

Dual Model Misspecification in Generalized Linear Models with Error in Variables

Xianzheng Huang

Abstract We study maximum likelihood estimation of regression parameters in generalized linear models for a binary response with error-prone covariates when the distribution of the error-prone covariate or the link function is misspecified. We revisit the remeasurement method proposed by Huang, Stefanski, and Davidian (2006) for detecting latent-variable model misspecification and examine its operating characteristics in the presence of link misspecification. Furthermore, we propose a new diagnostic method for assessing assumptions on the link function. Combining these two methods yields informative diagnostic procedures that can identify which model assumption is violated and also reveal the direction in which the true latent-variable distribution or the true link function deviates from the assumed one.

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). Studies in these applications often involve covariates that cannot be measured precisely or directly. For example, in the Framingham Heart Study (Kannel et al., 1986), a logistic regression model was used to relate the indicator for the presence of coronary heart disease with covariates such as one's smoking status, body mass index, age, serum cholesterol level, and long-term systolic blood pressure (SBP). Among these covariates, measures of one's serum cholesterol level were imprecise, and the actual observed blood pressure of a subject is merely a noisy surrogate of the long-term SBP, which cannot be measured directly. Taking the structural model point of view to account for measurement error as opposed to the functional model point of view (Carroll et al., 2006, Section 2.1), one needs to

Xianzheng Huang
Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, U.S.A.,
e-mail: huang@stat.sc.edu

assume a model for the latent true covariates in order to derive the observed data likelihood function. Together the latent-covariate model, the model that relates the true covariates with their noisy surrogates, and the GLM as the conditional model of the response given the true covariates, one has the complete specification of a structural measurement error model for the observed data. From that point on, one can draw parametric inference on the regression parameters straightforwardly.

Like most model-based inference, the validity of inference derived from the structure measurement error model relies on the assumed latent-variable model as well as the posited GLM. In the measurement error community there is a general concern with imposing models for unobserved covariates, as one can easily make inappropriate assumptions on unobservable covariates that often lead to misleading inference (Huang, Stefanski, and Davidian, 2006). The widely entertained GLMs for a binary response often assume one of the popular links such as logistic, probit, and complementary log-log. The choice of these popular links is mostly encouraged by ease of interpretation, the familiarity among practitioners, and its convenient implementation using standard statistical software. However, for one particular application, a link function outside of this popular suite of links may be able to capture the underlying association between the response and covariates more accurately. Li and Duan (1989) studied the properties of regression analysis under a misspecified link function in general regression settings. Czado and Santner (1992) focused on the effects of link misspecification on regression analysis based on GLMs for a binary response. Without considering measurement error in covariates, 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 biased and inefficient.

In this article, we address both sources of model misspecification and propose diagnostic procedures to assess these model assumptions. There are only a handful of diagnostic methods available for testing either one of these assumptions (e.g., Pregibon, 1980; Brown, 1982; Stukel, 1988; Huang, Stefanski, and Davidian, 2009), and most existing tests for GLM, with or without error-prone covariates, are omnibus tests designed for testing overall goodness-of-fit (GOF) rather than assessing specific assumptions of a hierarchical model (e.g., Tsiatis, 1980; Fowlkes, 1987; Hosmer and Lemeshow, 1989; le Cessie and van Houwelingen, 1991; Ma et al., 2011). To the best of our knowledge, there is no existing work that address the dual misspecification considered in our study. Huang, Stefanski, and Davidian (2006) proposed the so-called remeasurement method, referred to as RM henceforth, to detect latent-variable model misspecification in structural measurement error models. This method also has successes in testing latent-variable model assumptions in the bigger class of joint models (Huang, Stefanski, and Davidian, 2009), and was later improved to adapt to more challenging data structures (Huang, 2009). To detect link misspecification without involving error-prone covariates, Pregibon (1980) proposed a test derived from linearizing the discrepancy between the assumed link and the true link. His test was developed under the assumption that the assumed link and the true link belong to the same family, which can be a stringent assumption.

Moreover, his test fails easily if the local linear expansion of the true link about the assumed link is a poor approximation of the true link. For logistic regression models in the absence of measurement error, Hosmer et al. (1997) compared nine GOF tests for three types of model misspecification, including link misspecification, and found none of these tests have satisfactory power to detect link misspecification.

Inspired by the rationale behind RM, we propose a new diagnostic method initially aiming to detect link misspecification, called the reclassification method, or RC for short. This new method is described in Section 2, where we first define generic notations in a structural measurement error model, followed by a brief review of RM. Both RM and RC are motivated by theoretical findings on the effects of either type of misspecification on MLEs. For illustration purposes, we focus on one particular assumed structural measurement error model throughout the study and formulate a class of true flexible models. Under such formulation we present properties of the MLEs in the presence of one or both sources of misspecification in Section 3. In Section 4 we report finite-sample simulation studies to illustrate the performance of the proposed diagnostic procedures. Two real-life data examples are used to demonstrate the implementation of these methods in Section 5. Finally, discussions on our findings and follow-up research directions ensue in Section 6.

2 Models and Two Diagnostic Methods

2.1 Models

Denote by Y_i the binary response of subject i , for $i = 1, \dots, n$, and the true distribution of Y_i conditioning on covariates X_i is specified by a GLM,

$$P(Y_i = 1 | X_i; \beta) = H(\beta_0 + \beta_1^t X_i), \quad (1)$$

where $\beta = (\beta_0, \beta_1^t)^t$ is the vector of regression coefficients, and $H(s)$ is the inverse link function, assumed to be a nondecreasing and differentiable function of s . For a succinct exposition, we assume a scalar error-prone covariate X_i in the sequel, and the observed covariate, W_i , relates to X_i via a classical measurement error model (Carroll et al., 2006, Section 1.2), for $i = 1, \dots, n$,

$$W_i = X_i + U_i, \quad (2)$$

where $U_i \sim N(0, \sigma_u^2)$ is the nondifferential measurement error (Carroll et al., 2006, Section 2.5). Estimation of σ_u^2 is straightforward when replicate measures of each X_i ($i = 1, \dots, n$) are available (Carroll et al., 2006, equation (4.3)). For notational simplicity, σ_u^2 is assumed known in the majority of this article. Lastly, suppose that $\{X_i\}_{i=1}^n$ is a random sample from a distribution specified by the probability density function (pdf) $f_x^{(t)}(x; \tau)$, indexed by parameters τ . The three component models, (1), (2), and $f_x^{(t)}(x; \tau)$, constitute the structural measurement error model, based on which one has the correct likelihood function of the observed data for subject i ,

(Y_i, W_i) , given by $f_{Y,W}^{(t)}(Y_i, W_i; \Omega^{(t)}, \sigma_u^2) = \int \{H(\beta_0 + \beta_1 x)\}^{Y_i} \{1 - H(\beta_0 + \beta_1 x)\}^{1-Y_i} \sigma_u^{-1} \phi\{(W_i - x)/\sigma_u\} f_x^{(t)}(x; \tau) dx$, where $\phi(s)$ is the pdf of the standard normal distribution, and $\Omega^{(t)} = (\beta', \tau')$ is the vector of all unknown parameters under the correct model specification.

Suppose that one assumes the link function to be $J(s)$, which may differ from $H(s)$ in (1), and one posits a model for X_i with pdf give by $f_x(x; \eta)$, indexed by parameters η . Then one has the assumed likelihood function of the observed data for subject i , denoted by $f_{Y,W}(Y_i, W_i; \Omega, \sigma_u^2)$, similarly derived as above, where $\Omega = (\beta', \eta')$ is the p -dimensional vector of all unknown parameters under the assumed model.

2.2 Remeasurement Method and Reclassification Method

It was shown in Huang, Stefanski, and Davidian (2006) that, when the model for the true covariate, that is, the X -model, is misspecified, the MLE of β is usually inconsistent with bias depending on the measurement error variance. By exploiting this dependence, they proposed further contaminating $\{W_i\}_{i=1}^n$ to generate $W_{b,i}^* = W_i + \sqrt{\lambda} \sigma_u Z_{b,i}$, for $b = 1, \dots, B$, $i = 1, \dots, n$, where λ is a user-specified positive constant and $Z_{b,i}$'s are independent pseudo errors from $N(0, 1)$. Note that the measurement error variance associated with $\{W_{b,i}^*, b = 1, \dots, B\}_{i=1}^n$ is equal to $(1 + \lambda) \sigma_u^2$. They then constructed a test statistic based on the difference between the MLE of β , $\hat{\beta}$, computed using the raw data, $\{(Y_i, W_i)\}_{i=1}^n$, and the counterpart MLE, $\hat{\beta}_r$, obtained from the remeasured data, $\{(Y_i, W_i^*)\}_{i=1}^n$, where $W_i^* = (W_{1,i}^*, \dots, W_{B,i}^*)$, for $i = 1, \dots, n$. Take β_1 as an example, the test statistic associated with β_1 is defined by $T_{\beta_1} = (\hat{\beta}_1 - \hat{\beta}_{1r}) / \hat{v}_{\beta_1}$, where \hat{v}_{β_1} is an estimator of the standard error of $\hat{\beta}_1 - \hat{\beta}_{1r}$. Each so-constructed test statistic for a parameter in Ω follows a Student's t distribution with $n - p$ degrees of freedom asymptotically under the null hypothesis that the two MLEs being compared converge to the same limit as $n \rightarrow \infty$. If the value of a test statistic deviates significantly from zero, one finds evidence that the assumed latent-variable model is inappropriate. Derivations of the standard error estimator and the proof of the null distribution, omitted here, are given in Huang, Stefanski, and Davidian (2006).

It is assumed in this existing work that all aspects of the structural measurement error model are correctly specified except for the X -model. But one may legitimately question the adequacy of the assumed link in the GLM. And if the link is indeed misspecified, one may wonder if RM can also detect the link misspecification and how its ability to reveal latent-variable model misspecification is affected by this additional misspecification. As an important step in RM, pseudo measurement error are added to the observed covariates $\{W_i\}_{i=1}^n$ to produce the remeasured data. A natural extension of this idea is to add measurement error to the responses $\{Y_i\}_{i=1}^n$. For binary data, measurement error lead to misclassification of the binary responses. Parallel with adding noise to W to detect latent-variable model misspecification, we

propose to detect link misspecification by adding noise to Y , producing the so-called reclassified data. Now one may think of $\hat{\beta}_r$ as the MLE of β obtained from the reclassified data. If $\hat{\beta}$ is biased due to link misspecification, then $\hat{\beta}_r$ is usually also biased. If the bias of $\hat{\beta}_r$ depends on some parameter in the user-specified reclassification model according to which the reclassified data are created, then $\hat{\beta}_r$ can differ noticeably from $\hat{\beta}$. Such difference can serve as evidence of link misspecification. And test statistics like those constructed in RM can be used to quantify the significance of the difference. We refer to this strategy as the reclassification method, or, RC for short.

Under regularity conditions, the MLE of β follows a normal distribution asymptotically, despite the source of model misspecification (White, 1982) and the type of measurement error. Because both RM and RC rely on the discrepancy between the MLEs of β before and after pseudo measurement error are added (to W or Y), one important clue to answering the question, “Does RM/RC work?”, is the means of these asymptotic normal distributions associated with the MLEs from data with measurement error (in X or Y) in the presence of different model misspecification. The next section is devoted to studying these asymptotic quantities, i.e., the limiting MLEs of β .

3 Limiting Maximum Likelihood Estimators

3.1 Estimating Equations

Denote by β_m and β_c the limiting MLEs of β associated with the raw data and the reclassified data, respectively, as $n \rightarrow \infty$. By the theory of maximum likelihood estimation in the presence of model misspecification (White, 1982), β_m and β_c uniquely satisfy the following score equations respectively,

$$E_w [E_{Y|W} \{(\partial/\partial\beta)f_{Y,W}(Y_i, W_i; \Omega, \sigma_u^2)|_{\beta=\beta_m}\}] = 0, \quad (3)$$

$$E_w [E_{Y^*|W} \{(\partial/\partial\beta)f_{Y^*,W}(Y_i^*, W_i; \Omega, \sigma_u^2)|_{\beta=\beta_c}\}] = 0, \quad (4)$$

where $f_{Y^*,W}(Y_i^*, W_i; \Omega, \sigma_u^2)$ is the likelihood of the reclassified data for subject i , (Y_i^*, W_i) , and the subscripts attached to “ E ” signify that the expectations are defined with respect to the relevant true model.

In order to focus on inference for β , we treat the parameters in the assumed X -model, η , as known constants in (3) and (4). Although in practice one has to estimate η along with β , this seemingly unrealistic treatment of η does not make the follow-up theoretical findings less practically valuable if η can be estimated consistently (in some sense). Consistent estimation of η in the presence of model misspecification is often possible in many scenarios. For example, when both the assumed and the true X -models can be fully parameterized via some moments (included in η) up to a finite order, the interpretation of η remains meaningful even if the assumed X -model

differs from the true model, and hence one can still conceptualize the “true” value of η , which are simply the moments of the true X -distribution. Moreover, such η usually can be consistently estimated, say, using the method of moments based on $\{W_i\}_{i=1}^n$, even in the presence of dual misspecification.

In general, the above estimating equations cannot be solved explicitly, thus closed form expressions of their solutions, β_m and β_c , are usually unattainable. Without sacrificing too much the generality of the theoretical investigation, we next formulate the assumed model and true models that make these limiting MLEs more transparent.

3.2 Assumed and True Models

For tractability, we fix the assumed structural measurement error model at the probit-normal model, which is one of the favorite toy examples entertained in the measurement error literature. In this model, one posits a probit link in the primary model (1) and assumes $X \sim N(\mu_x, \sigma_x^2)$. As for the true model, we formulate a class of the so-called mixture-probit-normal models, which contains the probit-normal model as a special member. In this class of true models, the link function $H(s)$ is the cdf of a two-component mixture normal, referred to as the mixture probit. With a mixture probit link, the primary model is a GLM given by

$$P(Y_i = 1|X_i; \beta) = \alpha \Phi\left(\frac{\beta_0 + \beta_1 X_i - \mu_1}{\sigma_1}\right) + (1 - \alpha) \Phi\left(\frac{\beta_0 + \beta_1 X_i - \mu_2}{\sigma_2}\right), \quad (5)$$

where $\alpha \in [0, 1]$, μ_k and $\sigma_k > 0$ ($k = 1, 2$) are chosen such that the corresponding mixture normal, $\alpha N(\mu_1, \sigma_1^2) + (1 - \alpha)N(\mu_2, \sigma_2^2)$, is of zero mean and unit variance. The true X -model in this class is a mixture normal.

To achieve explicit likelihood for the reclassified data without being overly restrictive in the creation of reclassified data, we consider reclassification models of the form $P(Y_i^* = Y_i|W_i) = \pi_i$, for $i = 1, \dots, n$, according to which the reclassified responses, $\{Y_i^*\}_{i=1}^n$, are generated. Combining the assumed raw-data likelihood, $f_{y,w}(Y_i, W_i; \Omega, \sigma_u^2)$, and the reclassification model yields the likelihood of (Y_i^*, W_i) under the probit-normal model, $f_{y^*,w}(Y_i^*, W_i; \Omega, \sigma_u^2)$.

Under the formulated assumed and true models, all needed ingredients for deriving the score equations in (3) and (4) become available in closed form. These ingredients include the true mean of Y_i and Y_i^* given W_i , the assumed-model likelihood for the raw data, $f_{y,w}(Y_i, W_i; \Omega, \sigma_u^2)$, and that for the reclassified data, $f_{y^*,w}(Y_i^*, W_i; \Omega, \sigma_u^2)$, the true-model likelihood for both types of data, $f_{y,w}^{(t)}(Y_i, W_i; \Omega^{(t)}, \sigma_u^2)$ and $f_{y^*,w}^{(t)}(Y_i^*, W_i; \Omega^{(t)}, \sigma_u^2)$. The explicit expressions of these quantities are provided in Appendix A in the Supplementary Materials. Some interesting findings regarding β_m and β_c are presented next, in which we only consider cases where $\beta_1 \neq 0$. The special case with $\beta_1 = 0$ is discussed in Appendix B in the Supplementary Materials, where the expressions of β_m and β_c are derived.

This is a rare case where (3) and (4) can be solved explicitly, and also a rare case where the MLE of β_1 is consistent despite the type of model misspecification. When $\beta_1 \neq 0$, although (3) and (4) cannot be solved explicitly, we are able to make use the aforementioned intermediate results in Appendix A to study the limiting MLEs.

3.3 Limiting MLEs from Data with Measurement Error Only in X

Fixing the assumed model at the probit-normal model, we consider combinations of five true links and five true X -distributions in the formulation of the true model. The five true links are, (L0) probit link, and four mixture probit links with the following parameter configurations: (L1) $\alpha = 0.3$, $\mu_1 = 0.3$, $\sigma_1 = 0.1$; (L2) $\alpha = 0.3$, $\mu_1 = -0.3$, $\sigma_1 = 0.1$; (L3) $\alpha = 0.7$, $\mu_1 = 0.5$, $\sigma_1 = 0.2$; (L4) $\alpha = 0.7$, $\mu_1 = -0.5$, $\sigma_1 = 0.2$. The upper panels of Figure 1 depict these five links. For two link functions, $H_1(s)$ and $H_2(s)$, we say that $H_1(s)$ and $H_2(s)$ are symmetric of each other if $H_1(s) = 1 - H_2(-s)$. Among the four mixture probit links, (L1) and (L2) are symmetric of each other, and (L3) and (L4) are symmetric of each other, with the latter two links deviating from probit more than the former two. The five true X -distributions are, (D0) $N(0, 1)$, and four mixture normals with mean zero and variance one formulated by varying the mixing proportion ζ , skewness ξ , and excessive kurtosis κ as follows: (D1) $\zeta = 0.3$, $\xi = -1$, $\kappa = 2$; (D2) $\zeta = 0.3$, $\xi = 1$, $\kappa = 2$; (D3) $\zeta = 0.1$, $\xi = -1.5$, $\kappa = 2$; (D4) $\zeta = 0.1$, $\xi = 1.5$, $\kappa = 2$. The lower panels of Figure 1 show the pdf's of these five distributions. Among the four mixture normal distributions, (D1) and (D2) are symmetric of each other, and (D3) and (D4) are symmetric of each other, with the latter pair deviating from normal further than the former pair. In the true GLM in (5), we set $\beta_0 = 0$ and $\beta_1 = 1$. For ease of presentation, we use “ \wedge ” to connect a true X -model with a true link to refer to a true model specification. For example, (D1) \wedge (L3) refers to the true model with X following a distribution specified by (D1) and the link configured according to (L3).

Under each of the above true model specifications, we numerically solve (3) for β_m . Figure 2 presents β_m under different true models as σ_u^2 increases from 0 to 1. This range of σ_u^2 yields a reliability ratio ω that drops from 1 to 0.5, where $\omega = \sigma_x^2 / (\sigma_x^2 + \sigma_u^2)$. The top panels of Figure 2, where the true X -model coincides with the assumed, show that β_m only changes slightly as σ_u^2 increases in the presence of link misspecification. This suggests that, unless information in both the raw data and the remeasured data are rich enough to allow detection of the weak dependence of β_m on σ_u^2 , RM will have low power to detect link misspecification despite the amount of bias in β_m due to link misspecification. When the true X -model deviates from normal (see the middle and the bottom panels of Figure 2), although the dependence of β_{1m} on σ_u^2 is stronger than before, β_{1m} changes noticeably mainly over a narrow range of σ_u^2 . This phenomenon for cases with dual misspecification indicates that, although RM has been shown to be effective in diagnosing latent-variable model misspecification, its power in this regard can be substantially compromised by the coexistence of link misspecification.

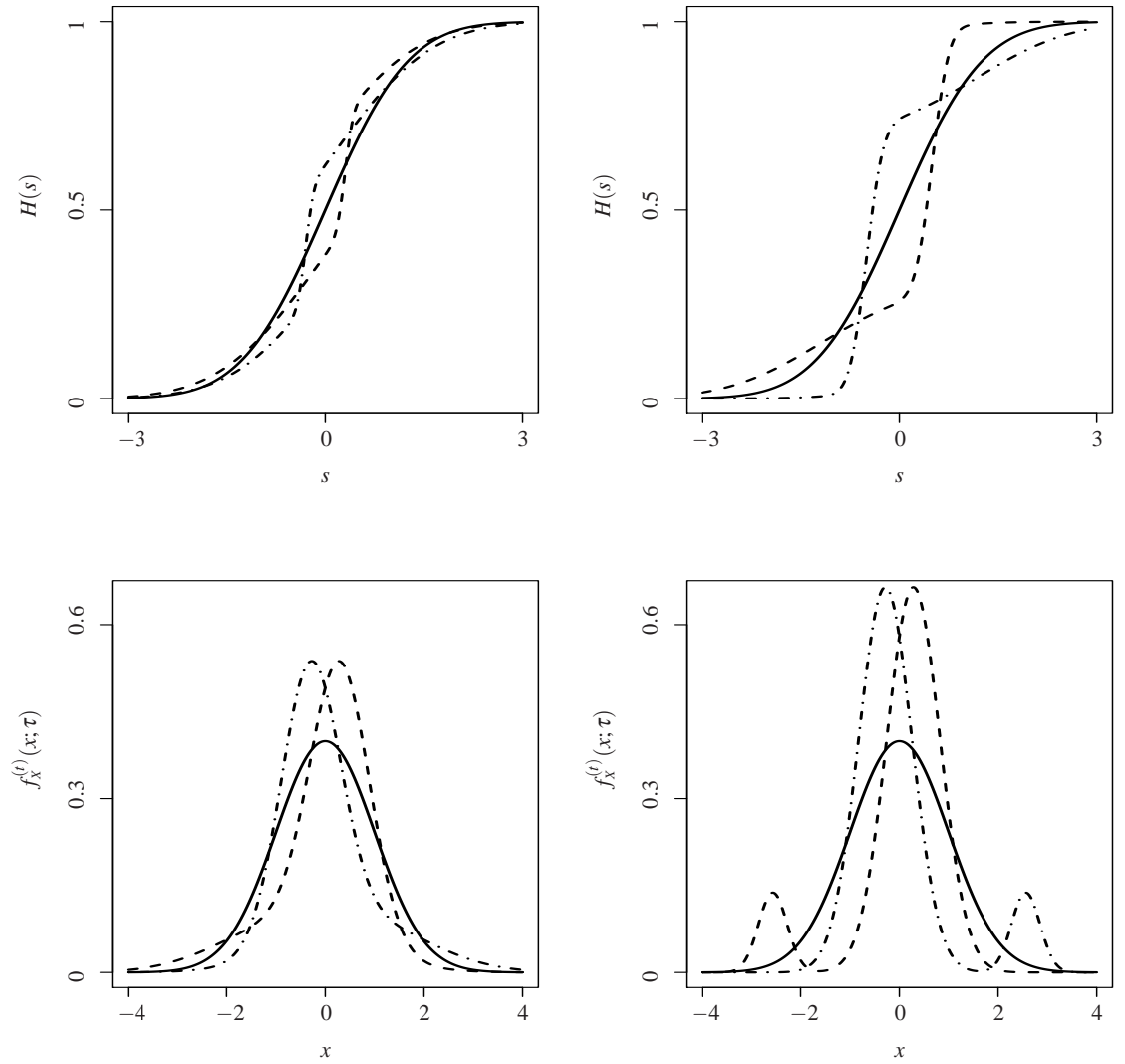


Fig. 1 Upper panels give four mixture probit links formulated in Section 3.3, where the upper left panel gives link (L1) (dashed line) and link (L2) (dot-dashed line), and the upper right panel gives link (L3) (dashed line) and link (L4) (dot-dashed line). Solid lines are the probit link. Lower panels show four mixture normal density functions formulated in Section 3.3, where the lower left panel gives distributions (D1) (dashed line) and (D2) (dot-dashed line), and the lower right panel gives distributions (D3) (dashed line) and (D4) (dot-dashed line). Solid lines are the density function of $N(0, 1)$.

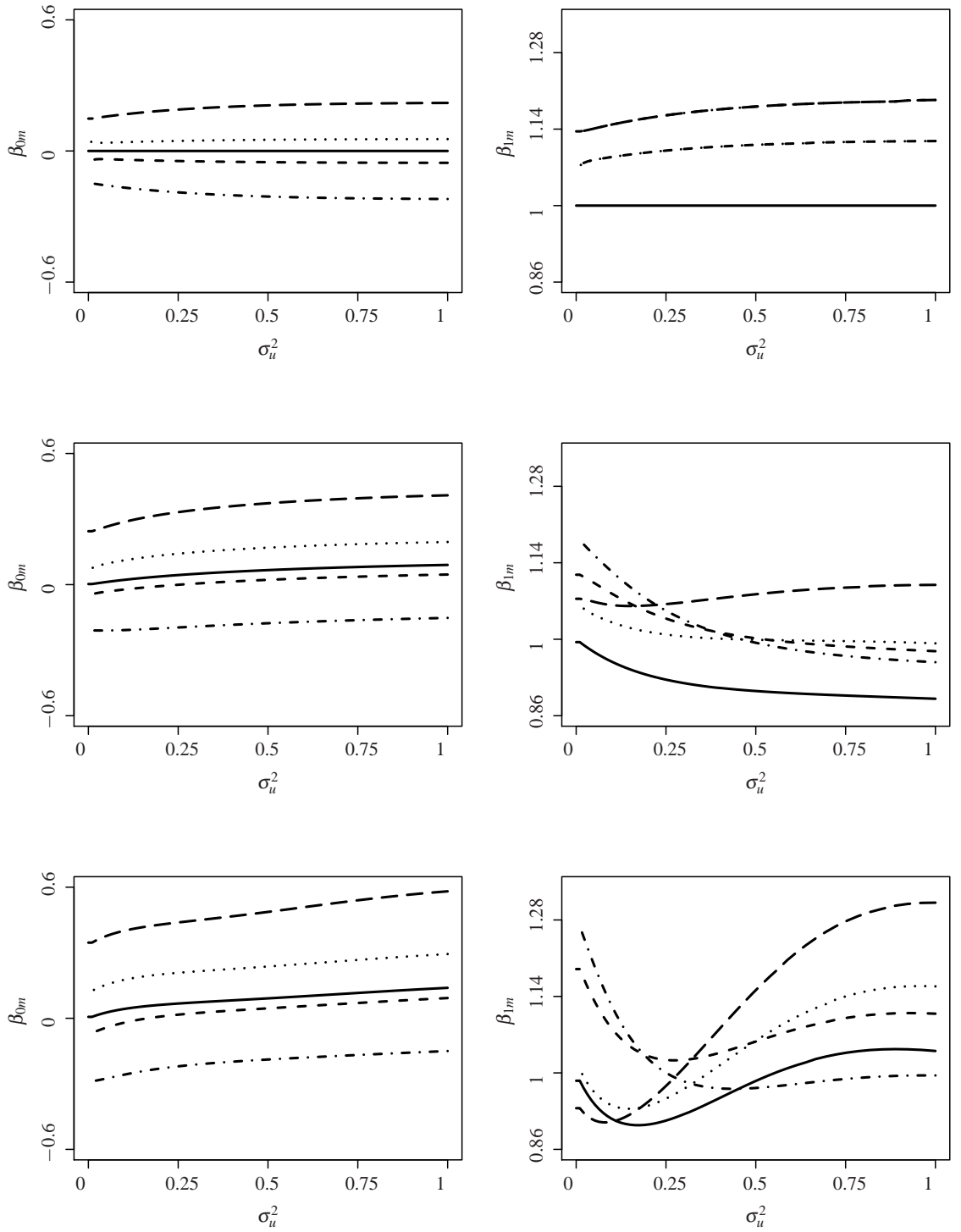


Fig. 2 Plots of β_{0m} (left column) and β_{1m} (right column) versus σ_u^2 when fixing the true X-model at $N(0, 1)$ (top row), D1 (middle row), and D3 (bottom row), respectively, then varying the true link among the five links: probit (solid lines), L1 (short dashed lines), L2 (dotted lines), L3 (dot-dashed lines), and L4 (long dashed lines).

Besides Figure 2, we show analytically in Appendix C in the Supplementary Materials that, under certain conditions, β_{1m} is unchanged by a symmetric flip of either the true X -distribution or the true link, and only β_{0m} is affected. This property is stated next.

Proposition 3.1. *Let $f_1(x)$ and $f_2(x)$ be two pdf's specifying two true X -distributions that are symmetric of each other, and let $H_1(s)$ and $H_2(s)$ be two true links that are symmetric of each other. Denote by $\beta_m^{(jk)}$ the limiting MLE of β based on data with measurement error only in X when the true model is $f_j(x) \wedge H_k(s)$, for $j, k = 1, 2$. If $E(X) = \beta_0 = 0$, then $\beta_{0m}^{(11)} = -\beta_{0m}^{(22)}$ and $\beta_{1m}^{(11)} = \beta_{1m}^{(22)}$.*

Note that Proposition 3.1 includes two special cases: one is when $H_1(s) \neq H_2(s)$ and $f_1(x) = f_2(x) = f(x)$, where $f(x)$ is a pdf symmetric around zero; the other is when $f_1(x) \neq f_2(x)$ and $H_1(s) = H_2(s) = H(s)$, where $H(s)$ is the cdf associated with a distribution symmetric around zero. This is because $f_1(x) = f_2(x) = f(x)$ implies $f_1(x) = f_2(-x)$, since $f(x) = f(-x)$, and thus $f_1(x)$ and $f_2(x)$ are symmetric of each other. Similarly, $H_1(s) = H_2(s) = H(s)$ implies $H_1(s) = 1 - H_2(-s)$, as $H(s) = 1 - H(-s)$, hence $H_1(s)$ and $H_2(s)$ are symmetric of each other. This proposition implies that β_{0m} can distinguish two true X -models that are symmetric of each other, and can also tell apart two true links that are symmetric of each other. For the purpose of model diagnosis, one can exploit this and other properties of β_{0m} to obtain a directional test based on RM that can identify the direction of model misspecification. This potential of RM is supported by the following observations of β_{0m} under the conditions stated in Proposition 3.1:

- (M1) Despite the skewness of the true link, when the true X -model is not normal, β_{0m} is increasing in σ_u^2 when the true X -model is left-skewed, and it is decreasing in σ_u^2 when the true X -model is right-skewed.
- (M2) When the true X -model is normal and the true link is not probit, β_{0m} is increasing in σ_u^2 when the true link is right-skewed, and it is decreasing in σ_u^2 when the true link is left-skewed.

The middle and bottom panels of Figure 2, which are associated with two left-skewed true X -models, illustrate the first half of (M1), and the second half of (M1) is indicated by Proposition 3.1. Viewing a link function as a cdf, we say that a link function is left-skewed if the corresponding pdf is left-skewed. Among the four considered mixture probit links, (L1) and (L3) are left-skewed and (L2) and (L4) right-skewed. The top panel of Figure 2 illustrates (M2). In Section 4.4, we propose a directional test based on RM that utilizes the properties of β_{0m} summarized in (M1) and (M2).

3.4 Limiting MLEs Based on Reclassified Data

Under the same configurations for the assumed/true models as in Section 3.3, we solve (4) numerically for β_c based on reclassified data generated according to the

reclassification model $P(Y_i^* = Y_i|W_i) = \Phi(W_i + \gamma)$, for $i = 1, \dots, n$, where γ is a constant. Figure 3 presents β_c when $\gamma = 0$, which shows stronger dependence on σ_u^2 compared to Figure 2, especially for β_{0c} . This implies that, if one applies RM to the reclassified data, T_{β_0} can be much more significant than the counterpart test statistic from RM only (without adding noise to Y).

Viewing β_c as a function of γ and thinking of β_c as $\beta_c(\gamma)$ symbolically, Figure 4 presents $\beta_c(-2) - \beta_c(0)$ as σ_u^2 varies. This figure reveals that the changes in β_c as γ changes can be substantial when σ_u^2 is small. This phenomenon suggests that RC alone (without adding further noise to W) can have good power to detect X -model misspecification or link misspecification, and the power is higher when the error contamination in X is milder. If the X -model is correctly specified, both β_{0c} and β_{1c} can change substantially as γ varies when σ_u^2 is fixed at a lower level, including 0. Hence, in the absence of measurement error in X , and thus without involving RM, RC alone is expected to possess some power to detect moderate to severe link misspecification.

In Appendix D in the Supplementary Materials, we show that, if the reclassification model is $P(Y_i^* = Y_i|W_i) = \pi(W_i)$, where $\pi(t)$ is an even function, then β_c has the same property of β_m under the same conditions stated in Proposition 3.1.

4 Testing Procedures

The investigation in Section 3 on the limiting MLEs of β based on data with measurement error in X or Y in the presence of X -model misspecification or link misspecification are helpful for understanding the operating characteristics of the test statistics, T_{β_0} and T_{β_1} . Although the assumed model is the probit-normal model throughout Section 3, many findings there, such as some phenomenon summarized in Proposition 3.1, also have empirical justifications from simulation studies carried out under other assumed models, such as the logit-normal model. More importantly, when the true model is not in the class of mixture-probit-mixture-normal, but the assumed model is probit-normal, most of the phenomena described in Section 3 that motivate the upcoming testing strategies are still observed in extensive simulations we carried out. This latter point is practically more important because, although one cannot control what the true model looks like in reality, one can choose an assumed model for the purpose of “guessing” some features of the true model. Some of these simulation studies are presented in this section. More specifically, in this section, besides computing finite sample MLEs under different assumed models, we first study via simulation the operating characteristics of the aforementioned test statistics resulting from three diagnostic methods: first, RM; second, RC; third, a hybrid method that combines RM and RC. Then we propose more informative testing procedures that can disentangle two sources of misspecification and point at the direction of misspecification.

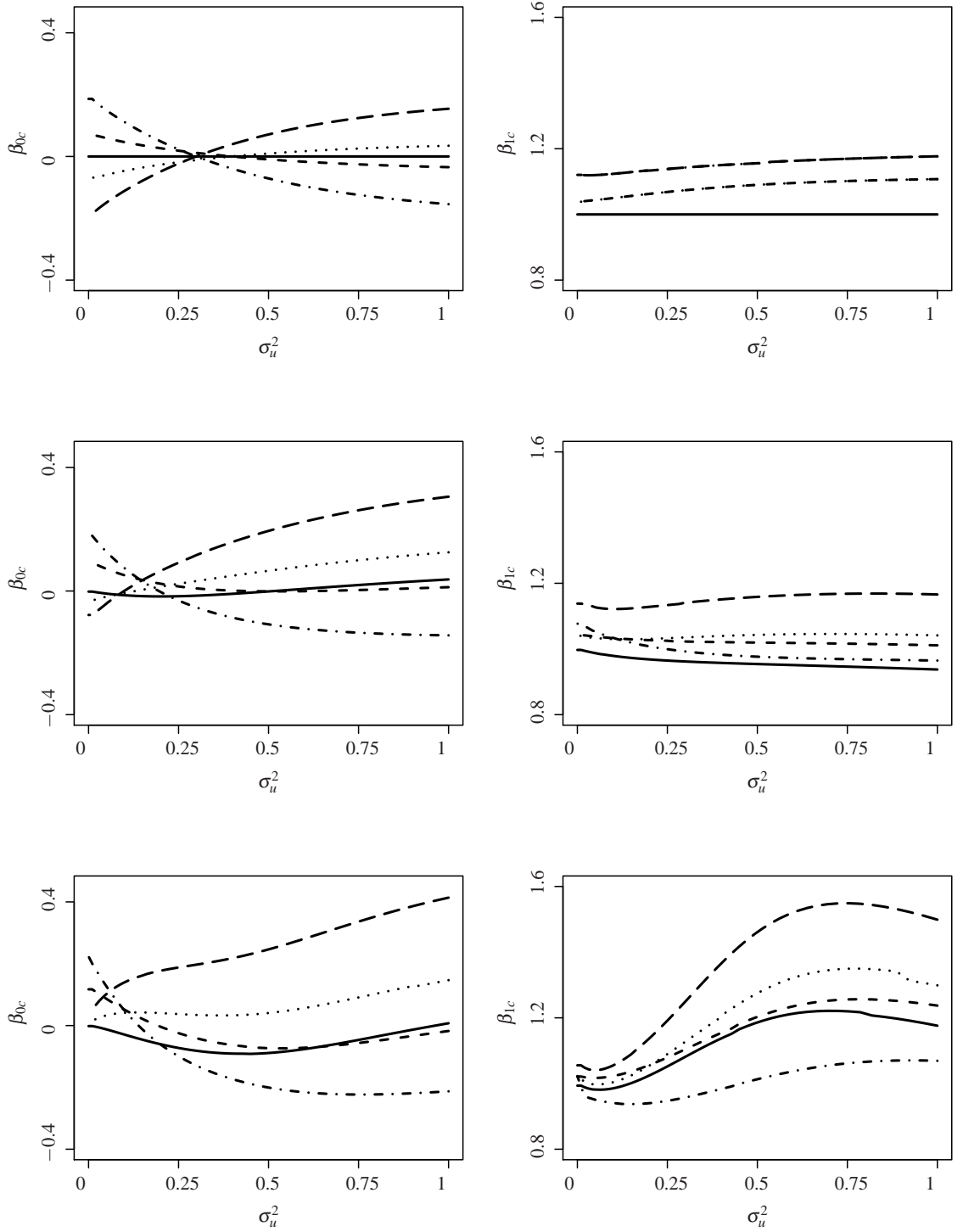


Fig. 3 Plots of β_{0c} (left column) and β_{1c} (right column) when $P(Y_i^* = Y_i|W_i) = \Phi(W_i)$, for $i = 1, \dots, n$, versus σ_u^2 , with the true X -model being $N(0, 1)$ (top), (D1) (middle), and (D3) (bottom), and the true link being probit (solid lines), (L1) (short dashed lines), (L2) (dotted lines), (L3) (dot-dashed lines), and (L4) (long dashed lines).

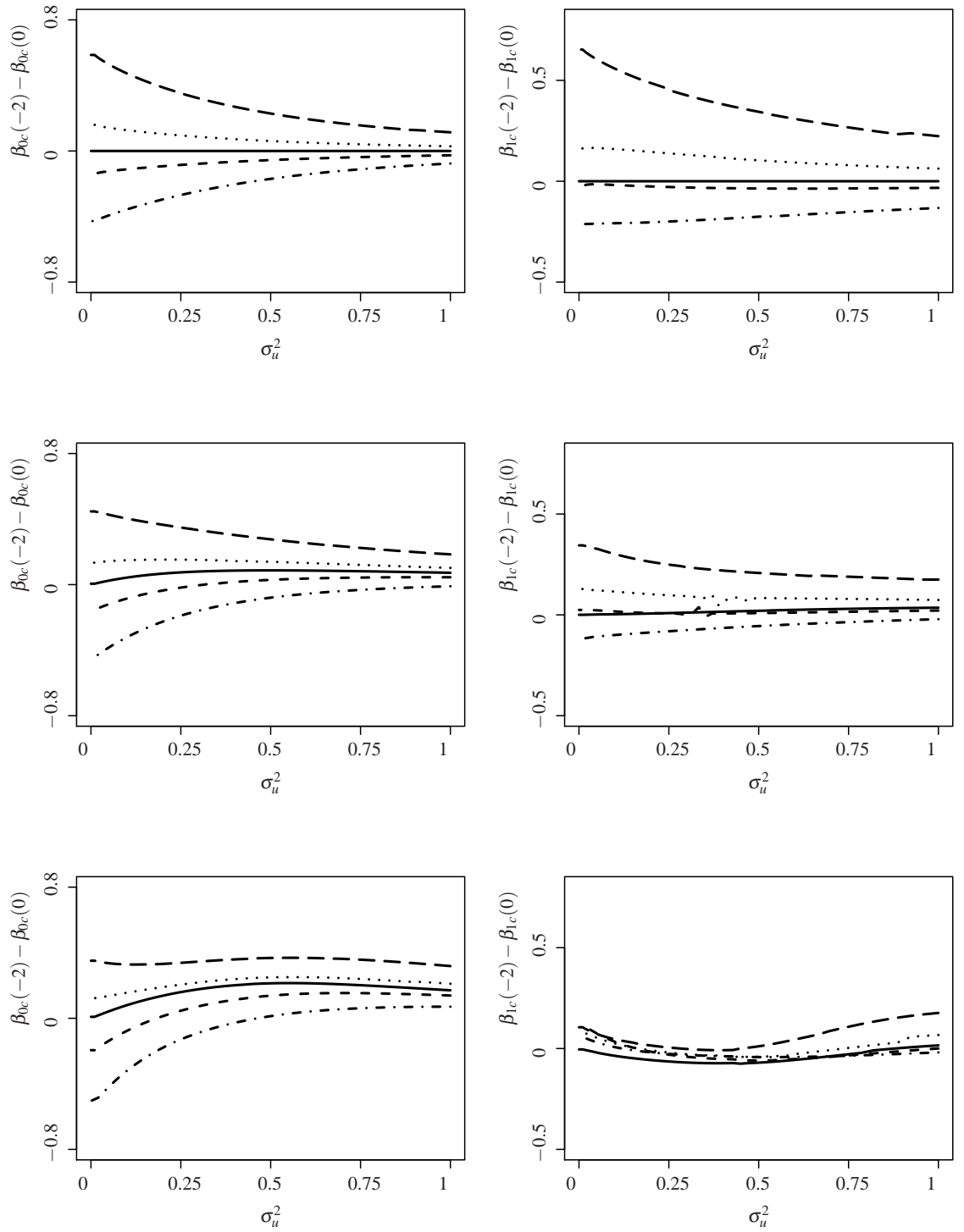


Fig. 4 Differences, $\beta_{0c}(-2) - \beta_{0c}(0)$ (left column) and $\beta_{1c}(-2) - \beta_{1c}(0)$ (right column), versus σ_u^2 with the true X -model being $N(0, 1)$ (top), (D1) (middle), and (D3) (bottom), and the true link being probit (solid lines), (L1) (short dashed lines), (L2) (dotted lines), (L3) (dot-dashed lines), and (L4) (long dashed lines).

4.1 Simulation Design

Fixing the sample size n at 500, we create the raw data, $\{(Y_i, W_i)\}_{i=1}^n$, from different true models resulting from varying four factors in the simulation experiments. The first factor is the assumed model, taking two levels: probit-normal and logit-normal. The second factor is the true X -model, taking five levels (D0)–(D4) as defined in Section 3.3. The third factor is the true link function. When the assumed model is probit-normal, we consider the five true links, (L0)–(L4), i.e., the probit and mixture-probit links formulated in Section 3.3. When the assumed model is logit-normal, two generalized logit links (Stukel, 1988) are used as the true links, referred to as (L5) and (L6). These two generalized logit links are symmetric of each other, with (L5) left-skewed and (L6) right-skewed, as depicted in Figure 5. The fourth factor is the value of σ_u^2 used to generate $\{W_i\}_{i=1}^n$ according to (2), with four values leading to reliability ratio ω ranging from 0.7 to 1 at increments of 0.1. Under each simulation setting, 1000 Monte Carlo (MC) replicates are generated. After each replicate is generated, assuming a probit-normal model, we compute T_{β_0} and T_{β_1} associated with the aforementioned three diagnostic methods.

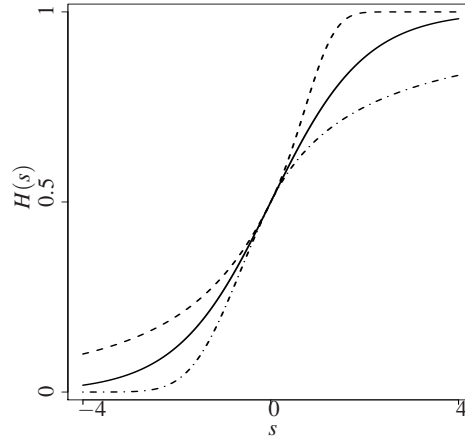


Fig. 5 Two generalized logit links, (L5) (dashed line) and (L6) (dot-dashed line), in comparison with the logit link (solid line).

When implementing RM, $\hat{\beta}_r$ is the MLE from the remeasured data $\{(Y_i, W_i^*)\}_{i=1}^n$, where $W_i^* = (W_{1,i}^*, \dots, W_{B,i}^*)$, in which $W_{b,i}^* = W_i + \sigma_u Z_{b,i}$, with $Z_{b,i} \sim N(0, 1)$ independent across $b = 1, \dots, B$, $i = 1, \dots, n$, and $B = 100$. When carrying out RC, $\hat{\beta}_r$ is the estimate computed from the reclassified data, $\{(Y_i^*, W_i)\}_{i=1}^n$, where the reclassified responses, $Y_i^* = (Y_{1,i}^*, \dots, Y_{B,i}^*)$, for $i = 1, \dots, n$, are generated according to $P(Y_{b,i}^* = Y_i | W_i) = \Phi(W_i)$. When employing the hybrid method, we first generate $\{W_{b,i}^*, b = 1, \dots, B\}_{i=1}^n$ as in RM above, then the reclassified responses are generated

according to $P(Y_{b,i}^* = Y_i | W_{b,i}^*) = \Phi(W_{b,i}^*)$; finally one obtains $\hat{\beta}_r$ based on the hybrid data that have measurement error in both X and Y , $\{(Y_{b,i}^*, W_{b,i}^*), b = 1, \dots, B\}_{i=1}^n$. Using a significance level of 0.05, we monitor how often the value of a test statistic turns out significant, leading to rejection of a null hypothesis, which states that two MLEs being compared in the test statistic have the same limit as $n \rightarrow \infty$.

4.2 Simulation Results

Table 1 presents the rejection rate of each test statistic under each simulation setting across 1000 MC replicates for a representative subset of all considered true-model configurations. This subset of true models includes, (D3) \wedge (L0), (D0) \wedge (L3), (D3) \wedge (L3), (D4) \wedge (L3), and (D3) \wedge (L4). Among these five true-model configurations, (D3) \wedge (L0) represents the scenario where only the X -model is misspecified, (D0) \wedge (L3) represents the case where only the link is misspecified, and the latter three configurations represent cases with dual misspecification. Albeit not included in Table 1, we observe rejection rates for all tests well controlled at around 0.05 when the true model is (D0) \wedge (L0), that is, when there is no model misspecification. Some noteworthy observations regarding RM and RC from the simulation are summarized in the following three remarks.

Remark 1: When $\sigma_u^2 = 0$, that is, the covariate is measured without error ($\omega = 1$), RM can detect neither source of misspecification. This is due to the definition of the remeasured data, $W_{b,i}^* = W_i + \sqrt{\lambda} \sigma_u Z_{b,i}$, resulting in the remeasured data identical to the raw data when $\sigma_u^2 = 0$. In contrast, when $\sigma_u^2 = 0$, RC has impressive power to detect link misspecification, whether or not the X -model is also misspecified.

Remark 2: When $\sigma_u^2 \neq 0$, the power of RM to detect X -model misspecification surpasses that of RC if this is the only source of misspecification; but when only the link is misspecified, the test based on T_{β_0} from RC is the clear winner in detecting link misspecification, whose power increases as σ_u^2 decreases.

Remark 3: Although RM is designed for detecting X -model misspecification, and RC is proposed aiming at detecting link misspecification, each of them can be influenced in nontrivial ways by the other source of misspecification. Take RM as an example. When only the X -model is misspecified, such as case (D3) \wedge (L0) in Table 1, RM is expectedly effective in picking up this type of misspecification. But its power is mostly weakened by the additional link misspecification as in case (D3) \wedge (L3). Note that, when the true model is (D3) \wedge (L3), the directions of the two misspecification are the same in the sense that the true X -model is left-skewed and so is the true link. This tampering effect on the power of RM due to the added link misspecification is not observed for T_{β_0} when the dual misspecification are of opposite directions, such as in case (D3) \wedge (L4), although the power of T_{β_1} is substantially compromised there. Similar nontrivial patterns are observed for RC when X -model

misspecification is added on top of link misspecification. In summary, whether or not the added misspecification compromises the power of a method to detect the type of misspecification it is originally designed for depends on how the two types of misspecification interact.

Finally, the hybrid method is the same as RC when $\sigma_u^2 = 0$. And, according to Table 1, when $\sigma_u^2 \neq 0$, the hybrid method performs similarly as RC when only the link is misspecified. In other cases, the power of the hybrid method mostly lies between that of RM and RC. We recommend use the hybrid method with caution due to the amount of information loss when creating the hybrid data.

Table 1 Rejection rates across 1000 Monte Carlo replicates of each test statistic under each testing procedure considered in Section 4 at different levels of reliability ratio ω . “HB” refers to the hybrid method

True model		$\omega = 0.7$			$\omega = 0.8$			$\omega = 0.9$			$\omega = 1$		
Assumed model: Probit-Normal													
		RM	RC	HB	RM	RC	HB	RM	RC	HB	RM	RC	HB
(D3) \wedge (L0)	T_{β_0}	0.99	0.53	0.85	1.00	0.40	0.75	1.00	0.23	0.38	0	0.09	0.09
	T_{β_1}	0.47	0.51	0.39	0.81	0.52	0.41	0.93	0.34	0.23	0	0.07	0.07
(D0) \wedge (L3)	T_{β_0}	0.18	0.75	0.72	0.31	0.92	0.92	0.48	0.98	0.99	0	1.00	1.00
	T_{β_1}	0.04	0.08	0.06	0.05	0.09	0.08	0.08	0.08	0.09	0	0.06	0.06
(D3) \wedge (L3)	T_{β_0}	0.82	0.09	0.08	0.88	0.44	0.16	0.76	0.80	0.90	0	0.70	0.70
	T_{β_1}	0.58	0.09	0.05	0.67	0.11	0.16	0.59	0.36	0.62	0	0.63	0.63
(D4) \wedge (L3)	T_{β_0}	1.00	0.89	0.97	1.00	0.89	0.96	1.00	0.96	0.96	0	1.00	1.00
	T_{β_1}	0.03	0.53	0.66	0.10	0.74	0.83	0.45	0.81	0.84	0	0.80	0.80
(D3) \wedge (L4)	T_{β_0}	1.00	0.85	0.97	1.00	0.89	0.96	1.00	0.95	0.96	0	1.00	1.00
	T_{β_1}	0.03	0.55	0.69	0.11	0.74	0.82	0.45	0.82	0.83	0	0.74	0.74
Assumed model: Logit-Normal													
		RM	RC	HB	RM	RC	HB	RM	RC	HB	RM	RC	HB
(D0) \wedge (L5)	T_{β_0}	0.03	0.37	0.40	0.03	0.57	0.58	0.04	0.76	0.76	0	0.92	0.92
	T_{β_1}	0.04	0.06	0.07	0.04	0.06	0.08	0.04	0.09	0.10	0	0.10	0.10
(D3) \wedge (L5)	T_{β_0}	0.67	0.05	0.18	0.89	0.09	0.06	0.95	0.32	0.18	0	0.87	0.87
	T_{β_1}	0.04	0.08	0.14	0.08	0.06	0.08	0.13	0.04	0.04	0	0.12	0.12
(D4) \wedge (L5)	T_{β_0}	0.81	0.87	0.98	0.98	0.90	0.98	1.00	0.91	0.97	0	0.90	0.90
	T_{β_1}	0.36	0.78	0.75	0.67	0.85	0.87	0.91	0.87	0.88	0	0.74	0.74
(D3) \wedge (L6)	T_{β_0}	0.73	0.86	0.99	0.94	0.89	0.98	1.00	0.87	0.95	0	0.80	0.80
	T_{β_1}	0.29	0.68	0.73	0.56	0.77	0.84	0.82	0.75	0.80	0	0.62	0.62

4.3 Sequential Tests

Although we caution use of the hybrid method in practice, sequentially using test results from RM and those from RC can help to disentangle two types of misspecification. We now illustrate some sequential testing procedures when the covariate is measured with error. To distinguish the test statistics from two methods, denote by $T_{\theta}^{(m)}$ and $T_{\theta}^{(c)}$ the test statistics associated with RM and RC, respectively, where θ denotes a generic parameter. Suppose one implements RM, with only W -data further contaminated, and then implements RC, with only Y -data contaminated (and the W -data left as originally observed). Implementing these two methods sequentially yields four test statistics of interest, $T_{\beta_0}^{(m)}$, $T_{\beta_1}^{(m)}$, $T_{\beta_0}^{(c)}$, and $T_{\beta_1}^{(c)}$. In light of the operating characteristics of these test statistics revealed in Section 4.2, we consider the following three sequential testing strategies.

First, if $T_{\beta_0}^{(m)}$ is highly significant and $T_{\beta_0}^{(c)}$ is insignificant, one may interpret this as evidence that the X -model is misspecified and the assumed link may be adequate for the observed data. For instance, when the true model is $(D3) \wedge (L0)$, using this testing criterion, one concludes “only the X -model is misspecified” 55%, 70%, and 84% of the time when $\omega = 0.7, 0.8, 0.9$, respectively, based on the simulation results in Section 4.2. When summarizing the preceding rejection rates, we apply the Bonferroni correction for multiple testing and use a significance level of $0.025 (= 0.05/2)$ now that two test statistics are used simultaneously.

Second, if $T_{\beta_1}^{(m)}$ turns out insignificant whereas $T_{\beta_0}^{(c)}$ is highly significant, one may view this as indication that the assumed X -model may be appropriate but the assumed link is inadequate. Revisiting the simulation results in Section 4.2, when the true model is $(D0) \wedge (L3)$, using this sequential testing strategy, one concludes “only the link is misspecified” 67%, 86%, and 94% of the time when $\omega = 0.7, 0.8, 0.9$, respectively.

Third, having observed promising power from the above two sequential tests, one would hope that having both $T_{\beta_0}^{(m)}$ and $T_{\beta_0}^{(c)}$ significant can be interpreted as an indication of dual misspecification. Unfortunately, due to the complicated interaction between the two misspecification described in *Remark 3* in Section 4.2, this criterion is a reliable indicator of dual misspecification only when two misspecification are of opposite directions. For example, when the true model is $(D4) \wedge (L3)$, the criterion of both $T_{\beta_0}^{(m)}$ and $T_{\beta_0}^{(c)}$ being significant is met 79%, 85%, and 93% of the time across 1000 MC replicates when $\omega = 0.7, 0.8, 0.9$, respectively. Similar high power is also observed when the true model is $(D3) \wedge (L4)$. However, if the true model is $(D3) \wedge (L3)$, the rejection rates according to this same criterion drop to 1%, 13%, and 29% when $\omega = 0.7, 0.8, 0.9$, respectively.

Despite the complication arising from dual misspecification, empirical evidence from the above three sequential tests give much encouragement to use the combination of two tests from two diagnostic methods, such as $T_{\beta_0}^{(m)}$ (or $T_{\beta_1}^{(m)}$) and $T_{\beta_0}^{(c)}$, in order to learn more from the data regarding the two model assumptions.

Table 2 Rejection rates associated with a one-sided test based on $T_{\beta_0}^{(m)}$ at significance level 0.05 under different true model configurations defined in Section 3.3 at different levels of reliability ratio ω . Codes beneath the true model codes, [L] and [R], indicate left-sided and right-sided tests, respectively

ω	(D0) \wedge (L1) [R]	(D0) \wedge (L3) [R]	(D0) \wedge (L4) [L]	(D4) \wedge (L0) [R]	(D3) \wedge (L3) [L]
0.7	0.09	0.29	0.26	1.00	0.95
0.8	0.10	0.45	0.42	1.00	0.98
0.9	0.13	0.62	0.60	1.00	0.96
ω	(D1) \wedge (L0) [L]	(D3) \wedge (L0) [L]	(D1) \wedge (L1) [L]	(D2) \wedge (L1) [R]	(D4) \wedge (L3) [R]
0.7	0.87	1.00	0.83	0.93	1.00
0.8	0.98	1.00	0.95	1.00	1.00
0.9	1.00	1.00	0.99	1.00	1.00

4.4 Directional Tests

The properties of β_{0m} described in (M1)–(M2) in Section 3.3 suggest that the sign of $T_{\beta_0}^{(m)}$ can indicate in which direction the true X -model deviates from normal or the true link function deviates from probit. More specifically, if there is strong evidence against a normal X -distribution, then, despite what the true link is, a significantly negative (positive) $T_{\beta_0}^{(m)}$ implies that the true X -distribution is left-skewed (right-skewed). This is supported by (M1). On the other hand, suppose one has evidence to suggest that the assumed normal X -model is likely appropriate, but suspects that the assumed probit link may be inadequate, then one further gains evidence to support a right-skewed link if $T_{\beta_0}^{(m)} < 0$, and left-skewed otherwise. This is justified by (M2).

As empirical evidence, Table 2 presents the rejection rates (at significance level 0.05) from the same simulation study described in Section 4.1 but associated with a one-sided test based on $T_{\beta_0}^{(m)}$, assuming one knows a priori the right side of the test (as we do in simulations). The high rejection rates for the cases with X -model misspecification tabulated in Table 2 indicate that, if one is mostly interested in the skewness of the true X -distribution, the sign of $T_{\beta_0}^{(m)}$ is indeed an effective indicator of the direction of skewness, regardless whether or not (and how) the link function is misspecified. In the absence of X -model misspecification, $T_{\beta_0}^{(m)}$ requires milder error contamination in X in order to more effectively reveal the direction of skewness of the true link.

5 Application to Real Data Examples

We now apply the above testing procedures to two data examples, beginning with a data set from the Framingham Heart Study briefly described in Section 1.

5.1 Framingham Heart Study

The data considered in this example consist of information on 1615 subjects, who were followed for the development of coronary heart disease over six examination periods. Denote by Y_i the binary indicator of the first evidence of coronary heart disease for subject i within an eight-year follow-up period from the second examination period, for $i = 1, \dots, 1615$. At each of the second and third examination periods, each subject's SBP was measured twice. We first center all observed SBP measures from the second examination. Then, for subject i ($= 1, \dots, 1615$), we compute the average of the two (centered) SBP measures divided by 100 from the second examination, and use it as W_i , the error-contaminated version of the unobservable (centered) long-term SBP, X_i . Using the two replicate measures in the second exam and applying equation (4.3) in Carroll et al. (2006) gives an estimated ω for the so-defined W as around 0.92. Assuming a probit-normal structural measurement error model for the observed data $\{(Y_i, W_i)\}_{i=1}^{1615}$, we apply RM with $\lambda = 1$ and $B = 100$. The resulting test statistics are $T_{\beta_0}^{(m)} \approx 2.349$ (0.019) and $T_{\beta_1}^{(m)} \approx -2.387$ (0.017), with the corresponding p -values in parentheses. These test results yield significant evidence that the normality assumption on X is inadequate. This finding is not new (see, e.g., Huang, Stefanski, and Davidian, 2006; Huang, 2009). What is new here is that, because now $T_{\beta_0}^{(m)}$ is significantly positive (at significance level 0.05), using the directional test described in Section 4.4, we also find evidence that the true X -distribution is right-skewed. This new finding (from a model diagnostics standpoint) agrees with the kernel density estimate for X in Wang and Wang (2011, Figure 5), who applied the deconvoluting kernel density estimation (Stefanski and Carroll, 1990) to estimate the density of X based on W -data.

We also apply the RC method using the reclassification model, $P(Y_i^* = Y_i | W_i) = \Phi(W_i)$, for $i = 1, \dots, 1615$, to generate the reclassified data. The resultant test statistics are $T_{\beta_0}^{(c)} \approx -1.474$ (0.141) and $T_{\beta_1}^{(c)} \approx 1.474$ (0.141), with the associated p -values in parentheses. Based on these we conclude that the current data do not give sufficient evidence to imply that the probit link is inappropriate for this application. To this end, we are comfortable with the probit link in the GLM and lean toward a right-skewed distribution for X as opposed to normal.

5.2 Beetle Mortality

Pregibon (1980) studied the association between mortality of adult beetles and exposure to gaseous carbon disulfide. Using his test for link specification, he found strong evidence to support an asymmetric link as opposed to the logit link. The data include logarithm 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 i th beetle, and denote by X_i the standardized (via centering and scaling) logarithm of dosage this beetle was exposed to, for $i = 1, \dots, 481$. Here, the covariate of interest, $\log(\text{dosage})$, is free of measurement error, making assumptions on X -model irrelevant to estimating β . Hence, we first focus on using RC to assess the adequacy of a probit GLM relating Y and X . The reclassification model used for this purpose is $P(Y_i^* = Y_i | X_i) = 0.2$, for $i = 1, \dots, 481$. The values of the test statistics of interest are, with the corresponding p -values in parentheses, $T_{\beta_0}^{(c)} \approx 3.184$ (0.002) and $T_{\beta_1}^{(c)} \approx 0.241$ (0.810). This round of test provides strong evidence that the assumed probit link may be inappropriate.

Log transformation is a popular transformation used by data analysts in biology and medicine to alleviate departure from normality of data. We are now curious about what RM can tell us about the normality assumption on the covariate, $\log(\text{dosage})$, in this study. To make RM applicable, we create noisy surrogate covariate data, $\{W_i\}_{i=1}^{481}$, according to (2) with an estimated ω to be 0.8. Using the new data, $\{(Y_i, W_i)\}_{i=1}^{481}$, pretending to be the "raw" observed data, and assuming a probit-normal model, we implement RM, RC, and the hybrid method, successively. When carrying out RM, the remeasured data, $\{W_{b,i}^*, b = 1, \dots, 100\}_{i=1}^{481}$, are generated according to $W_{b,i}^* = W_i + \sigma_u Z_{b,i}$ with $Z_{b,i} \sim N(0, 1)$, for $b = 1, \dots, 100$, $i = 1, \dots, 481$. For RC and the hybrid method, the reclassified responses are generated according to $P(Y_{b,i}^* = Y_i | W_i) = 0.2$ and $P(Y_{b,i}^* = Y_i | W_{b,i}^*) = 0.2$, respectively, for $b = 1, \dots, 100$, $i = 1, \dots, 481$. The resultant test statistics are listed below, with p -values in parentheses. From RM, $T_{\beta_0}^{(m)} \approx -0.618$ (0.537) and $T_{\beta_1}^{(m)} \approx -1.724$ (0.085); from RC, $T_{\beta_0}^{(c)} \approx 2.904$ (0.004) and $T_{\beta_1}^{(c)} \approx 0.090$ (0.929); and finally, from the hybrid method, $T_{\beta_0} \approx 2.903$ (0.004) and $T_{\beta_1} \approx 0.086$ (0.932). The pattern of these three sets of tests is mostly consistent with what is observed in Section 4.2 when only the link is misspecified. Following the sequential testing strategy proposed in Section 4.3, with the insignificant $T_{\beta_1}^{(m)}$ and the highly significant $T_{\beta_0}^{(c)}$ (at significance level 0.025), one can also conclude that the current data only provide strong evidence against the assumed link but not the normality assumption on $\log(\text{dosage})$. In addition, using the directional test described in Section 4.4, although insignificant, the negative sign of $T_{\beta_0}^{(m)}$ may be an indication that the true link is right-skewed.

For illustration purposes, we drop the log transformation on the dosage levels in the raw data and view the standardized dosage as the true covariate X . Then we repeat the same data generation procedure to create the (hypothetical) error-contaminated observed data, $\{(Y_i, W_i)\}_{i=1}^{481}$, based on which we further generate the

remeasure data and the reclassified data as above, and implement RM, RC, and the hybrid method. The test statistics are: from RM, $T_{\beta_0}^{(m)} \approx -1.192$ (0.234) and $T_{\beta_1}^{(m)} \approx -4.067$ (0.000); from RC, $T_{\beta_0}^{(c)} \approx 1.938$ (0.053) and $T_{\beta_1}^{(c)} \approx -1.320$ (0.188); from the hybrid method, $T_{\beta_0} \approx 1.253$ (0.211) and $T_{\beta_1} \approx -0.843$ (0.400). Now the test based on $T_{\beta_1}^{(m)}$ from RM indicates that the assumed normality on “dosage” is highly suspicious. The nearly significant $T_{\beta_0}^{(c)}$ (at significance level 0.05) from RC may also suggest the probit link questionable, although the evidence is weaker than the previous round of testing from RC when $\log(\text{dosage})$ is the true covariate. This seems to suggest that the power of RC to detect link misspecification is somewhat compromised by the coexistence of an inappropriate assumed X -model. Finally, using the directional test proposed in Section 4.4, the fact that $T_{\beta_0}^{(m)} < 0$, although insignificant, may be evidence that the true distribution of dosage is left-skewed.

6 Discussion

In this study we tackle the challenging problem of model diagnostics for GLM with error-prone covariates, where there are two potential sources of model misspecification. Motivated by the rationale behind the remeasurement method (RM) designed for assessing latent-variable model assumptions, we propose the reclassification method (RC) mainly for detecting a misspecified link in GLM. We carry out rigorous theoretical investigation to study the properties of MLEs for the regression coefficients in GLM when only the link is misspecified, and also when both the assumed link and the assumed latent-variable distribution differ from the truth. These estimators include MLEs resulting from data with measurement error only in the covariate, and also MLEs based on data with measurement error in the binary response. These properties of the estimators justify use of RM and RC for assessing different model assumptions, and further motivate more informative sequential/directional tests that can reveal how the true link or true latent-variable model deviates from the assumed one.

Although starting from Section 3.2 we focus on the (mixture-)probit-normal model as the assumed/true models, the theoretical findings in Sections 3.3 and 3.4 have broader implications beyond this formulation. For example, when the assumed link is logistic and the true link belongs to the class of generalized logistic links (Stukel, 1988), most properties of β_m and β_c stated in Sections 3.3 and 3.4, especially the phenomena demonstrated in Figures 2, 3, and 4, are still observed. The catch then is that none of the needed ingredients listed in Section 3.2 for deriving the scores equations (A.9) and (A.10) are available in closed form. Consequently, one then has to go through far more brute-force numerical procedures to obtain β_m and β_c on a case-by-case basis, making it difficult to summarize their properties, and much more challenging (if not impossible) to analytically establish the clean and neat results such as those in Proposition 3.1 and Appendix D. In other words,

the assumed/true models formulated in Section 3.2 help us make great strides toward understanding the asymptotic properties of the MLEs in the presence of model misspecification, and the findings under this formulation provide answers to more general questions like “What happen to the MLE when one assumes a symmetric (not necessarily normal/probit) X -model/link whereas the true X -model/link is asymmetric?”. Because of the generality of their implications, we expect similar operating characteristics of the proposed testing procedures as described in Section 4.2 outside of the (mixture-)probit-normal formulation.

When multiple model assumptions are in question simultaneously, a potential obstacle for model diagnostics, and for inference in general, is non-identifiability. For example, in the framework of generalized linear mixed models (GLMM), it is only meaningful to test a posited model for the random effects when one assumes that the model for the response given the random effects is correct because these two models cannot be identified/validated simultaneously (Alonso, Litière, and Laenen, 2010; Verbeke and Molenberghs, 2010). In the context of our study, although the true covariate X in the primary model is a latent variable like random effects in GLMM, the existence of an observed surrogate W , which relates to X via a known model, clears the obstacle of non-identifiability encountered in GLMM, and thus it is possible to assess the assumed primary model and the assumed latent-variable model simultaneously. Concrete evidence of such identifiability is partly given by Proposition 3.1.

In the actual implementation of RC, one open question relates to the choice of reclassification model. In this work, we choose this model mostly for ease of deriving the reclassified-data likelihood and also try to avoid too much information loss in the reclassified responses. An interesting follow-up research topic is to find some optimal ways of creating reclassified data to maximize the power of RC. This direction of research will require involvement of the asymptotic variance of the MLE of β , a quantity yet to be studied besides the asymptotic means which we focus on in this article. Other practical concerns worth addressing in the future research are incorporation of multivariate error-prone covariates and relaxing the normality assumption on the measurement error.

Appendix

Appendix A: Likelihood and Score Functions Referenced in Section 3.2

A.1 Likelihood and Score Functions Under the Assumed Model

If one posits a probit link in the primary model and assumes $X \sim N(\mu_x, \sigma_x^2)$, the observed-data likelihood for subject i is

$$f_{Y,W}(Y_i, W_i; \Omega, \sigma_u^2) = e_i [\Phi\{h_i(\beta)\}]^{Y_i} [\Phi\{-h_i(\beta)\}]^{1-Y_i}, \quad \text{for } i = 1, \dots, n, \quad (\text{A.1})$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of $N(0, 1)$, and

$$e_i = \frac{1}{\sqrt{\sigma_u^2 + \sigma_x^2}} \phi\left(\frac{W_i - \mu_x}{\sqrt{\sigma_u^2 + \sigma_x^2}}\right), \quad (\text{A.2})$$

$$h_i(\beta) = \left(\beta_0 + \beta_1 \frac{\sigma_x^2 W_i + \sigma_u^2 \mu_x}{\sigma_u^2 + \sigma_x^2}\right) \left(1 + \frac{\beta_1^2 \sigma_u^2 \sigma_x^2}{\sigma_u^2 + \sigma_x^2}\right)^{-1/2}. \quad (\text{A.3})$$

If the reclassification model is

$$P(Y_i^* = Y_i | W_i) = \pi_i, \quad \text{for } i = 1, \dots, n, \quad (\text{A.4})$$

the likelihood of the i th reclassified data, (Y_i^*, W_i) , under the assumed model is

$$f_{Y^*,W}(Y_i^*, W_i; \Omega, \sigma_u^2) = e_i [\pi_i \Phi\{h_i(\beta)\} + (1 - \pi_i) \Phi\{-h_i(\beta)\}]^{Y_i^*} \times \\ [(1 - \pi_i) \Phi\{h_i(\beta)\} + \pi_i \Phi\{-h_i(\beta)\}]^{1-Y_i^*}. \quad (\text{A.5})$$

Differentiating the logarithm of (A.1) with respect to β yields the normal scores associated with β based on the raw data with measurement error only in X ; and, similarly, differentiating the logarithm of (A.5) with respect to β gives the counterpart normal scores for the reclassified data with measurement error in both X and Y . These two sets of scores are respectively

$$\psi_m(\beta; Y_i, W_i) = h_i'(\beta) \phi\{h_i(\beta)\} \Phi^{-1}\{-h_i(\beta)\} \left[\frac{Y_i}{\Phi\{h_i(\beta)\}} - 1\right], \quad (\text{A.6})$$

$$\psi_c(\beta; Y_i^*, W_i) = h_i'(\beta) \phi\{h_i(\beta)\} \Phi^{-1}\{-h_i(\beta)\} d_i^{-1}(\beta) \times \\ \left[\frac{Y_i^* (2\pi_i - 1) \Phi\{-h_i(\beta)\}}{1 - d_i(\beta)} + 1 - d_i(\beta) - \pi_i\right], \quad (\text{A.7})$$

where

$$d_i(\beta) = (1 - \pi_i) \Phi\{h_i(\beta)\} + \pi_i \Phi\{-h_i(\beta)\}, \quad (\text{A.8})$$

and $h_i'(\beta) = (\partial/\partial\beta)h_i(\beta)$ consists of the following two elements,

$$\frac{\partial h_i(\beta)}{\partial \beta_0} = \left(1 + \frac{\beta_1^2 \sigma_u^2 \sigma_x^2}{\sigma_u^2 + \sigma_x^2}\right)^{-1/2}, \\ \frac{\partial h_i(\beta)}{\partial \beta_1} = \frac{(\sigma_x^2 W_i + \sigma_u^2 \mu_x) \{(\partial/\partial\beta_0)h_i(\beta)\}^{-1} - \beta_1 \sigma_u^2 \sigma_x^2 h_i(\beta)}{\sigma_u^2 + \sigma_x^2 + \beta_1^2 \sigma_u^2 \sigma_x^2}.$$

A close inspection of the scores in (A.6) and (A.7) reveals some values of π_i that one should avoid when specifying the reclassification model in (A.4). First, note that the score function in (A.7) is identically zero if $\pi_i = 0.5$ for all $i = 1, \dots, n$. Consequently, β is non-estimable from the reclassified data generated according to

$P(Y_i^* = Y_i|W_i) = 0.5$ for all $i = 1, \dots, n$. This is not surprising as, with all π_i 's equal to 0.5, $\{Y_i^*\}_{i=1}^n$ virtually contains no information of the true responses. Second, the two sets of scores are equal when $\pi_i = 0$ for $i = 1, \dots, n$, or, $\pi_i = 1$ for $i = 1, \dots, n$. This is also expected as this is the case where $\{Y_i^*\}_{i=1}^n$ literally contains the same information as $\{Y_i\}_{i=1}^n$, and hence MLEs of β from these two data sets are identical, whether or not the assumed model is correct. Therefore, for the purpose of model diagnosis, we avoid setting π_i in (A.4) identically as 0.5, or 0, or 1, for all $i = 1, \dots, n$.

A.2 Score Estimating Equations

Under regularity conditions, the limiting MLE of β based on the raw data and that based on the reclassified data as $n \rightarrow \infty$, β_m and β_c , uniquely satisfy the following score equations respectively,

$$E_{Y,W} \{\psi_m(\beta_m; Y_i, W_i)\} = 0, \quad (\text{A.9})$$

$$E_{Y^*,W} \{\psi_c(\beta_c; Y_i^*, W_i)\} = 0, \quad (\text{A.10})$$

where the subscripts attached to $E\{\cdot\}$ signify that the expectations are defined with respect to the relevant true model.

Using iterated expectations, one can show that (A.9) boils down the following set of two equations,

$$E_W \left[\phi\{h_i(\beta_m)\} \frac{p_i - \Phi\{h_i(\beta_m)\}}{\Phi\{h_i(\beta_m)\}\Phi\{-h_i(\beta_m)\}} \right] = 0, \quad (\text{A.11})$$

$$E_W \left[W_i \phi\{h_i(\beta_m)\} \frac{p_i - \Phi\{h_i(\beta_m)\}}{\Phi\{h_i(\beta_m)\}\Phi\{-h_i(\beta_m)\}} \right] = 0, \quad (\text{A.12})$$

where p_i is the mean of Y_i given W_i under the true model, that is, $p_i = P^{(t)}(Y_i = 1|W_i)$ evaluated at β (the true parameter value), for $i = 1, \dots, n$. Similarly, one can deduce that (A.10) is equivalent to the following system of equations,

$$E_W \left[\phi\{h_i(\beta_c)\} \frac{(1 - 2\pi_i)\{1 - d_i(\beta_c) - q_i\}}{d_i(\beta_c)\{1 - d_i(\beta_c)\}} \right] = 0, \quad (\text{A.13})$$

$$E_W \left[W_i \phi\{h_i(\beta_c)\} \frac{(1 - 2\pi_i)\{1 - d_i(\beta_c) - q_i\}}{d_i(\beta_c)\{1 - d_i(\beta_c)\}} \right] = 0, \quad (\text{A.14})$$

where q_i is the mean of Y_i^* given W_i under the true model, that is,

$$q_i = P^{(t)}(Y_i^* = 1|W_i) = \pi_i p_i + (1 - \pi_i)(1 - p_i), \quad \text{for } i = 1, \dots, n. \quad (\text{A.15})$$

A.3 Likelihood Function Under the True Model

Under the mixture-probit-normal model specified in Section 3.2, the likelihood of (Y_i, W_i) is

$$f_{Y,W}^{(t)}(Y_i, W_i; \Omega^{(t)}, \sigma_u^2) = \rho e_{1i} p_{1i}^{Y_i} (1 - p_{1i})^{1 - Y_i} + (1 - \rho) e_{2i} p_{2i}^{Y_i} (1 - p_{2i})^{1 - Y_i},$$

where, for $\ell = 1, 2$,

$$\begin{aligned} e_{\ell i} &= \frac{1}{\sqrt{\sigma_u^2 + \sigma_{x\ell}^2}} \phi \left(\frac{W_i - \mu_{x\ell}}{\sqrt{\sigma_u^2 + \sigma_{x\ell}^2}} \right), \\ p_{\ell i} &= \alpha \Phi(h_{\ell 1i}) + (1 - \alpha) \Phi(h_{\ell 2i}), \\ h_{\ell ki} &= \left(\beta_0 - \mu_k + \beta_1 \frac{\sigma_{x\ell}^2 W_i + \sigma_u^2 \mu_{x\ell}}{\sigma_u^2 + \sigma_{x\ell}^2} \right) \left(\sigma_k^2 + \frac{\beta_1^2 \sigma_u^2 \sigma_{x\ell}^2}{\sigma_u^2 + \sigma_{x\ell}^2} \right)^{-1/2}, \quad \text{for } k = 1, 2. \end{aligned}$$

It follows that, as the true mean of Y_i given W_i ,

$$p_i = P^{(t)}(Y_i = 1 | W_i) = \frac{\rho e_{1i} p_{1i} + (1 - \rho) e_{2i} p_{2i}}{\rho e_{1i} + (1 - \rho) e_{2i}}, \quad \text{for } i = 1, \dots, n. \quad (\text{A.16})$$

Evaluating (A.15) at this p_i , one obtains the true mean of Y_i^* given W_i , that is, $q_i = P^{(t)}(Y_i^* = 1 | W_i)$, and further deduces that the true-model likelihood of the reclassified data (Y_i^*, W_i) is, for $i = 1, \dots, n$,

$$f_{Y^*,W}^{(t)}(Y_i^*, W_i; \Omega^{(t)}, \sigma_u^2) = \{\rho e_{1i} + (1 - \rho) e_{2i}\} q_i^{Y_i^*} (1 - q_i)^{1 - Y_i^*}.$$

Appendix B: Limiting Maximum Likelihood Estimators when $\beta_1 = 0$

When $\beta_1 = 0$, the limiting MLEs of β are given in the following proposition.

Proposition B.0.1. *Suppose that the true primary model is a GLM with a mixture probit link and $\beta_1 = 0$. Under the assumed probit-normal model, $\beta_c = \beta_m = (\beta_{0m}, 0)^t$, where*

$$\beta_{0m} = \Phi^{-1} \left\{ \alpha \Phi \left(\frac{\beta_0 - \mu_1}{\sigma_1} \right) + (1 - \alpha) \Phi \left(\frac{\beta_0 - \mu_2}{\sigma_2} \right) \right\}. \quad (\text{B.1})$$

The proof is given next, which does not depend on the true X -model or the reclassification model. Proposition B.0.1 indicates that, if $\beta_1 = 0$, β_m does not depend on σ_u^2 , suggesting that RM cannot detect either misspecification. Also, β_c does not depend on π_i , which defeats the purpose of creating reclassified data, hence RC does

not help in model diagnosis either. This implication should not raise much concern because, after all, now $\beta_{1m} = \beta_{1c} = \beta_1 (= 0)$, suggesting that MLEs of β_1 remain consistent despite model misspecification.

Proof: By the uniqueness of the solution to (A.9), it suffices to check if $\beta_m = (\beta_{0m}, 0)^t$ solves (A.11)–(A.12), where β_{0m} is given in (B.1).

Because $\beta_1 = 0$,

$$\begin{aligned}
p_i &= P^{(t)}(Y_i = 1|W_i) \\
&= \frac{f^{(t)}(Y_i = 1, W_i; \Omega^{(t)}, \sigma_u^2)}{f_w^{(t)}(W_i; \tau, \sigma_u^2)} \\
&= \frac{\int P^{(t)}(Y_i = 1|x; \beta) f_{w|x}^{(t)}(W_i|x; \sigma_u^2) f_x^{(t)}(x; \tau) dx}{f_w^{(t)}(W_i; \tau, \sigma_u^2)} \\
&\quad [\text{Note that } P^{(t)}(Y_i = 1|x; \beta) \text{ is free of } x \text{ when } \beta_1 = 0.] \\
&= \frac{P^{(t)}(Y_i = 1|x; \beta) f_w^{(t)}(W_i; \tau, \sigma_u^2)}{f_w^{(t)}(W_i; \tau, \sigma_u^2)} \\
&= \alpha \Phi\left(\frac{\beta_0 - \mu_1}{\sigma_1}\right) + (1 - \alpha) \Phi\left(\frac{\beta_0 - \mu_2}{\sigma_2}\right). \tag{B.2}
\end{aligned}$$

Suppose one assumes for now that $\beta_{1m} = 0$, then by, (A.3), $h_i(\beta_m) = \beta_{0m}$. With both $h_i(\beta_m)$ and p_i in (B.2) free of W_i , (A.11) reduces to $p_i - \Phi\{h_i(\beta_m)\} = 0$, or, $\Phi(\beta_{0m}) = p_i$. Therefore, $\beta_{0m} = \Phi^{-1}(p_i)$, which proves (B.1). And with $p_i - \Phi\{h_i(\beta_m)\} = 0$, (A.12) holds automatically. This completes proving the result regarding β_m .

Next we show that β_m established above also solves (A.13)–(A.14), that is, $\beta_c = \beta_m$. Suppose $\beta_{1c} = 0$, then $h_i(\beta_c) = \beta_{0c}$, and $d_i(\beta_c) = (1 - \pi_i)\Phi(\beta_{0c}) + \pi_i\Phi(-\beta_{0c})$. Note that, inside (A.13), with $q_i = \pi_i p_i + (1 - \pi_i)(1 - p_i)$ and $d_i(\beta_c) = (1 - \pi_i)\Phi(\beta_{0c}) + \pi_i\Phi(-\beta_{0c})$, one has $1 - d_i(\beta_c) - q_i = (1 - 2\pi_i)\{p_i - \Phi(\beta_{0c})\}$. Therefore, if $\beta_{0c} = \Phi^{-1}(p_i)$, then $1 - d_i(\beta_c) - q_i = 0$ and (A.13) holds for all π_i . Furthermore, $1 - d_i(\beta_c) - q_i = 0$ immediately makes (A.14) hold. This shows that $\beta_c = \beta_m$.

This completes the proof for Proposition B.0.1. \square

Appendix C: Proof of Proposition 3.1

The following four results are crucial for proving Proposition 3.1. For clarity, we incorporate the dependence of $h_i(\beta)$ in (A.3) on W_i by re-expressing this function as $h(\beta_0, \beta_1, w)$, with the subscript i suppressed.

- (R1) If $\mu_x = 0$, then $h(-\beta_{0m}, \beta_{1m}, -w) = -h(\beta_{0m}, \beta_{1m}, w)$.
- (R2) If $\mu_x = 0$, then $\phi\{h(-\beta_{0m}, \beta_{1m}, -w)\} = C\phi\{h(\beta_{0m}, \beta_{1m}, w)\}$, where C does not depend on w .

- (R3) If $f_1(x) = f_2(-x)$ and $f_U(u) = f_U(-u)$, then $f_w^{(1)}(w) = f_w^{(2)}(-w)$, where $f_U(u)$ is the pdf of the measurement error U , $f_w^{(1)}(w)$ and $f_w^{(2)}(w)$ are the pdf of W when the pdf of X is $f_1(x)$ and $f_2(x)$, respectively.
- (R4) If $f_1(x) = f_2(-x)$, $f_U(u) = f_U(-u)$, $H_1(s) = 1 - H_2(-s)$, $\mu_x = 0$, and $\beta_0 = 0$, then $p^{(22)}(-w) = 1 - p^{(11)}(w)$, where $p^{(jk)}(w)$ denotes the conditional mean of Y_i given $W_i = w$ under the true model $f_j(x) \wedge H_k(s)$, for $j, k = 1, 2$.

The first two results, (R1) and (R2), follow directly from the definition of $h_i(\beta)$ in (A.3); (R3) can be easily proved by using the convolution formula based on the error model given in equation (2) in the main article. The proof for (R4) is given next.

Proof. By the definition of $p^{(jk)}(w)$, one has, with $\beta_0 = 0$,

$$p^{(11)}(w) = P^{(t)}(Y_i = 1 | W_i = w) = \int_{-\infty}^{\infty} H_1(\beta_1 x) f_U(w - x) f_1(x) dx / f_w^{(1)}(w).$$

Similarly, $p^{(22)}(-w)$ is equal to

$$\begin{aligned} & \int_{-\infty}^{\infty} H_2(\beta_1 x) f_U(-w - x) f_2(x) dx / f_w^{(2)}(-w) \\ &= \int_{-\infty}^{\infty} \{1 - H_1(-\beta_1 x)\} f_U(-w - x) f_1(-x) dx / f_w^{(1)}(w), \text{ by (R3),} \\ &= \int_{-\infty}^{\infty} f_U(-w - x) f_1(-x) dx / f_w^{(1)}(w) - \int_{-\infty}^{\infty} H_1(-\beta_1 x) f_U(-w - x) f_1(-x) dx / f_w^{(1)}(w) \\ &= \int_{-\infty}^{\infty} f_U(-w + s) f_1(s) ds / f_w^{(1)}(w) - \int_{-\infty}^{\infty} H_1(\beta_1 s) f_U(-w + s) f_1(s) ds / f_w^{(1)}(w) \\ &= \int_{-\infty}^{\infty} f_U(w - s) f_1(s) ds / f_w^{(1)}(w) - \int_{-\infty}^{\infty} H_1(\beta_1 s) f_U(w - s) f_1(s) ds / f_w^{(1)}(w) \\ &= 1 - p^{(11)}(w). \end{aligned}$$

This completes the proof of (R4).

Now we are ready to show Proposition 3.1. In essence, we will show that, if (β_{0m}, β_{1m}) solves (A.11)–(A.12) when the true model is $f_1(x) \wedge H_1(s)$, then $(-\beta_{0m}, \beta_{1m})$ solves (A.11)–(A.12) when the true model is $f_2(x) \wedge H_2(s)$. More specifically, evaluating (A.11) and (A.12) at its solution under the true model $f_1(x) \wedge H_1(s)$, we will show that the following two equations,

$$\int_{-\infty}^{\infty} \phi\{h(\beta_{0m}, \beta_{1m}, w)\} \frac{p^{(11)}(w) - \Phi\{h(\beta_{0m}, \beta_{1m}, w)\}}{\Phi\{h(\beta_{0m}, \beta_{1m}, w)\} \Phi\{-h(\beta_{0m}, \beta_{1m}, w)\}} f_w^{(1)}(w) dw = 0, \quad (\text{C.1})$$

$$\int_{-\infty}^{\infty} w \phi\{h(\beta_{0m}, \beta_{1m}, w)\} \frac{p^{(11)}(w) - \Phi\{h(\beta_{0m}, \beta_{1m}, w)\}}{\Phi\{h(\beta_{0m}, \beta_{1m}, w)\} \Phi\{-h(\beta_{0m}, \beta_{1m}, w)\}} f_w^{(1)}(w) dw = 0, \quad (\text{C.2})$$

imply the following two identities,

$$\int_{-\infty}^{\infty} \phi\{h(-\beta_{0m}, \beta_{1m}, w)\} \frac{p^{(22)}(w) - \Phi\{h(-\beta_{0m}, \beta_{1m}, w)\}}{\Phi\{h(-\beta_{0m}, \beta_{1m}, w)\} \Phi\{-h(-\beta_{0m}, \beta_{1m}, w)\}} f_w^{(2)}(w) dw = 0, \quad (\text{C.3})$$

$$\int_{-\infty}^{\infty} w \phi\{h(-\beta_{0m}, \beta_{1m}, w)\} \frac{p^{(22)}(w) - \Phi\{h(-\beta_{0m}, \beta_{1m}, w)\}}{\Phi\{h(-\beta_{0m}, \beta_{1m}, w)\} \Phi\{-h(-\beta_{0m}, \beta_{1m}, w)\}} f_w^{(2)}(w) dw = 0. \quad (\text{C.4})$$

Take (C.4) as an example, the left-hand side of it is equal to, by (R1)–(R4) and $\Phi(-t) = 1 - \Phi(t)$,

$$\begin{aligned} & \int_{-\infty}^{\infty} (-v) \phi\{h(-\beta_{0m}, \beta_{1m}, -v)\} \frac{p^{(22)}(-v) - \Phi\{h(-\beta_{0m}, \beta_{1m}, -v)\}}{\Phi\{h(-\beta_{0m}, \beta_{1m}, -v)\} \Phi\{-h(-\beta_{0m}, \beta_{1m}, -v)\}} f_w^{(2)}(-v) dv \\ &= -C \int_{-\infty}^{\infty} v \phi\{h(\beta_{0m}, \beta_{1m}, v)\} \frac{1 - p^{(11)}(v) - \Phi\{-h(\beta_{0m}, \beta_{1m}, v)\}}{\Phi\{-h(\beta_{0m}, \beta_{1m}, v)\} \Phi\{h(\beta_{0m}, \beta_{1m}, v)\}} f_w^{(1)}(v) dv \\ &= -C \int_{-\infty}^{\infty} v \phi\{h(\beta_{0m}, \beta_{1m}, v)\} \frac{1 - p^{(11)}(v) - 1 + \Phi\{h(\beta_{0m}, \beta_{1m}, v)\}}{\Phi\{h(\beta_{0m}, \beta_{1m}, v)\} \Phi\{-h(\beta_{0m}, \beta_{1m}, v)\}} f_w^{(1)}(v) dv \\ &= C \int_{-\infty}^{\infty} v \phi\{h(\beta_{0m}, \beta_{1m}, v)\} \frac{p^{(11)}(v) - \Phi\{h(\beta_{0m}, \beta_{1m}, v)\}}{\Phi\{h(\beta_{0m}, \beta_{1m}, v)\} \Phi\{-h(\beta_{0m}, \beta_{1m}, v)\}} f_w^{(1)}(v) dv \\ &= 0, \text{ according to (C.2).} \end{aligned}$$

Following similar derivations, one can show that the left-hand side of (C.3) is equal to

$$-C \int_{-\infty}^{\infty} \phi\{h(\beta_{0m}, \beta_{1m}, v)\} \frac{p^{(11)}(v) - \Phi\{h(\beta_{0m}, \beta_{1m}, v)\}}{\Phi\{h(\beta_{0m}, \beta_{1m}, v)\} \Phi\{-h(\beta_{0m}, \beta_{1m}, v)\}} f_w^{(1)}(v) dv,$$

which is also equal to 0 according to (C.1). Therefore, $\beta_{0m}^{(11)} = -\beta_{0m}^{(22)}$ and $\beta_{1m}^{(11)} = \beta_{1m}^{(22)}$. This completes the proof of Proposition 3.1.

Appendix D: A Counterpart Proposition of Proposition 3.1 for β_c

Proposition D.0.2. *Let $f_1(x)$ and $f_2(x)$ be two pdf's specifying two true X -distributions that are symmetric of each other, and let $H_1(s)$ and $H_2(s)$ be two true links that are symmetric of each other. Denote by $\beta_c^{(jk)}$ the limiting MLE of β based on reclassified data generated according to $P(Y_i^* = Y_i | W_i) = \pi(W_i)$ when the true model is $f_j(x) \wedge H_k(s)$, for $j, k = 1, 2$. If $\mu_x = \beta_0 = 0$ and $\pi(t)$ is an even function, then $\beta_{0c}^{(11)} = -\beta_{0c}^{(22)}$ and $\beta_{1c}^{(11)} = \beta_{1c}^{(22)}$.*

The following two lemmas are needed in the proof, one lemma concerning $d_i(\beta)$ defined in (A.8), and the other relates to q_i defined in (A.15). To elaborate the de-

pendence of $d_i(\beta)$ on W_i in (A.8), we re-express this function as $d(\beta_0, \beta_1, w)$, with the subscript i suppressed.

Lemma 1. *If $\mu_x = 0$ and $\pi(t)$ is an even function, then $d(-\beta_{0c}, \beta_{1c}, -w) = 1 - d(\beta_{0c}, \beta_{1c}, w)$.*

Proof: By (A.8),

$$\begin{aligned} & d(-\beta_{0c}, \beta_{1c}, -w) \\ &= \{1 - \pi(-w)\} \Phi\{h(-\beta_{0c}, \beta_{1c}, -w)\} + \pi(-w) \Phi\{-h(-\beta_{0c}, \beta_{1c}, -w)\} \\ &= \{1 - \pi(w)\} \Phi\{-h(\beta_{0c}, \beta_{1c}, w)\} + \pi(w) \Phi\{h(\beta_{0c}, \beta_{1c}, w)\} \\ & \quad [\text{Next use (R1) and the fact that } \pi(t) = \pi(-t).] \\ &= \{1 - \pi(w)\} [1 - \Phi\{h(\beta_{0c}, \beta_{1c}, w)\}] + \pi(w) [1 - \Phi\{-h(\beta_{0c}, \beta_{1c}, w)\}] \\ &= 1 - d(\beta_{0c}, \beta_{1c}, w). \end{aligned}$$

This completes the proof of Lemma 1. \square

Lemma 2. *If $f_1(x) = f_2(-x)$, $f_v(u) = f_v(-u)$, $H_1(s) = 1 - H_2(-s)$, $\mu_x = 0$, $\beta_0 = 0$, and $\pi(t)$ is an even function, then $q^{(22)}(-w) = 1 - q^{(11)}(w)$, where $q^{(jk)}(w)$ denotes the conditional mean of Y_i^* given $W_i = w$ under the true model $f_j(x) \wedge H_k(s)$, for $j, k = 1, 2$.*

Proof: By (A.15),

$$\begin{aligned} & q^{(22)}(-w) \\ &= \{1 - \pi(-w)\} \{1 - p^{(22)}(-w)\} + \pi(-w) p^{(22)}(-w) \\ &= \{1 - \pi(w)\} p^{(11)}(w) + \pi(w) \{1 - p^{(11)}(w)\}, \text{ by (R4) and } \pi(-t) = \pi(t), \\ &= 1 - q^{(11)}(w). \end{aligned}$$

This completes the proof of Lemma 2. Following similar derivations, one can show that $q^{(12)}(-w) = 1 - q^{(21)}(w)$.

Now we are ready to show that, if (β_{0c}, β_{1c}) solves (A.13)–(A.14) under the true model $f_1(x) \wedge H_1(s)$, then $(-\beta_{0c}, \beta_{1c})$ solves (A.13)–(A.14) under the true model $f_2(x) \wedge H_2(s)$. Given that (β_{0c}, β_{1c}) solves (A.13) and (A.14) under the true model $f_1(x) \wedge H_1(s)$, one has, by elaborating (A.13) and (A.14),

$$\begin{aligned} & \int_{-\infty}^{\infty} \frac{\phi\{h(\beta_{0c}, \beta_{1c}, w)\}}{d(\beta_{0c}, \beta_{1c}, w) \{1 - d(\beta_{0c}, \beta_{1c}, w)\}} \{1 - 2\pi(w)\} \\ & \left\{1 - q^{(11)}(w) - d(\beta_{0c}, \beta_{1c}, w)\right\} f_w^{(1)}(w) dw = 0, \end{aligned} \quad (\text{D.1})$$

$$\begin{aligned} & \int_{-\infty}^{\infty} w \frac{\phi\{h(\beta_{0c}, \beta_{1c}, w)\}}{d(\beta_{0c}, \beta_{1c}, w) \{1 - d(\beta_{0c}, \beta_{1c}, w)\}} \{1 - 2\pi(w)\} \\ & \left\{1 - q^{(11)}(w) - d(\beta_{0c}, \beta_{1c}, w)\right\} f_w^{(1)}(w) dw = 0. \end{aligned} \quad (\text{D.2})$$

Now we check if $(-\beta_{0c}, \beta_{1c})$ solves (A.13)–(A.14) under the true model $f_2(x) \wedge H_2(s)$. Plugging $(-\beta_{0c}, \beta_{1c})$ in (A.13) gives, where we set $v = -w$ in the first equality,

$$\begin{aligned}
& \int_{-\infty}^{\infty} \frac{\phi \{h(-\beta_{0c}, \beta_{1c}, w)\}}{d(-\beta_{0c}, \beta_{1c}, w) \{1 - d(-\beta_{0c}, \beta_{1c}, w)\}} \{1 - 2\pi(w)\} \\
& \quad \{1 - q^{(22)}(w) - d(-\beta_{0c}, \beta_{1c}, w)\} f_w^{(2)}(w) dw \\
= & \int_{-\infty}^{\infty} \frac{\phi \{h(-\beta_{0c}, \beta_{1c}, -v)\}}{d(-\beta_{0c}, \beta_{1c}, -v) \{1 - d(-\beta_{0c}, \beta_{1c}, -v)\}} \{1 - 2\pi(-v)\} \\
& \quad \{1 - q^{(22)}(-v) - d(-\beta_{0c}, \beta_{1c}, -v)\} f_w^{(2)}(-v) dv \\
& \quad [\text{Next use (R1)–(R3), Lemma 1, Lemma 2, and } \pi(t) = \pi(-t).] \\
= & \int_{-\infty}^{\infty} \frac{C \phi \{h(\beta_{0c}, \beta_{1c}, v)\}}{\{1 - d(\beta_{0c}, \beta_{1c}, v)\} d(\beta_{0c}, \beta_{1c}, v)} \{1 - 2\pi(v)\} \\
& \quad \{-1 + q^{(11)}(v) + d(\beta_{0c}, \beta_{1c}, v)\} f_w^{(1)}(v) dv \\
= & -C \int_{-\infty}^{\infty} \frac{\phi \{h(\beta_{0c}, \beta_{1c}, v)\}}{d(\beta_{0c}, \beta_{1c}, v) \{1 - d(\beta_{0c}, \beta_{1c}, v)\}} \{1 - 2\pi(v)\} \\
& \quad \{1 - q^{(11)}(v) - d(\beta_{0c}, \beta_{1c}, v)\} f_w^{(1)}(v) dv \\
= & 0, \text{ by (D.1).}
\end{aligned}$$

Similarly, one can show that (D.2) implies

$$\begin{aligned}
& \int_{-\infty}^{\infty} w \frac{\phi \{h(-\beta_{0c}, \beta_{1c}, w)\}}{d(-\beta_{0c}, \beta_{1c}, w) \{1 - d(-\beta_{0c}, \beta_{1c}, w)\}} \{1 - 2\pi(w)\} \\
& \quad \{1 - q^{(22)}(w) - d(-\beta_{0c}, \beta_{1c}, w)\} f_w^{(2)}(w) dw = 0.
\end{aligned}$$

Hence, $(-\beta_{0c}, \beta_{1c})$ does solve (A.13)–(A.14) under the true model $f_2(x) \wedge H_2(s)$. In other words, $\beta_{0c}^{(11)} = -\beta_{0c}^{(22)}$ and $\beta_{1c}^{(11)} = \beta_{1c}^{(22)}$. Following parallel arguments as above one can show that $\beta_{0c}^{(12)} = -\beta_{0c}^{(21)}$ and $\beta_{1c}^{(12)} = \beta_{1c}^{(21)}$. This completes the proof of Proposition D.0.2. \square

References

- Alonso, A., Litière, S., and Laenen, A. (2010). A note on the indeterminacy of the random-effects distribution in hierarchical models. *The American Statistician* **64**, 318–324.
- Brown, C. C. (1982). On a goodness-of-fit test for the logistic model based on score statistics. *Communications in Statistics* **11**, 1087–1105.

- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Non-linear Models: A Modern Perspective*, second edition. Chapman & Hall/CRC, Boca Raton.
- Czado, C. and Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference* **33**, 213-31.
- Fowlkes, E. B. (1987). Some diagnostics for binary regression via smoothing. *Biometrika* **74**, 503–515.
- Fuller, W. A. (1987). *Measurement error models*. John Wiley and Sons, New York.
- Hosmer, D. W., Hosmer, T., le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* **16**, 965–980.
- Hosmer, D. W., Lemeshow, S. (1989). *Applied logistic regression*. Wiley, New York.
- Huang, X. (2009). An improved test of latent-variable model misspecification in structural measurement error models for group testing data. *Statistics in Medicine* **28**, 3316–3327.
- Huang, X., Stefanski, L. A., and Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika* **93**, 53–64.
- Huang, X., Stefanski, L. A., and Davidian, M. (2009). Latent-model robustness in joint modeling for a primary endpoint and a longitudinal process. *Biometrics* **65**, 719–727.
- Kannel, W. B., Neaton, J. D., Wentworth, D., Thomas, H. E., Stamler, J., Hulley, S. B., and Kjelsberg, M. O. (1986). Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for MRFIT. *American Heart Journal* **112**, 825–36.
- le Cessie, S. and van Houwelingen, J. C. (1991). A goodness-of-fit test for binary data based on smoothing residuals. *Biometrics* **47**, 1267–1282.
- Li, K. and Duan N. (1989). Regression Analysis Under Link Violation. *The Annals of Statistics* **17**, 1009–1052.
- Ma, Y., Hart, J. D., Janicki, R., and Carroll, R. J. (2011). Local and omnibus goodness-of-fit tests in classical measurement error models. *Journal of the Royal Statistical Society: Series B* **73**, 81–98.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. Chapman & Hall/CRC, .
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A* **135**, 370–384.
- Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society: Series C* **29**, 15–24.
- Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21**, 169–184.
- Stukel, T. A. (1988). Generalized logistic models. *Journal of American Statistical Association* **83**, 426–431.
- Tsiatis, A. A. (1980). A note on a goodness-of-fit test for logistic regression model. *Biometrika* **67**, 250–251.

- Verbeke, G. and Molenberghs, G. (2010). Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statistical Modelling* **10**, 391–419.
- Wang, X. and Wang, B. (2011). Deconvolution estimation in measurement error models: the R package decon. *Journal of Statistical Software* **39**, 1–24.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*. **50** 1–25.