1\}I(t \geq 0) - 10\log(1 + \alpha_2 t)I(t < 0)$ with $(\alpha_1, \alpha_2) = (0.1, -0.1)$. This link function is depicted (as the dashed curve) in contrast to the logit link (as the solid curve) in Fig. 2. It is an example of generalized logit links (Stukel 1988) to be introduced more formally in Sect. 4. The true values of the regression coefficients are $\boldsymbol{\beta} = (-1, 1)^{\mathrm{T}}$. The upper panels show MLEs for $\boldsymbol{\beta}$ and π based on reclassified data induced from this one raw

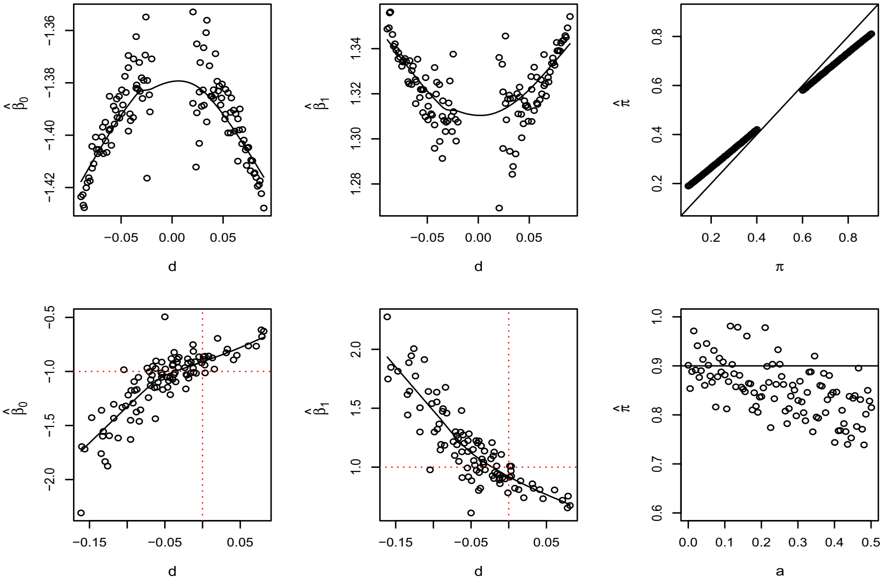
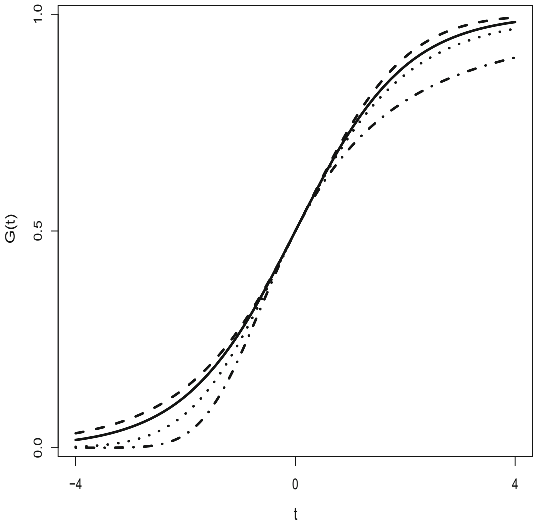


Fig. 1 Upper panels show $\hat{\boldsymbol{\beta}}$ versus $d = \hat{\pi} - \pi$ and $\hat{\pi}$ versus π based on reclassified data induced from a data set of size $n = 600$ generated according to a GLM with a generalized logit link with $(\alpha_1, \alpha_2) = (0.1, -0.1)$, where the reclassified data correspond to π varying over the range $[0.1, 0.4] \cup [0.6, 0.9]$. Lower panels show $\hat{\boldsymbol{\beta}}$ versus d and $\hat{\pi}$ versus a as a varies within [0, 0.5] based on one reclassified data with $\pi = 0.9$ at each level of a induced from a data set of size $n = 600$ generated from a GLM with the generalized logit link with $(\alpha_1, \alpha_2) = (a, -a)$. In the four panels regarding $\hat{\boldsymbol{\beta}}$, solid lines imposing on the scatter plots result from the loess fit, red dotted lines are the reference lines highlighting $d = 0$ and the truth of $\boldsymbol{\beta}$. The solid line in the plot of $\hat{\pi}$ versus π is a 45° reference line. The solid line in the plot of $\hat{\pi}$ versus a is the reference line signifying the true value of π, 0.9

Fig. 2 Three generalized logit links, with $(\alpha_1, \alpha_2) = (0.1, -0.1)$ (dashed line), $(-0.1, 0.2)$ (dotted line), and $(-0.5, 0.5)$ (dot-dashed line), contrasting with the logit link (solid line)



data set with π varying over the range $[0.1, 0.4] \cup [0.6, 0.9]$, while always assuming a logistic model for Y given X. The nearly symmetric pattern with respect to the center point of the presented range of the horizontal axis shown these estimates indicates that varying π over only the upper region above 0.5, such as [0.6, 0.9], can lead to a simpler and more effective extrapolation. More importantly, $\pi = 0.5$ is a singularity where $\boldsymbol{\beta}$ is not identifiable based on the corresponding reclassified data, even though the scatter plot of $\hat{\pi}$ versus π suggests that $\hat{\pi}$ approaches 0.5 (the truth) as π tends to 0.5, and hence $d = \hat{\pi} - \pi$ approaches zero. This is consistent with the implication of the estimating equations in (2). By (4), $\mu_0 = E(Y^*|X = x; \boldsymbol{\beta}) = 0.5$ for all x and $\boldsymbol{\beta}$; and thus, by (3), $p = 0.5$ in conjunction with any value for $\boldsymbol{b}$ solves (2) for all $\boldsymbol{\beta}$. This is why $\hat{\boldsymbol{\beta}}$ becomes ill-behaved as d approaches zero (due to π approaching 0.5). Despite this singularity point of $\pi = 0.5$, the first two upper panels do imply certain dependence of $\hat{\boldsymbol{\beta}}$ on d that motivates the implicit bias reduction method. To show such dependence from a different angle, we design another experiment where we fix π at 0.9 when generating reclassified data, and vary the true GLM from which the raw binary responses are simulated, where the link functions in these models are generalized logit links with $(\alpha_1, \alpha_2) = (a, -a)$, in which a varies from 0 to 0.5. As pointed out in Sect. 4, the generalized logit link with $(\alpha_1, \alpha_2) = (0, 0)$ is simply the logit link, producing a case where the assumed logistic model coincides with the true model. Using the reclassified data induced from each raw data set at a fixed a-level, we obtain the MLEs for $\boldsymbol{\beta}$ and π shown in the lower panels in Fig. 1. Like those seen in the upper panels, the dependence of $\hat{\boldsymbol{\beta}}$ on d is evident, and the former becomes closer to the truth as d gets closer to zero, but without the concern of ill-behaved $\hat{\boldsymbol{\beta}}$ near the singularity point of $\pi = 0.5$ observed in the upper panels.

Certainly, in a given application with one observed data set for (Y, X), one cannot vary the underlying true model as we did to create plots in the lower panels in

Fig. 1. In order to empirically manifest the dependence of $\hat{\boldsymbol{\beta}}$ on d, one creates movements in d by varying π when generating reclassified data based on one given raw data set while avoiding the singularity point of $\pi = 0.5$. One concern of extrapolating at $d = 0$ in RC-4 is that this direction of extrapolation is equivalent to pushing π towards 0.5. Although this is a legitimate concern, the hope here is that the dependence of $\hat{\boldsymbol{\beta}}$ on d observed in the lower panels of Fig. 1 can be preserved well enough over the majority of the lower or upper half range of π so that one can utilize the preserved dependence over such range to learn the underlying dependence of $\hat{\boldsymbol{\beta}}$ and d as model misspecification diminishes (corresponding to a shrinks to zero in the lower panels of Fig. 1.)

This proposed method shares some similarity with a bias reduction method well received in the measurement error community, known as the simulation extrapolation (SIMEX) method (Cook and Stefanski 1994; Stefanski and Cook 1995). Consider a more general setting where one has a method to consistently estimate a parameter, say, θ, based on data $\{(Y_i, X_i)\}_{i=1}^n$. Suppose that $\{X_i\}_{i=1}^n$ are unobserved, and the actual observed covariate values are $\{W_i\}_{i=1}^n$, where $W_i = X_i + U_i$, in which U_i is the nondifferential measurement error (Carroll et al. 2006, section 2.5) for $i = 1, \ldots, n$. If one ignores measurement error, one would apply the same estimation method to data $\{(Y_i, W_i)\}_{i=1}^n$ to estimate θ, resulting in a naive estimator denoted by $\tilde{\theta}$, which is typically inconsistent. To reduce bias in $\tilde{\theta}$, the SIMEX method exploits further contaminated covariate data as in the following algorithm, where $\lambda_1 < \lambda_2 < \ldots < \lambda_K$ are a sequence of user-specified positive constants.

SM-1 For each $(j, k) \in \{1, \ldots, J\} \times \{1, \ldots, K\}$, generate further contaminated covariate data, $\{W^*_{ijk} = W_i + \sqrt{\lambda_k} U^*_{ij}, \; i = 1, \ldots, n\}_{j=1}^J$, where $\{U^*_{ij}\}_{i=1}^n$ are user-simulated pseudo errors that follow the same distribution as $\{U_i\}_{i=1}^n$, and are independent across $j = 1, \ldots, J$.

SM-2 For each $(j, k) \in \{1, \ldots, J\} \times \{1, \ldots, K\}$, compute the naive estimate for θ using data $\{(Y_i, W^*_{ijk})\}_{i=1}^n$, denoted by $\hat{\theta}_{k,j}$. Compute $\hat{\theta}(\lambda_k) = J^{-1} \sum_{j=1}^J \hat{\theta}_{k,j}$, for $k = 1, \ldots, K$.

SM-3 View $\{\hat{\theta}(\lambda_k), \lambda_k\}_{k=0}^K$ as $K + 1$ realizations of the response-predictor pair $(\hat{\theta}(\lambda), \lambda)$, where $\lambda_0 = 0$ and $\hat{\theta}(0) = \tilde{\theta}$. Use these realizations to carry out regression analysis assuming a user-specified regression function.

SM-4 Use the regression results from SM-3 to extrapolate the response $\hat{\theta}(\lambda)$ at $\lambda = -1$, leading to a SIMEX estimate for θ.

Heuristically, the case with $\lambda = -1$ is of interest because $\mathrm{Var}(W_i + \sqrt{\lambda} U^*_{ij} | X_i) = (1 + \lambda)\mathrm{Var}(W_i | X_i)$, which is equal to zero at $\lambda = -1$, corresponding to data without measurement error. Thus, with a well chosen extrapolant, $\hat{\theta}(-1)$ is expected to resemble the consistent estimate one would obtain had one used data $\{(Y_i, X_i)\}_{i=1}^n$ to estimate θ.

Both SIMEX and the implicit bias reduction method can be easily generalized to models with a vector covariate X, with more caution when choosing an extrapolant, since the extrapolant is typically unknown. In the practice of SIMEX, simple extrapolant such as the quadratic extrapolant has shown to work well in many scenarios. Plots of $\hat{\boldsymbol{\beta}}$ in Fig. 1 along with the fitted loess curve (Cleveland and Devlin 1988) also suggest that a quadratic extrapolant may be adequate in RC-3 and RC-4 in the proposed implicit bias reduction algorithm. This will be the extrapolant used in our second proposed method in the simulation study. The proposed method differs from SIMEX in two aspects. First, in SM-1, in order to simulate pseudo measurement error, one needs to estimate the distribution of $\{U_i\}_{i=1}^n$ based on external data or replicate measures. In contrast, in RC-1, one knows the right model according to which coarsened data are induced from the original data. Second, the absence of measurement error translates to a correct model specification in the context of SIMEX; but model misspecification remains even if one eliminates data coarsening (by setting $\pi = 1$) in our context. This, along with the fact that λ is not estimated in SIMEX but π is estimated along with $\boldsymbol{\beta}$ in our method, makes it more challenging to develop an estimator for the variance of $\hat{\boldsymbol{\beta}}^{(2)}$ following the strategy for SIMEX estimators proposed in Stefanski and Cook (1995). One may adopt bootstrap methods to estimate the variance of $\hat{\boldsymbol{\beta}}^{(2)}$, even though we do not pursue this issue in the current study.

4 Simulation study

4.1 Finite sample performance of $\hat{\boldsymbol{\beta}}^{(1)}$

A simulation study is conducted to compare the first proposed estimator $\hat{\boldsymbol{\beta}}^{(1)}$ with the naive estimator $\tilde{\boldsymbol{\beta}}$ when one assumes a logistic model whereas the truth is a generalized logistic model (Stukel 1988), with $\boldsymbol{\beta} = (-1, 1)^{\mathrm{T}}$. Here, $H(t) = 1/\{1 + \exp(-t)\}$ and $G(t) = 1/[1 + \exp\{-h(t)\}]$, where

$$h(t) = \begin{cases} \alpha_1^{-1}\{\exp(\alpha_1 t) - 1\}I(\alpha_1 > 0) + tI(\alpha_1 = 0) - \alpha_1^{-1}\log(1 - \alpha_1 t)I(\alpha_1 < 0), & \text{if } t \geq 0, \\ -\alpha_2^{-1}\{\exp(-\alpha_2 t) - 1\}I(\alpha_2 > 0) + tI(\alpha_2 = 0) + \alpha_2^{-1}\log(1 + \alpha_2 t)I(\alpha_2 < 0), & \text{if } t < 0. \end{cases}$$

Note that, if $\alpha_1 = \alpha_2 = 0$, $G(t)$ reduces to the logit link; if $\alpha_1 = \alpha_2$, $G(t)$ is symmetric; otherwise, $G(t)$ is asymmetric. The generalized logit link with $(\alpha_1, \alpha_2) = (-0.1, 0.2)$ used in this experiment is depicted as the dotted curve in Fig. 2. We consider three covariate distributions, all with mean zero and variance one: $N(0, 1)$, uniform$(-\sqrt{3}, \sqrt{3})$, and a shifted gamma distribution with skewness equal to $\sqrt{2}$. These covariate distribution configurations encompass symmetric and asymmetric distributions, as well as distributions with bounded support and unbounded support. Given the simulated covariate values $\{X_i\}_{i=1}^n$, $\{Y_i\}_{i=1}^n$ are generated from the generalized logistic model, where $n = 400$, 600, 800, 1000. Based on each simulated data set $\{(Y_i, X_i)\}_{i=1}^n$, we carry out logistic regression to obtain $\tilde{\boldsymbol{\beta}}$; then two reclassified data sets are generated according to (1) with $(\pi_1, \pi_2) = (0.7, 0.9)$. Using these two coarsened data sets, $\hat{\boldsymbol{\beta}}^{(1)}$ is computed according to (13) and (14). This experiment is repeated 1000 times at each simulation setting.

Table 1 presents summary statistics of the simulation results, including Monte Carlo averages of the considered estimates, mean absolute deviations (MAD) of these estimates from the corresponding truth, and empirical coverage probabilities (CP) of 95% confidence intervals. The 95% confidence intervals for β_0 based on $\tilde{\beta}_0$ is obtained by invoking the asymptotic normality of MLE, leading to $\tilde{\beta}_0 \pm 1.96 \times \text{s.e.}(\tilde{\beta}_0)$ as the interval bounds, where $\text{s.e.}(\tilde{\beta}_0)$ is the estimated standard error of $\tilde{\beta}_0$ resulting from the sandwich variance estimation for M-estimators (Boos and Stefanski 2013, section 7.2.1). A 95% confidence internal for β_1 based on $\tilde{\beta}_1$ is similarly obtained using each simulated data set. Also assuming asymptotic normality for $\hat{\boldsymbol{\beta}}^{(1)}$, we construct 95% confidence intervals based on the first proposed estimate, with estimated standard errors associated with each point estimate obtained via a bootstrap method involving 100 bootstrap samples. Under the current designed link misspecification, $\tilde{\boldsymbol{\beta}}$ is noticeably compromised, and $\hat{\boldsymbol{\beta}}^{(1)}$ exhibits impressive bias reduction, at the price of inflated variability. Due to the inflated variability, the MAD associated with $\hat{\boldsymbol{\beta}}^{(1)}$ is typically around three times as high as that associated with $\tilde{\boldsymbol{\beta}}$ in the current simulation settings. Thanks to the bias correction, and also in part due to the inflated variability, the confidence intervals based on $\hat{\boldsymbol{\beta}}^{(1)}$ stay much closer to the nominal level compared to those based on $\tilde{\boldsymbol{\beta}}$, of which coverage probabilities drop quickly as sample size increases.

As discussed in Sect. 3.1, $\hat{\boldsymbol{\beta}}^{(1)}$ can deteriorate in the presence of severe link misspecification, and its quality depends on factors irrelevant to the primary model configuration, such as the distribution of X and the choice of (π_1, π_2). Among all three covariate distributions experimented in the presented simulation study, we see the proposed estimator achieve bias reduction to some extent. When choosing (π_1, π_2), we suggest selecting two values in (0.5, 1) so that the reclassified responses do not lose too much information in the original responses. Another guideline we have found practically useful in the empirical study is to choose (π_1, π_2) so that the common denominator appearing in (13) and (14), i.e., $(\hat{\pi}_1 - \pi_1)/(2\pi_1 - 1) - (\hat{\pi}_2 - \pi_2)/(2\pi_2 - 1)$, is not too close to zero, say, is above 0.01. Even with the above practical considerations one needs to bear in mind when applying the proposed explicit bias reduction method, it is still an appealing and convenient way to correct the naive estimates for bias because of the closed-form expressions for such correction once the naive estimates are computed via straightforward maximum likelihood estimation.

Even though all derivations in Sects. 2.2 and 3.1 still go through when β_1 is a vector slope parameter, with Taylor approximation in (6) and other derivations relating to β_1 done elementwise, we do not recommend generalizing the explicit bias reduction method to models with a vector covariate for reasons pointed out at the end of Sect. 3.1.

Table 1 Averages of parameter estimates for $\boldsymbol{\beta}$, denoted by $\tilde{\boldsymbol{\beta}}$, and the corresponding mean absolute deviations (MAD) and empirical coverage probabilities (CP) of 95% confidence intervals across 1000 Monte Carlo replicates when assuming a logistic model for Y given X, and the counterpart quantities associated with the proposed explicit bias-reduced estimates, $\hat{\boldsymbol{\beta}}^{(1)}$. The true parameter values are $\boldsymbol{\beta} = (-1, 1)^{\mathrm{T}}$. Numbers in parentheses are Monte Carlo standard errors (s.e.) associated with the averages

	Estimate (s.e.)	MAD (s.e.)	CP	Estimate (s.e.)	MAD (s.e.)	CP
$X \sim N(0,1)$						
	$n = 400$			$n = 600$		
$\tilde{\beta}_0$	−1.177 (0.004)	0.185 (0.004)	0.780	−1.168 (0.003)	0.172 (0.003)	0.687
$\tilde{\beta}_1$	1.166 (0.005)	0.185 (0.004)	0.833	1.162 (0.004)	0.171 (0.004)	0.762
$\hat{\beta}_0^{(1)}$	−0.977 (0.036)	0.672 (0.029)	0.955	−0.924 (0.034)	0.652 (0.027)	0.933
$\hat{\beta}_1^{(1)}$	1.129 (0.047)	0.889 (0.038)	0.952	1.050 (0.043)	0.827 (0.034)	0.944
	$n = 800$			$n = 1000$		
$\tilde{\beta}_0$	−1.178 (0.003)	0.179 (0.003)	0.554	−1.179 (0.003)	0.179 (0.003)	0.439
$\tilde{\beta}_1$	1.175 (0.003)	0.179 (0.003)	0.634	1.170 (0.003)	0.173 (0.003)	0.580
$\hat{\beta}_0^{(1)}$	−0.936 (0.034)	0.645 (0.027)	0.943	−0.977 (0.034)	0.626 (0.027)	0.931
$\hat{\beta}_1^{(1)}$	1.010 (0.044)	0.808 (0.035)	0.944	1.028 (0.039)	0.769 (0.031)	0.948
$X \sim \text{uniform}(-\sqrt{3}, \sqrt{3})$						
	$n = 400$			$n = 600$		
$\tilde{\beta}_0$	−1.210 (0.004)	0.214 (0.004)	0.698	−1.193 (0.004)	0.196 (0.003)	0.606
$\tilde{\beta}_1$	1.201 (0.005)	0.211 (0.004)	0.749	1.185 (0.004)	0.189 (0.003)	0.670
$\hat{\beta}_0^{(1)}$	−1.017 (0.037)	0.711 (0.029)	0.951	−0.930 (0.035)	0.687 (0.028)	0.945
$\hat{\beta}_1^{(1)}$	1.025 (0.041)	0.777 (0.033)	0.948	0.967 (0.043)	0.778 (0.036)	0.938
	$n = 800$			$n = 1000$		
$\tilde{\beta}_0$	−1.201 (0.003)	0.202 (0.003)	0.467	−1.200 (0.003)	0.200 (0.003)	0.377
$\tilde{\beta}_1$	1.192 (0.003)	0.194 (0.003)	0.537	1.190 (0.003)	0.191 (0.003)	0.455
$\hat{\beta}_0^{(1)}$	−0.962 (0.034)	0.651 (0.027)	0.942	−0.962 (0.032)	0.629 (0.025)	0.933
$\hat{\beta}_1^{(1)}$	0.936 (0.036)	0.655 (0.030)	0.946	0.989 (0.036)	0.664 (0.029)	0.949
$X \sim$ shifted gamma						
	$n = 400$			$n = 600$		
$\tilde{\beta}_0$	−1.179 (0.004)	0.187 (0.004)	0.749	−1.179 (0.003)	0.182 (0.003)	0.633
$\tilde{\beta}_1$	1.138 (0.005)	0.164 (0.004)	0.877	1.137 (0.004)	0.151 (0.003)	0.837
$\hat{\beta}_0^{(1)}$	−1.036 (0.039)	0.803 (0.030)	0.948	−0.985 (0.035)	0.687 (0.027)	0.936
$\hat{\beta}_1^{(1)}$	1.060 (0.048)	0.921 (0.039)	0.960	1.064 (0.046)	0.835 (0.037)	0.953
	$n = 800$			$n = 1000$		
$\tilde{\beta}_0$	−1.177 (0.003)	0.180 (0.003)	0.527	−1.172 (0.003)	0.173 (0.003)	0.460
$\tilde{\beta}_1$	1.138 (0.003)	0.147 (0.003)	0.749	1.128 (0.003)	0.134 (0.003)	0.737
$\hat{\beta}_0^{(1)}$	−0.970 (0.033)	0.624 (0.026)	0.937	−0.973 (0.030)	0.576 (0.024)	0.932
$\hat{\beta}_1^{(1)}$	1.049 (0.043)	0.774 (0.036)	0.952	1.016 (0.038)	0.694 (0.031)	0.959

4.2 Finite sample performance of $\hat{\boldsymbol{\beta}}^{(2)}$

To demonstrate the performance of the implicit bias reduction method, we carry out simulation experiments under similar settings as those in Sect. 4.1, except that the true link function in the data generating process takes a sequence of generalized logit links. More specifically, we consider true links as generalized logit link $G(t)$ with $(\alpha_1, \alpha_2) = (-a, a)$, where a varies from 0.1 to 0.5 at increments of 0.1. The generalized logit link that deviates from the logit link the most in this sequence, at $a = 0.5$, is shown in Fig. 2. When implementing this method, we estimate $\boldsymbol{\beta}$ based on reclassified data generated according to (1) with π varying from 0.6 to 0.9 at increments of 0.005, with $J = 100$ in the algorithm described in Sect. 3.2; and we use the quadratic extrapolant in RC-3 and RC-4 to obtain $\hat{\boldsymbol{\beta}}^{(2)}$. Based on 1000 Monte Carlo replicates, Fig. 3 provides pictorial comparisons between $\hat{\boldsymbol{\beta}}^{(2)}$ and $\tilde{\boldsymbol{\beta}}$ when $n = 800$, with covariate $X \sim N(0, 1)$. Figure 4 shows the same comparisons when X follows a shifted gamma distribution with mean zero, variance one, and skewness $\sqrt{2}$. The substantial bias reduction achieved by $\hat{\boldsymbol{\beta}}^{(2)}$ using a quadratic extrapolant is evident in both figures. Although, like $\hat{\boldsymbol{\beta}}^{(1)}$, $\hat{\boldsymbol{\beta}}^{(2)}$ is more variable than $\tilde{\boldsymbol{\beta}}$ (but much less variable than $\hat{\boldsymbol{\beta}}^{(1)}$), which is expected when coarsened data are used to compute

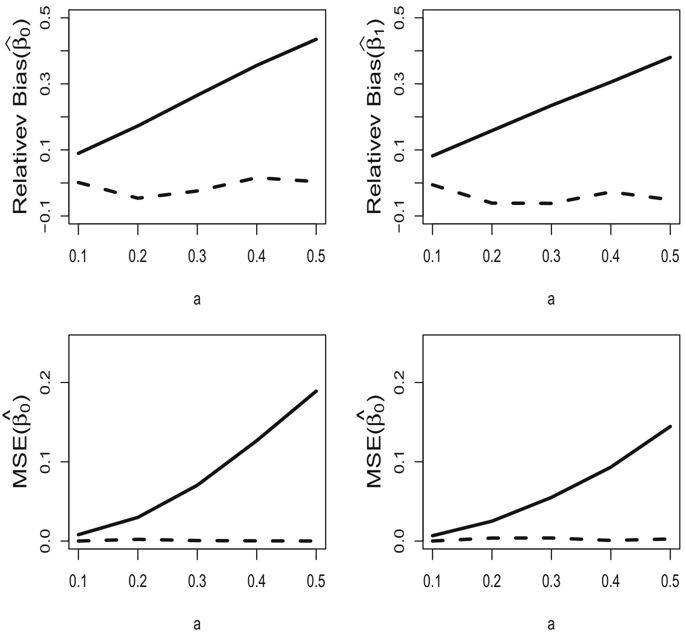


Fig. 3 Upper panels show Monte Carlo averages of relative bias of $\hat{\boldsymbol{\beta}}^{(2)}$ (dashed line) and those of $\tilde{\boldsymbol{\beta}}$ (solid line). Lower panels show MSE of $\hat{\boldsymbol{\beta}}^{(2)}$ (dashed line) and MSE of $\tilde{\boldsymbol{\beta}}$ (solid line). True links are generalized logit links with $(\alpha_1, \alpha_2) = (-a, a)$. The covariate X follows $N(0, 1)$

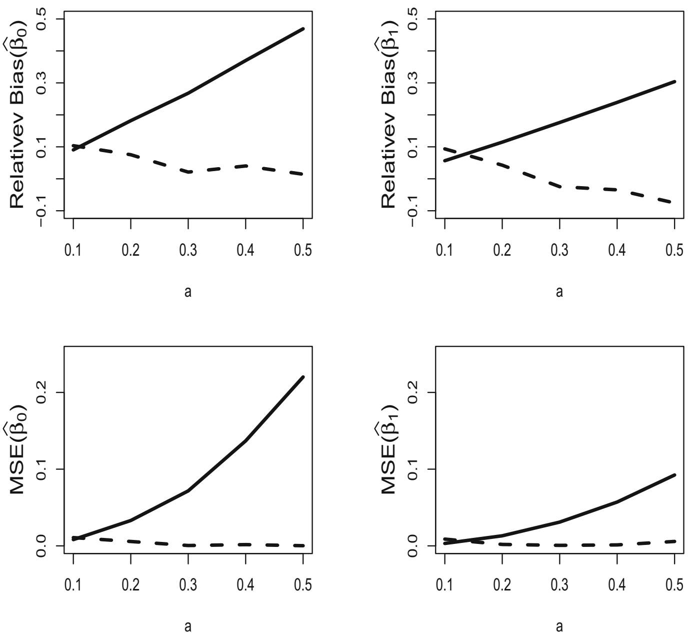


Fig. 4 Upper panels show Monte Carlo averages of relative bias of $\hat{\boldsymbol{\beta}}^{(2)}$ (dashed line) and those of $\tilde{\boldsymbol{\beta}}$ (solid line). Lower panels show MSE of $\hat{\boldsymbol{\beta}}^{(2)}$ (dashed line) and MSE of $\tilde{\boldsymbol{\beta}}$ (solid line). True links are generalized logit links with $(\alpha_1, \alpha_2) = (-a, a)$. The covariate X follows a shifted gamma distribution with mean zero, variance one, and skewness $\sqrt{2}$

the bias-reduced estimator and an additional parameter needs to be estimated simultaneously. Accounting for both bias and variance, the mean squared error (MSE) of $\hat{\boldsymbol{\beta}}^{(2)}$ is significantly lower than that of $\tilde{\boldsymbol{\beta}}$.

Applying the implicit bias reduction method to regression models with multiple covariates does not add extra complication, even though a larger sample size is needed to obtain improved inference. Figure 5 shows the comparison of $\hat{\boldsymbol{\beta}}^{(2)}$ and $\tilde{\boldsymbol{\beta}}$ obtained from logistic regression with two covariates. In particular, we generate 300 Monte Carlo replicate data sets, each of size $n = 1000$, from the true models with the aforementioned sequence of generalized logit models, which involves one continuous covariate $X_1 \sim N(0, 1)$ and one binary covariate X_2 that takes value one with probability 0.5. The true regression coefficients are $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}} = (-1, 0.5, 0.5)^{\mathrm{T}}$. In this case, the proposed method with the quadratic extrapolant shows signs of over correcting $\tilde{\boldsymbol{\beta}}$ for bias when the link misspecification is mild (when $a = 0.1$ and 0.2), producing estimates more biased than $\tilde{\boldsymbol{\beta}}$. This may suggest the need to explore a different extrapolant. In practice, we recommend one plot $\hat{\boldsymbol{\beta}}^{(2)}(d)$ (elementwise) versus d to gain some visual hints on the choice of an extrapolant. Regardless, using the quadratic extrapolant in this

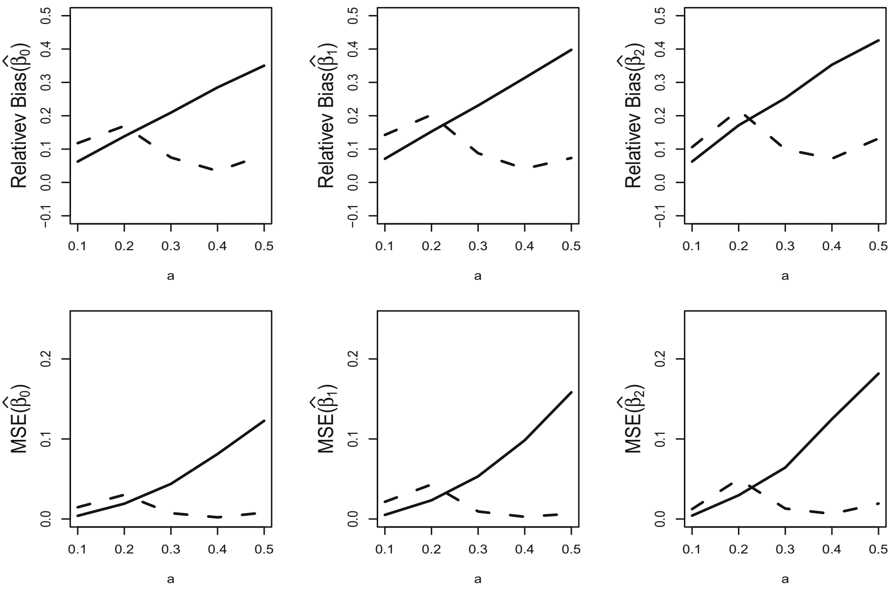


Fig. 5 Upper panels show Monte Carlo averages of relative bias of $\hat{\boldsymbol{\beta}}^{(2)}$ (dashed line) and those of $\tilde{\boldsymbol{\beta}}$ (solid line). Lower panels show MSE of $\hat{\boldsymbol{\beta}}^{(2)}$ (dashed line) and MSE of $\tilde{\boldsymbol{\beta}}$ (solid line). True links are generalized logit links with $(\alpha_1, \alpha_2) = (-a, a)$. The covariates X_1 follows $N(0,\ 1)$ and X_2 follows Bernoulli(0.5)

experiment, the gain from the proposed method again stand out once the link misspecification is more severe (e.g., when $a > 0.2$).

Up to this point, the true links $G(t)$ in the data generating processes in the simulation studies for the two proposed methods are all asymmetric. In "Appendix A" in the supplementary materials we present additional simulation study where symmetric generalized logit links are used in the data generating process. These additional results provide convincing empirical evidence that both proposed methods yield estimators less biased than $\tilde{\boldsymbol{\beta}}$ when the assumed link and the true link are both symmetric.

5 A test for link misspecification

Here we propose a simple t test for link misspecification using byproducts of the proposed bias reduction methods. Hosmer et al. (1997) compared nine tests for link misspecification in the context of logistic regression and found none of them exhibit satisfactory power. Among these tests, eight of them are goodness-of-fit (GOF) tests in nature constructed based on prediction error, and one is Stukel's score test based on fitting a generalized logit model (Stukel 1988). This score test is a test of $H_0 : (\alpha_1, \alpha_2) = (0, 0)$, where the score is the normal score derived from the likelihood of a generalized logistic model. Compared with the eight GOF tests, Stukel's score test exhibits the highest power to detect link misspecification. We believe the reason for this is that, although inference on covariate effects can be very misleading in the presence of link misspecification, the impact on predictions is often more subtle. Hence, a residual-based GOF test tends to be less sensitive to link misspecification.

A more sensitive indicator of link misspecification is readily available from the proposed estimation procedures for $\boldsymbol{\beta}$, which is simply $\hat{\pi}$, since $\hat{\pi}$ inconsistently estimate π in the presence of link misspecification. Hence, fixing π at a value one chooses, one can easily construct a t test with test statistic $t = (\hat{\pi} - \pi)/\hat{v}$, where $\hat{v}$ is the sandwich standard error estimator associated with $\hat{\pi}$. Following the asymptotic theory of MLE, it is straightforward to show that the null distribution of the test statistic is a t distribution with $n - \dim(\boldsymbol{\beta}) - 1$ degrees of freedom, where $\dim(\boldsymbol{\beta})$ denotes the dimension of $\boldsymbol{\beta}$. A test statistic value that deviates significantly from zero indicates an inadequate assumed link.

Like Stukel's score test, our proposed test is not based on prediction error. To compare the operating characteristics of Stukel's test and our test, we carry out simulation study under settings similar as those in the experiments in Sect. 4, with the sample size n varying from 200 to 1000 at increments of 200. We consider four true links in this experiment, including the logit link and three generalized logit links with $(\alpha_1, \alpha_2) = (-0.5, 0.5), (0.5, 0.5), (1, 1)$. The setting with the logit link as the truth allows one to assess the size of a test. We let $\pi = 0.9$ in the t test. Setting the significance level at 0.05, Fig. 6 depicts the empirical power defined by the rejection rate across 1000 MC replicates associated with each test versus n under each of the true link configurations. Clearly, our test outperforms the Stukel's test in all scenarios, and both retain the right size. Stukel's test appears to be more promising when the true link is asymmetric, whereas the proposed t test is more powerful in the presence of more severe link misspecification even when the true link is also symmetric as the assumed link. "Appendix B" in the supplementary materials provides QQ plots of the test statistics collected from the case without link misspecification, which suggest close agreement between the claimed null distribution of the proposed test statistic and a t distribution. Due to the typically moderate to large sample size n when comparing with the dimension of $\boldsymbol{\beta}$ in most applications where this proposed test is designed for, one may simply view the standard normal as the null distribution of the test statistic in practice.

6 A real data example

We now entertain a classic data example reported in Bliss (1935) that has been analyzed by many researchers since then, who were mostly concerned about the adequacy of the logistic model for this date set. The data were collected in an experiment where the association between mortality of adult beetles and exposure to gaseous carbon disulfide is of interest. In particular, the data include logarithm (with base 10) of dosages of carbon disulfide exposure for a total of 481 adult beetles, and the status (being killed or surviving) of each beetle after five hours' exposure. Let Y_i denote the indicator of being killed after exposure to carbon disulfide for the ith beetle, and denote by X_i the $\log_{10}$dose this beetle was exposed to, for $i = 1, \ldots, 481$. Pregibon (1980) applied his test for link specification and found strong evidence to

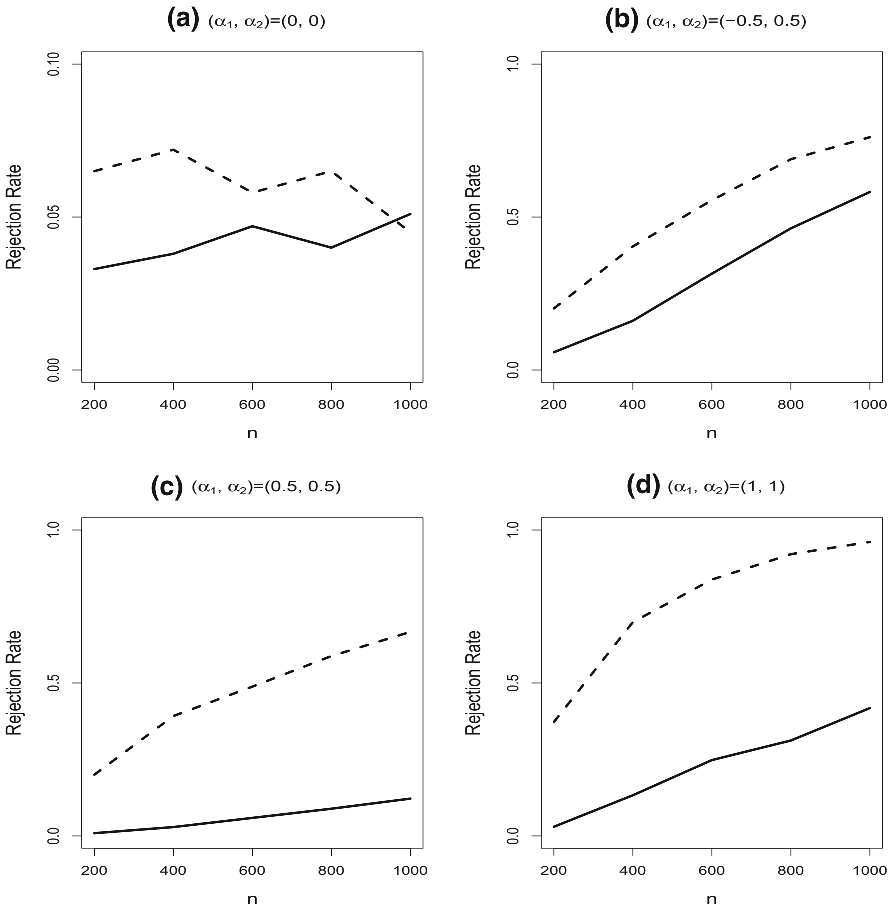


Fig. 6 Rejection rates of Stukel's test (solid lines) and the proposed t test (dashed lines) across 1000 Monte Carlo replicates versus the sample size n when the true link is logit (in (a)) and when the true link is a generalized logit link with three different configurations for (α_1, α_2) (in (b), (c), (d))

support an asymmetric link as opposed to the logit link. Aranda-Ordaz (1981) showed that the complementary log-log link is more appropriate for the data. Stukel (1988) first used the likelihood ratio GOF test, resulting in a p-value of 0.0815; she then followed up with her score test and found stronger evidence against the logit link, with a p-value of 0.0125; lastly, she used the likelihood ratio test to compare the logistic model and her proposed skewed generalized logistic model for this data, and obtained a p-value of 0.0077.

Using our test proposed in Sect. 5, with $\pi = 0.95$, we also reject the assumed logit link, with p-value equal to 0.0001. We then repeatedly estimate $\boldsymbol{\beta}$ assuming the logit link and using reclassified data simulated from the raw data according to (1) with π ranging from 0.75 to 0.95 at increments of 0.001. Applying a quadratic extrapolant on the sequence of $\boldsymbol{\beta}(\hat{d})$ leads to $\hat{\boldsymbol{\beta}}^{(2)} = (-57.99, 32.62)^{\mathrm{T}}$, with the estimated (via bootstrap) standard errors equal to (10.23, 5.41). In Stukel's analysis based on a generalized logistic model, her estimates of β_0 and β_1 are -47.4 (6.47) and 26.6 (3.66), respectively, with the corresponding estimated standard error in parentheses. And the standard logistic regression gives $\tilde{\boldsymbol{\beta}} = (-60.71, 34.27)^{\mathrm{T}}$, with the estimated standard errors equal to (5.18, 2.91). In comparison, $\hat{\boldsymbol{\beta}}^{(2)}$ lies between $\tilde{\boldsymbol{\beta}}$ and that obtained by Stukel, who attempted to correct for the potentially inadequate logit link. This can suggest that the implicit bias reduction method effectively reduce some bias in $\tilde{\boldsymbol{\beta}}$ even though we still analyze the (reclassified) data assuming a logit link. We did not apply the explicit bias reduction method to this data because the support of X excludes zero and one, the two values one evaluates X at when deriving $\hat{\boldsymbol{\beta}}^{(1)}$.

7 Discussion

The conventional model building and inference routine is to first test suspicious assumptions in a posited model using some diagnostic tools; and if inadequate model assumptions are detected, one makes attempts to correct the model and draw inference again using the updated model. If one has little ground for verifying or correcting the posited model, one often resorts to semi-/non-parametric methods to draw inference. As rich as the body of existing semi-/non-parametric methods, most of them are computationally demanding and can be inefficient. In this study we present a different take on parametric inference that leads to more reliable inference even without guessing the "right" model. Moreover, we unify parametric inference and model diagnosis under the same framework based on simulated reclassified data. If the test for the assumed link does not reject the null, one has some reassurance for $\tilde{\boldsymbol{\beta}}$; otherwise, one may adopt the proposed bias-reduced estimates. In fact, we would not recommend one use the proposed bias reduction methods when one lacks sufficient evidence to indicate presence of model misspecification. The first bias reduction estimator for β_0 in (13) comes from (12), which becomes an identity that sheds no light on β_0 when the assumed model is not misspecified since now one has, on the left-hand side of (12), $b_0(0, \pi_1) - b_0(0, \pi_2) = 0$ for all π_1 and π_2, and the factor following $R(\beta_0)$ on the right-hand side is also zero. Consequently, (13) does not yield a sensible estimator for β_0. For the second bias reduction estimator for $\boldsymbol{\beta}$, the problem with it in the absence of model misspecification lies in the fact that d_k is expected to be close to zero for all k's, and the sequence $\{\hat{\boldsymbol{\beta}}(d_k), d_k\}_{k=1}^{K}$ will provide little information on the dependence of $\hat{\boldsymbol{\beta}}$ on d, and thus regressing $\hat{\boldsymbol{\beta}}(d_k)$ on d_k can be subject to high variability.

We provide in "Appendix C" in the supplementary materials the SAS PROC IML code for implementing the proposed estimation methods and the test. Computationally, the implicit bias reduction method is more demanding than the explicit bias reduction method mainly because the former entails computing the MLEs of unknown parameters for a large number of times. This makes using bootstrap methods, such as those described in Section A.9.4 in Carroll et al. (2006), to estimate the variance of $\hat{\boldsymbol{\beta}}^{(2)}$ more cumbersome. To develop new variance estimation methods for these estimators that are computationally less burdensome is among our upcoming research agenda.

Although we consider the assumed link in GLM as the source of model misspecification in the current study, the implicit bias reduction method and the proposed t test are also applicable when a different model assumption is violated, such as those considered in Hosmer et al. (1997). To illustrate the rationale of the proposed methods, we keep the reclassification mechanism as simple as (1), but different data coarsening mechanisms are certainly worth systematic investigation. This is especially needed in order to generalize these methods to GLM for responses other than a binary response. Even in the context of our study, different coarsening mechanisms that lead to induced data allowing stronger identifiability of π can be beneficial, as Copas (1988) pointed out, who used a missclassification model similar to ours, that π is hard to estimate unless when n is very large. This inherent weak identifiability of π causes little numerical difficulty in implementing the proposed methods when one uses the true value of π as the starting value when obtaining its MLE, which is feasible since the truth is known for this user-specified parameter. But, for a more complex GLM when such parameter becomes harder to estimate, one may generate coarsened data that allow part of the raw data free of error to alleviate the nonidentifiability issue. This is similar in spirit to having validation data or external data to allow identifiability of measurement error distributions. Furthermore, with the reclassfication mechanism given by (1), the proposed t test can be improved by using $\sup_{\pi \in (0.5,1)} (|\hat{\pi} - \pi|/\hat{v})$ as the test statistic, instead of fixing π at one value as is done in Sects. 5 and 6. The drawback of this supreme-type test statistic is that its null distribution is no longer as trivial as before, which may need to be estimated using some simulation-based methods. For instance, one may approximate the null distribution of $t_\pi = |\hat{\pi} - \pi|/\hat{v}$ by a Gaussian process indexed by π, and use some bootstrap method to obtain an estimate for the covariance function under an assumed covariance structure for the process.

The introduction of an extraneous parameter as a device to calibrate estimates for parameters of interest can be applied to other models more complex than GLM, which can be more vulnerable to model misspecification. Since "...all models are wrong but some are useful..." (George Box), rather than attempting to guess the right model, a more productive approach to draw inference is to embrace a useful wrong model and then strive for inference results that remain reliable under the wrong model. This is the very philosophy we follow in this study and also in our follow-up research.

Acknowledgements I would like to thank the Associate Editor and the two anonymous referees for their helpful suggestions and insightful comments that greatly improve of the manuscript.

Appendix: Proof of equation (2)

Because the assumed GLM is specified by $P(Y=1|X)=H(\eta)$, where $\eta=\beta_0+\beta_1 X$, and the reclassified response is generated according to $P(Y^*=Y|Y,X)=\pi$, one has

$$\begin{aligned} P(Y^*=1|X) &= P(Y^*=1,Y=0|X)+P(Y^*=1,Y=1|X)\\ &= P(Y=0|X)P(Y^*\neq Y|Y,X)+P(Y=1|X)P(Y^*=Y|Y,X)\\ &= \{1-H(\eta)\}(1-\pi)+H(\eta)\pi\\ &= (2\pi-1)H(\eta)+1-\pi. \end{aligned} \tag{15}$$

It follows that the likelihood function based on the assumed primary model for Y^* evaluated at one data point (Y^*,X) is $L(\pi,\boldsymbol{\beta})=P(Y^*=1|X)^{Y^*}\{1-P(Y^*=1|X)\}^{1-Y^*}$, and the log-likelihood function is $\ell(\pi,\boldsymbol{\beta})=Y^*\log P(Y^*=1|X)+(1-Y^*)\log\{1-P(Y^*=1|X)\}$.

Differentiating (15) with respect to each element in $(\pi,\boldsymbol{\beta})$ gives

$$\begin{aligned} \frac{\partial P(Y^*=1|X)}{\partial\pi} &= 2H(\eta)-1,\\ \frac{\partial P(Y^*=1|X)}{\partial\beta_0} &= (2\pi-1)H'(\eta),\\ \frac{\partial P(Y^*=1|X)}{\partial\beta_1} &= (2\pi-1)H'(\eta)X. \end{aligned} \tag{16}$$

Using (16), one can show that the three normal score functions associated with $\ell(\pi,\boldsymbol{\beta})$ are given by

$$\begin{aligned} \frac{\partial\ell(\pi,\boldsymbol{\beta})}{\partial\pi} &= Y^*\frac{2H(\eta)-1}{P(Y^*=1|X)}-(1-Y^*)\frac{2H(\eta)-1}{1-P(Y^*=1|X)},\\ \frac{\partial\ell(\pi,\boldsymbol{\beta})}{\partial\beta_0} &= Y^*\frac{(2\pi-1)H'(\eta)}{P(Y^*=1|X)}-(1-Y^*)\frac{(2\pi-1)H'(\eta)}{1-P(Y^*=1|X)},\\ \frac{\partial\ell(\pi,\boldsymbol{\beta})}{\partial\beta_1} &= Y^*\frac{(2\pi-1)H'(\eta)X}{P(Y^*=1|X)}-(1-Y^*)\frac{(2\pi-1)H'(\eta)X}{1-P(Y^*=1|X)}. \end{aligned}$$

To further simplify notations, let $\mu=P(Y^*=1|X)$. The above three score functions can be re-expressed as

$$
\begin{aligned}
\frac{\partial \ell(\pi, \boldsymbol{\beta})}{\partial \pi} &= \frac{Y^* - \mu}{\mu(1-\mu)}\{2H(\eta) - 1\},\\
\frac{\partial \ell(\pi, \boldsymbol{\beta})}{\partial \beta_0} &= \frac{Y^* - \mu}{\mu(1-\mu)}(2\pi - 1)H'(\eta),\\
\frac{\partial \ell(\pi, \boldsymbol{\beta})}{\partial \beta_1} &= \frac{Y^* - \mu}{\mu(1-\mu)}(2\pi - 1)H'(\eta)X.
\end{aligned}
\tag{17}
$$

The expectation of the first score in (17) with respect to the true distribution of (Y^*, X) is

$$
\begin{aligned}
E\left[\frac{Y^* - \mu}{\mu(1-\mu)}\{2H(\eta) - 1\}\right] &= E\left(E\left[\frac{Y^* - \mu}{\mu(1-\mu)}\{2H(\eta) - 1\}\middle| X\right]\right)\\
&= E\left[\frac{\mu_0 - \mu}{\mu(1-\mu)}\{2H(\eta) - 1\}\right],
\end{aligned}
$$

where η_0 is equal to η evaluated at the true value of $\boldsymbol{\beta}$, and $\mu_0 = (2\pi - 1)G(\eta_0) + 1 - \pi$, as defined in (4), is the mean of Y^* given X under the correct model evaluated at the true parameter values. Setting this expectation equal to zero gives the first estimating equation in (2). Similarly, the expectations of the second and the third score functions in (17) with respect to the true distribution of (Y^*, X) are given by

$$
\begin{aligned}
E\left[\frac{Y^* - \mu}{\mu(1-\mu)}(2\pi - 1)H'(\eta)\right] &= E\left[\frac{\mu_0 - \mu}{\mu(1-\mu)}H'(\eta)\right](2\pi - 1),\\
E\left[\frac{Y^* - \mu}{\mu(1-\mu)}(2\pi - 1)H'(\eta)X\right] &= E\left[\frac{\mu_0 - \mu}{\mu(1-\mu)}H'(\eta)X\right](2\pi - 1),
\end{aligned}
$$

respectively. Setting these two expectations equal to zero gives the second and the third equations in (2).

References

Aranda-Ordaz FJ (1981) On two families of transformations to additivity for binary response data. Biometrika 68(2):357–363

Bliss CI (1935) The calculation of the dosage–mortality curve. Ann Appl Biol 22(1):134–167

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

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C (2006) Measurement error in nonlinear models: a modern perspective. Chapman & Hall/CRC, Boca Raton

Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc 83(403):596–610

Cook JR, Stefanski LA (1994) Simulation-extrapolation estimation in parametric measurement error models. J Am Stat Assoc 89(428):1314–1328

Copas JB (1988) Binary regression models for contaminated data. J R Stat Soc Ser B (Methodol) 50(2):225–253

Czado C, Santner TJ (1992) The effect of link misspecification on binary regression inference. J Stat Plan Infer 33(2):213–231

Guerrero VM, Johnson RA (1982) Use of the box-cox transformation with binary response models. Biometrika 69(2):309–314

Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S (1997) A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 16(9):965–980

Jiang X, Dey DK, Prunier R, Wilson AM, Holsinger KE (2013) A new class of flexible link functions with application to species co-occurrence in cape floristic region. Ann Appl Stat 7(4):2180–2204

Kim S, Chen MH, Dey DK (2007) Flexible generalized t-link models for binary response data. Biometrika 95(1):93–106

McCullagh P, Nelder J (1989) Generalized linear models. Chapman & Hall/CRC, Boca Raton

Morgan BJ (1983) Observations on quantit analysis. Biometrics 39(4):879–886

Nelder JA, Wedderburn RW (1972) Generalized linear models. J R Stat Soc Ser A-G 135(3):370–384

Pregibon D (1980) Goodness of link tests for generalized linear models. J R Stat Soc C-Appl 29(1):15–24

Samejima F (2000) Logistic positive exponent family of models: virtue of asymmetric item characteristic curves. Psychometrika 65(3):319–335

Stefanski LA, Cook JR (1995) Simulation-extrapolation: the measurement error jackknife. J Am Stat Assoc 90(432):1247–1256

Stukel TA (1988) Generalized logistic models. J Am Stat Assoc 83(402):426–431

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50(1):1–25

Whittemore AS (1983) Transformations to linearity in binary regression. SIAM J Appl Math 43(4):703–710

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.