Research Article

Statistics in Medicine

Received 31 March 2008, Accepted 12 April 2010 Published online 26 May 2010 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3966

A note on dichotomization of continuous response variable in the presence of contamination and model misspecification

Yue Shentu[‡] and Minge Xie[*†]

The purpose of this note is to raise awareness of the complexity of the practice involving dichotomization. It is well known that the regular regression models are effective tools for analyzing Gaussian-type response variables, and researchers are often told that it is a 'bad idea' to practice dichotomization if continuous measurements are available. We demonstrate through special cases, however, that there is another side of the story if the response variable is contaminated. Although dichotomization causes loss of information, it can also reduce input of contamination. If the reduction of contamination input outweighs the loss of information, analysis based on dichotomization can sometimes provide better results. We derive formulas of bias and variance for binary regression estimators under a contamination model of unknown additive errors, and compare them with both the least squares and robust M-estimators from the corresponding linear regression analysis using continuous responses. As a case study, we study extensively the case in which the observed response is contaminated by an error with a mean and a variance proportional to the mean and the variance of the uncontaminated true response. Conditions under which dichotomization is preferred are obtained. A simulation study based on a real data setting is provided, which supports the theoretical developments.

Keywords: linear regression model; logistic regression; errors in responses; parameter estimation; quantal regression; robust method

1. Introduction

Dichotomization of continuous variables is 'widespread in clinical studies and other fields' [1]. For example, in high-throughput screening of chemical compounds at early stage of drug discovery, measurements of chemical potencies are routinely dichotomized in data analysis [2]. In medical research, certain clinical measurements, such as blood pressure or Hemoglobin A1c level, usually have conventional thresholds routinely used by physicians to make diagnosis [3]. In studies in social sciences or developmental psychology, it is 'common to use categorized outcomes' [4] and the categorized binary response variables are 'frequently modeled using binary outcome regression models such as ordinary logistic and ordinary probit regression' [4, 5]. It is well known theoretically that in ideal situations, when dichotomization takes place, some information contained in the original data is lost. The loss of information leads to loss of efficiency in estimation and loss of power in hypothesis testing; see, e.g. [1, 4, 6, 7]. A common defense of dichotomization involves, in one form or another, an argument that 'analysis and results are simplified' [7]. Or, there is also 'a general need in clinical practice to label individuals as having or not having an attribute (such as hypertensive, obese, high PSA) to determining diagnostic or therapeutic procedures' [1]. Justifications for dichotomization and choices of thresholds include 'following practices used in previous research' or 'using clinically significant cutpoints', among others [7]. In this note, we focus on a relevant question of practical interest: Does dichotomization or categorization always lead to worse statistical inference under a more complex situation? Using linear regression models, we demonstrate that when responses

Department of Statistics and Biostatistics, Hill Center for Mathematical Sciences, Rutgers University, Piscataway, NJ 08854, U.S.A.
[*]Correspondence to: Minge Xie, Department of Statistics and Biostatistics, Hill Center for Mathematical Sciences, Rutgers University, Piscataway, NJ 08854, U.S.A.
[†]E-mail: mxie@stat.rutgers.edu
[‡]Current address: Merck Research Laboratories, 126 E. Lincoln Avenue, P.O. Box 2000, Rahway, NJ 07065, U.S.A.

Contract/grant sponsor: NSF; contract/grant numbers: DMS0915139, SES0851521
Contract/grant sponsor: NSA; contract/grant number: H98230-08-1-0104
Contract/grant sponsor: DHS; contract/grant number: 2008-DN-077-ARI012-03

are contaminated by additive errors, binary regression based on dichotomized responses can sometimes produce better regression estimators than the corresponding regular linear regression and robust M-estimation approaches. The main goal of this research is to raise awareness on the complexity of the practice involving dichotomization.

Consider a standard linear regression model

$$Y_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i = \mu_i + \varepsilon_i, \quad i = 1 \ldots n, \tag{1}$$

where the error terms $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ are independently identically distributed random errors with a common variance σ^2. Let $Z_i = \mathbf{1}(Y_i > d)$ be the dichotomized response by a known threshold d, where $\mathbf{1}(\cdot)$ is the 0–1 indicator function. A binary probit model for the dichotomized response Z_i is

$$P(Z_i = 1|\mathbf{X}_i) = \Phi(\{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - d\}/\sigma), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. In quantal analysis models [8], the threshold d is often assumed to be a known constant and it can be absorbed into the intercept in the binary regression model (2). Without loss of generality, we assume throughout the paper that $d = 0$; discussions on different choices of thresholds can be found in Section 5. Also, the regression coefficients in model (2) are identifiable up to only $\boldsymbol{\beta}/\sigma$, and it is often assumed that $\sigma^2 = 1$ in quantal analysis models [8]. This assumption only affects the scale of the regression parameters, but not hypothesis testing results on the regression parameters (for example, if we perform the commonly used student t or likelihood ratio tests).

For $a = \frac{1}{1.702}$ and any t, we have $|\mathrm{e}^t/(1+\mathrm{e}^t) - \Phi(t\,a)| < 0.00946$; See, e.g. Demidenko [9, p. 336]. In data analysis, the probit regression model (2) with $\sigma = 1$ and $d = 0$ is little different from the most commonly used binary regression model

$$P(Z_i = 1|\mathbf{X}_i) = H(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}/a), \tag{3}$$

where $H(t) = \mathrm{e}^t/(1+\mathrm{e}^t)$ is the inverse logistic link function.

Assume that the response data are generated according to model (1). The least squares or the maximum likelihood estimators of the regression parameters $\boldsymbol{\beta}$ in model (1), say $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$, is the best linear unbiased estimator (BLUE). Under some regularity conditions [10], the maximum likelihood estimators in model (2) or (3), say $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$, is also consistent. See, Chapter 3 of Santner and Duffy [11] and Section 7.2 of Venables and Ripley [12] for discussions on the performance of the maximum likelihood estimators in binary regressions. The variance of $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ is typically smaller than the variance of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$. This result is one of the key pieces of theoretical evidence that dichotomization leads to loss of information, resulting in loss of efficiency in estimation and loss of power in statistical testing.

In this note, we focus on a situation when there are systematic noises in the measurements of the continuous response variables Y_i. Noisy data in continuous scale are very common in real-life data analysis. For example in micro-array studies, it is well known that intensity measurements for genes' expression levels are often very noisy, and in many cases the noise level is proportional to the true expression intensities. In survey sampling data in social and behavioral sciences, the real responses to survey questions may sometimes be misrepresented, censored or rounded up due to various reasons such as privacy concerns, misunderstanding of the designed questions, among others [13]. Also, in the aforementioned example of high-throughput screening practice of drug discovery, there are tremendous errors in the 'cheap, quick and dirty' measurements of chemical potencies [2]. Similar (may be less extensive) uncertainty is also in the measures of blood pressures or blood glucose levels in clinical studies [3]. Although dichotomization or categorization causes loss of information, it may also reduce the input of the random contamination. If the reduction in the input of the random contamination outweighs the loss of information due to dichotomization, the analysis based on dichotomization can provide better results.

Let us assume that the true response variable of interest Y_i follows model (1) with $\sigma = 1$, but it is confounded with an additive random contamination, say e_i. What is actually observed is a contaminated version of the response

$$Y_i^* = Y_i + e_i, \quad e_i \sim \mathrm{N}(\eta_i, \tau_i^2), \tag{4}$$

where we assume that e_i is independent of the regression error ε_i. In this case, the actual model for the observed response Y_i^* is

$$Y_i^* = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i + e_i, \quad i = 1 \ldots n. \tag{5}$$

Since we usually do not know about the actual contamination e_i, a working model is often used,

$$Y_i^* = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i^*, \quad i = 1 \ldots n, \tag{6}$$

where the error term ε_i^* is falsely assumed to be independently and identically distributed with mean 0 and a homoscedastic variance. Alternatively, if we perform dichotomization $Z_i^* = \mathbf{1}(Y_i^* > 0)$, a binary working model is usually

$$P(Z_i^* = 1|\mathbf{X}_i) = H\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}/a\right), \tag{7}$$

where the conditional mean $E(Z_i^*|\mathbf{X}_i)$ is falsely assumed to be $H\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}/a\right)$. Clearly, these two working models are misspecified models, especially when the mean of the random contamination $\eta_i = E(e_i) \neq 0$. This includes the case that some significant covariate variables are completely missing or omitted from the regression models, among others. We compare the regression parameter estimators in these two working models. Our objective is to investigate the impact of the random contamination and the model misspecification, and assess whether estimation based on model (7) could be more robust against the random contamination in certain situations.

The contamination model assumption (4) is similar in form to the 'error-in-variable-model' assumption (often on a covariate variable) in the measurement error literature; see, e.g. model (1.1) of [14]. However, we do not restrict the mean of the contamination error $\eta_i = Ee_i$ to be zero. Thus, the 'seemingly benign case' [15] with $\eta_i = Ee_i = 0$, an assumption that has been used in many publications in the measurement error literature, is a special case of ours. In practice, there are situations that $\eta_i \neq 0$, including those measurement error models described by Buonaccorsi [16] in radioimmunoassay practice. It is true that, when the error bias term $\eta_i \neq 0$, it is often difficult to estimate η_i without further knowledge or assumption on the contamination error e_i. But this is not a big concern in our development, since our task is to study the impact of dichotomization in the presence of unknown measuring errors and to raise awareness about the complexity of the practice involving dichotomization. This goal is different from that of establishing consistent estimation approaches in the usual measurement error literature. Nevertheless, our development later suggests that the impact of the contamination could be very large when $\eta_i \neq 0$, but it is much less under the common measurement error assumption that $\eta_i = 0$.

The contamination model (4) is also different from the standard gross contamination assumption often used in the conventional robust approaches (e.g. [17, 18]). The standard gross contamination model assumes only a few responses are impacted by extreme outliers [17, 18]. One often used gross contamination model is

$$Y_i^* = (1-\delta)Y_i + \delta e_i \quad \text{for a small constant } \delta, 0 < \delta < \tfrac{1}{2}, \tag{8}$$

where e_i is from a distribution that is quite different from the distribution of Y_i. Although the traditional robust procedures, such as M-estimation and others, are effective in dealing with contaminations of outlying points in (8), they are not particularly designed for contaminations modeled in (4). In this note, we include an M-estimation approach for the linear regression model (6). Our theoretical developments and numerical results suggest that conventional robust approaches, such as the M-estimation, are ineffective for the type of contaminations modeled in (4).

It is well known that we can construct examples in which the least-squares estimator or the robust estimator is the best, in terms of mean square errors. For instances, in the case when there is no or little contamination in the measurements, the least-square estimator is the best among unbiased linear estimators. In the case of gross contaminations modeled in (8), the traditional robust estimation methods, such as M-estimations, are better suited. However, little attention has been received in literature that dichotomization could sometimes outperform the least squares and even the robust estimators. This note provides such an example in the presence of unknown additive contamination errors. The goal is neither to promote nor to dismiss the dichotomization or any other procedures. It is to serve as a reminder of the complexity of the practice involving dichotomization.

We arrange the remainder of the paper as follows. In Section 2, we study estimations of regression parameters under the two working models (6) and (7). Formulas of bias and variance for the parameter estimators are given as the theoretical basis of further investigation. Section 3 considers two special cases in which the least squares and the M-estimators from model (6) and the maximum likelihood estimator from model (7) have close forms, asymptotically. They include the case of no errors and the case in which the mean and the variance of the contamination error e_i are proportional to the mean and the variance of the true response y_i. Section 4 contains simulation studies from a real data setting to illustrate what might happen in actual data analysis. Section 5 provides further discussion and remarks.

2. Regression parameter estimation with contaminated responses

In this section, we study the least-squares estimator under the working linear regression model (6), the maximum likelihood estimator under the working binary logistic model (7) and an M-estimator under the working linear regression model (6).

Statistics in Medicine

2.1. Least-squares estimation under working linear regression model

The least-squares estimator from the working model (6) can be expressed explicitly. It is just the standard least-squares estimator with the true responses replaced by the contaminated responses

$$\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbf{Y}^{*}, \tag{9}$$

where $\boldsymbol{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)^{\mathrm{T}}$ is the design matrix and $\mathbf{Y}^{*}=(Y_1^{*},\ldots,Y_n^{*})$ is the observed (contaminated) response vector. From (4), the actual distribution of $\varepsilon_i^{*}=\varepsilon_i+e_i$ is $\varepsilon_i^{*}\sim\mathrm{N}(\eta_i,1+\tau_i^2)$. Thus, the bias and the variance for the resulting estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ are

$$E(\hat{\boldsymbol{\beta}}^{(\mathrm{LS})})-\boldsymbol{\beta}_0=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\eta}\quad\text{and}\quad\mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{LS})})=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{\mathrm{LS}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}, \tag{10}$$

where $\boldsymbol{\beta}_0$ is the true parameter value, $\boldsymbol{\eta}=(\eta_1,\eta_2,\ldots,\eta_n)^{\mathrm{T}}$ and $\boldsymbol{W}_{\mathrm{LS}}=\mathrm{diag}(1+\tau_1^2,1+\tau_2^2,\ldots,1+\tau_n^2)$ is an $n\times n$ diagonal matrix with its ith diagonal element equal to $1+\tau_i^2$.

2.2. Maximum likelihood estimation under working logistic model

Assume that the maximum likelihood estimator of $\boldsymbol{\beta}$ exists in model (3) using uncontaminated data, and it is consistent and asymptotically normally distributed. We want to know the behavior of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$, the estimator of fitting the working binary logistic model (7) using contaminated data.

The set of estimating equations for $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ is obtained by setting the working score function to zero

$$S_D(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\{Z_i^{*}-H(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}/a)\}\mathbf{X}_i=0. \tag{11}$$

There is no close-form solution for $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$, and the iterative re-weighted least-squares estimation [19] is used to solve (11). The true relation between dichotomized Z_i^{*} and the covariates $\mathbf{X}_i$ is

$$E(Z_i^{*}|\mathbf{X}_i)=P(Z_i^{*}=1|\mathbf{X}_i)=P(Y_i^{*}>0|\mathbf{X}_i)=\Phi\left(\frac{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}+\eta_i}{\sqrt{1+\tau_i^2}}\right). \tag{12}$$

Hence, the set of estimating equations (11) is not Fisher consistent.

Suppose there exists a solution, say $\boldsymbol{\beta}^{*}$, to the following equations:

$$E\{S_D(\boldsymbol{\beta})\}=\frac{1}{n}\sum_{i=1}^{n}\left\{\Phi\left(\frac{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_i}{\sqrt{1+\tau_i^2}}\right)-H(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}/a)\right\}\mathbf{X}_i=0. \tag{13}$$

The solution $\boldsymbol{\beta}^{*}$ depends only on the covariates $\mathbf{X}_i$'s, the true $\boldsymbol{\beta}_0$ and the η_i and τ_i values, but not the random responses Y_i^{*}'s, Z_i^{*}'s or Y_i's. In general $\boldsymbol{\beta}^{*}$ and $\boldsymbol{\beta}_0$ are different, and the difference depends on the distances between the true expectations of Z_i^{*}, i.e. $E(Z_i^{*}|\mathbf{X}_i)$, and the falsely assumed expectations of Z_i^{*}, i.e. $H(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0/a)$, in the working model (7):

$$\begin{aligned}\bar{\boldsymbol{\delta}}_D=\boldsymbol{\delta}_D(\boldsymbol{\beta}_0)&=(E(Z_1^{*}|\mathbf{X}_1)-H(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_0/a),\ldots,E(Z_n^{*}|\mathbf{X}_n)-H(\mathbf{X}_n^{\mathrm{T}}\boldsymbol{\beta}_0/a))^{\mathrm{T}}\\&=\left(\Phi\left(\frac{\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_1}{\sqrt{1+\tau_1^2}}\right)-H(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_0/a),\ldots,\Phi\left(\frac{\mathbf{X}_n^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_n}{\sqrt{1+\tau_n^2}}\right)-H(\mathbf{X}_n^{\mathrm{T}}\boldsymbol{\beta}_0/a)\right)^{\mathrm{T}}.\end{aligned}$$

The first result of the following lemma provides a formula to describe the difference between $\boldsymbol{\beta}^{*}$ and $\boldsymbol{\beta}_0$ in terms of $\bar{\boldsymbol{\delta}}_D$. The second result of the lemma suggests that $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ and $\boldsymbol{\beta}^{*}$ are asymptotically equivalent. Although the condition can be much weaker, to simplify technical details we assume that the covariates $\mathbf{X}_i$'s are from compact sets and the means and the variances of the contamination errors e_i are bounded, in addition to the regularity conditions that the likelihood estimator exists.

Lemma 1
Under the conditions described above, we have

(i) $$\boldsymbol{\beta}^* - \boldsymbol{\beta}_0 = a(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_D^*\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\bar{\boldsymbol{\delta}}_D,$$

where $\boldsymbol{V}_D^* = \mathrm{diag}[H(\xi_1)\{1-H(\xi_1)\}, \ldots, H(\xi_n)\{1-H(\xi_n)\}]$, for some ξ_i between $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*/a$ and $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0/a$.

(ii) $$\hat{\boldsymbol{\beta}}^{(\mathrm{D})} - \boldsymbol{\beta}^* = O_p(n^{-1/2}).$$

A proof of this lemma is in the Appendix.

The following theorem provides a set of asymptotic expressions for the mean and the variance of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$. In particular, in the large sample setting, the mean of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ is equal to $\boldsymbol{\beta}^*$ and the variance of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ is in a 'sandwich' form with the true variance of Z_i^* sandwiched between the two copies of the assumed variance from the working model.

Theorem 1
Under the conditions of Lemma 1, the asymptotic mean and the variance of the estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ are

$$E\hat{\boldsymbol{\beta}}^{(\mathrm{D})} = \boldsymbol{\beta}^* + o(n^{-1/2}),$$
$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{D})}) = a^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_D\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}W_D\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_D\boldsymbol{X})^{-1} + o(n^{-1}), \tag{14}$$

where W_D and $\boldsymbol{V}_D$ are, respectively, the true variance of $\mathbf{Z}^* = (Z_1^*, \ldots, Z_n^*)^{\mathrm{T}}$ and the assumed variance under the working logistic model (7):

$$W_D = \mathrm{Var}(\mathbf{Z}^*) = \mathrm{diag}\{\mathrm{Var}(z_1), \ldots, \mathrm{Var}(z_n)\}$$
$$= \mathrm{diag}\left[\Phi\left(\frac{\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_1}{\sqrt{1+\tau_1^2}}\right)\left\{1-\Phi\left(\frac{\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_1}{\sqrt{1+\tau_1^2}}\right)\right\}, \ldots, \Phi\left(\frac{\mathbf{X}_n^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_n}{\sqrt{1+\tau_n^2}}\right)\left\{1-\Phi\left(\frac{\mathbf{X}_n^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_n}{\sqrt{1+\tau_n^2}}\right)\right\}\right],$$
$$\boldsymbol{V}_D = \mathrm{diag}[H(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}^*/a)\{1-H(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}^*/a)\}, \ldots, H(\mathbf{X}_n^{\mathrm{T}}\boldsymbol{\beta}^*/a)\{1-H(\mathbf{X}_n^{\mathrm{T}}\boldsymbol{\beta}^*/a)\}].$$

A proof of the theorem is in the Appendix.

2.3. *M-estimation under working linear regression model*

Let us consider a commonly used robust M-estimation method. According to the standard M-estimation procedure and under the working linear regression model (6), a commonly used M-estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ solves the following estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_c\left(\frac{Y_i^*-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{s}\right)\mathbf{X}_i = 0, \tag{15}$$

where $\Psi_c(t) = (-c)\mathbf{1}_{(t<-c)} + t\mathbf{1}_{(|t|\leqslant c)} + c\mathbf{1}_{(t>c)}$ is Huber's winsorizing function [12, 17]. The default choice is usually $c = 1.345$, which gives 95 per cent efficiency for normal samples. The scale s is a measure of the variation of the response Y_i^*. It can be either assumed to be known or estimated by an M-estimation method or the median absolute deviance (MAD) [12]. To simplify our theoretical discussion, we assume that s is the true standard deviance of Y_i so that $s = \{\mathrm{Var}(Y_i^*)\}^{1/2} = \sqrt{1+\tau_i^2}$. Then, $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ solves equations

$$S_M(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\Psi_c\left(\frac{Y_i^*-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{1+\tau_i^2}}\right)\mathbf{X}_i = 0. \tag{16}$$

In practice, if the scale s is estimated by an M-estimator, it is usually a consistent estimator of $\{\mathrm{Var}(Y_1^*)\}^{1/2}$, provided that $\tau_1 \equiv \cdots \equiv \tau_n$. In this case, the $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ obtained from the above two sets of M-estimating equations (15) and (16) are asymptotically equivalent. Since the Ψ_c function is a symmetric function, in the presence of biased extraneous noises (i.e. $\eta_i \neq 0$), these M-estimating equations for $\boldsymbol{\beta}$ usually are not Fisher consistent.

Suppose there exists a solution, say $\tilde{\boldsymbol{\beta}}$, to the following equations:

$$E\{S_M(\boldsymbol{\beta})\}=\frac{1}{n}\sum_{i=1}^{n}E\left\{\Psi_c\left(\frac{Y_i^*-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{1+\tau_i^2}}\right)\right\}\mathbf{X}_i=0. \tag{17}$$

In general, $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ are different. Write

$$\bar{\boldsymbol{\delta}}_M=\left(E\left\{\Psi_c\left(\frac{Y_1^*-\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{1+\tau_1^2}}\right)\right\},\ldots,E\left\{\Psi_c\left(\frac{Y_n^*-\mathbf{X}_n^{\mathrm{T}}\boldsymbol{\beta}}{\sqrt{1+\tau_n^2}}\right)\right\}\right)^{\mathrm{T}}.$$

The first result of the following lemma provides a formula to describe the difference between $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ in terms of $\bar{\boldsymbol{\delta}}_M$. The second result of the lemma suggests that the M-estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ and $\tilde{\boldsymbol{\beta}}$ are asymptotically equivalent.

Lemma 2
Under the conditions described in Lemma 1, we have

(i) $\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_M^*\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\bar{\boldsymbol{\delta}}_M$,
where $\boldsymbol{V}_M^*=\mathrm{diag}\{\tilde{v}_1/\sqrt{1+\tau_1^2},\ldots,\tilde{v}_n/\sqrt{1+\tau_n^2}\}$ with $\tilde{v}_i=\Phi(c-\xi_i)-\Phi(-c-\xi_i)$ for some ξ_i between $d_i=\eta_i/\sqrt{1+\tau_i^2}$ and $\tilde{d}_i=\{\eta_i+\mathbf{X}_i^{\mathrm{T}}(\boldsymbol{\beta}_0-\tilde{\boldsymbol{\beta}})\}/\sqrt{1+\tau_i^2}$.

(ii) $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}-\tilde{\boldsymbol{\beta}}=O_p(n^{-1/2})$.

A proof of the lemma is in the Appendix.

Theorem 2 below provides a set of asymptotic expressions for the mean and the variance of the robust M-estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$.

Theorem 2
Under the conditions of Lemma 1, the asymptotic mean and the variance of the robust estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ are

$$E\hat{\boldsymbol{\beta}}^{(\mathrm{M})}=\tilde{\boldsymbol{\beta}}+o(n^{-1/2}),$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{M})})=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_M\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_M\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_M\boldsymbol{X})^{-1}+o(n^{-1}),$$

where

$$\boldsymbol{W}_M=\mathrm{diag}\left\{\mathrm{Var}\left(\Psi_c\left(\frac{Y_1^*-\mathbf{X}_1^{\mathrm{T}}\tilde{\boldsymbol{\beta}}}{\sqrt{1+\tau_1^2}}\right)\right),\ldots,\mathrm{Var}\left(\Psi_c\left(\frac{Y_n^*-\mathbf{X}_n^{\mathrm{T}}\tilde{\boldsymbol{\beta}}}{\sqrt{1+\tau_n^2}}\right)\right)\right\}=\mathrm{diag}(g_1,\ldots,g_n),$$

$$\boldsymbol{V}_M=\mathrm{diag}\left\{\frac{v_1}{\sqrt{1+\tau_1^2}},\ldots,\frac{v_n}{\sqrt{1+\tau_n^2}}\right\}.$$

Here, $g_i=1-[\int_{c-\tilde{d}_i}^{\infty}\{x^2-(\tilde{d}_i-c)^2\}\phi(x)\,\mathrm{d}x+\int_{-\infty}^{-c-\tilde{d}_i}\{x^2-(\tilde{d}_i+c)^2\}\phi(x)\,\mathrm{d}x]-\{\int_{c-\tilde{d}_i}^{\infty}(x+\tilde{d}_i-c)\phi(x)\,\mathrm{d}x+\int_{-\infty}^{-c-\tilde{d}_i}(x+\tilde{d}_i+c)\phi(x)\,\mathrm{d}x\}^2$, $\tilde{d}_i=\{\eta_i+\mathbf{X}_i^{\mathrm{T}}(\boldsymbol{\beta}_0-\tilde{\boldsymbol{\beta}})\}/\sqrt{1+\tau_i^2}$ and $v_i=\Phi(c-\tilde{d}_i)-\Phi(-c-\tilde{d}_i)$.

The proof of the theorem can be found in the Appendix.

3. Case studies: comparison of three estimators

In this section, we consider two special cases of the contamination model (4). In the first case, there is no contamination. In the second case, the observed response is contaminated by an additive error whose mean and variance are proportional to the mean and the variance of the uncontaminated true response, respectively.

3.1. The case of no random contamination

In the case when there is no contamination (i.e. $\eta_i \equiv 0$ and $\tau_i \equiv 0$ for all i), $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ are just the least squares and robust M-estimators from the standard linear regression model (1), and $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ is the maximum likelihood estimator from the standard logistic model (3). Therefore,

$$E\hat{\boldsymbol{\beta}}^{(\mathrm{LS})} - \boldsymbol{\beta}_0 = 0 \quad \text{and} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}.$$

Since $\boldsymbol{\delta}_D \approx 0$ when $\eta_i \equiv 0$ and $\tau_i^2 \equiv 0$, a direct application of Lemma 1 and the approximation between logit and probit functions yield that $\boldsymbol{\beta}^* \approx \boldsymbol{\beta}_0$ and $\boldsymbol{W}_D \approx \boldsymbol{V}_D$. Thus, by Theorem 1,

$$E\hat{\boldsymbol{\beta}}^{(\mathrm{D})} - \boldsymbol{\beta}_0 \approx o(n^{-1/2}) \quad \text{and} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{D})}) \approx a^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_D\boldsymbol{X})^{-1} + o(n^{-1}).$$

Also, note that $\tilde{d}_i = 0$, when $\eta_i \equiv 0$ and $\tau_i^2 \equiv 0$. From Theorem 2,

$$E\hat{\boldsymbol{\beta}}^{(\mathrm{M})} - \boldsymbol{\beta}_0 = o(n^{-1/2}) \quad \text{and} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{M})}) = \tilde{c}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} + o(n^{-1}),$$

where $\tilde{c} = g_i / v_i^2 = \{1 - \int_{|x|>c}(x^2 - c^2)\phi(x)\,\mathrm{d}x\}/\{\Phi(c) - \Phi(-c)\}^2 = 1.053$ when $c = 1.345$.

We can show that $a^{-2}\boldsymbol{V}_D \leqslant \tilde{c}^{-1}\boldsymbol{I}_n \leqslant \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the identity matrix and the inequality between two symmetric matrices $\boldsymbol{A} \leqslant \boldsymbol{B}$ means that the matrix $\boldsymbol{B} - \boldsymbol{A}$ is non-negative definite. It follows that, when $\eta_i \equiv 0$, $\tau_i \equiv 0$ and n tends to infinity,

$$E\hat{\boldsymbol{\beta}}^{(\mathrm{D})} \approx E\hat{\boldsymbol{\beta}}^{(\mathrm{M})} = E\hat{\boldsymbol{\beta}}^{(\mathrm{LS})} = \boldsymbol{\beta}_0 \quad \text{and} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{D})}) \geqslant \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{M})}) \geqslant \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}),$$

where the first equation for the means and the first inequality for the variances are in an approximate sense, due to the approximation between the logistic function $H(t)$ and the probit function $\Phi(t)$.

In summary, in the case when there is no contamination, these three estimators are unbiased or asymptotically unbiased. The least-squares estimator is most efficient and the estimator from the dichotomized model is least efficient. These are well-known facts. In this case, both working models reflect the true stochastic relation between the response and the covariates, but less information is contained in the dichotomized responses.

3.2. A case study with a proportional type of random contaminations

We consider in this section that the contamination error e_i has a bias and variance proportional to the mean and the variance of the true response for all subjects, i.e.

$$\eta_i = \lambda\mu_i \text{ and } \tau_i^2 = \kappa\sigma^2, \quad i = 1, 2, \ldots, n, \text{ for some } \lambda \geqslant 0 \text{ and } \kappa > 0, \tag{18}$$

where $\sigma = 1$ as in Section 1. Under this setting, we can express the biases and the variances of $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$, $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ explicitly and the level of contamination can be simply described using λ and κ. Note that, (18) implies

$$E(Y_i^*)/E(Y_i) = 1 + \lambda \quad \text{and} \quad \mathrm{Var}(Y_i^*)/\mathrm{Var}(Y) = 1 + \kappa.$$

Thus, $1 + \lambda$ and $1 + \kappa$ can be interpreted as the ratios of means and variances between the observed (contaminated) response Y_i^* and the underlying true response Y_i.

Under the assumption (18), it follows that

$$E\hat{\boldsymbol{\beta}}^{(\mathrm{LS})} = (1 + \lambda)\boldsymbol{\beta}_0 \quad \text{and} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}) = (1 + \kappa)(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}. \tag{19}$$

On the other hand, by the approximation between logit and probit functions, we have

$$\Phi\left(\frac{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \eta_i}{\sqrt{1 + \tau_i^2}}\right) \approx H\left(\frac{(1 + \lambda)\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0}{a\sqrt{1 + \kappa}}\right).$$

We can prove from (13) that

$$\boldsymbol{\beta}^* \approx \frac{1 + \lambda}{\sqrt{1 + \kappa}}\boldsymbol{\beta}_0.$$

Thus, by Theorem 1, we have

$$E\hat{\boldsymbol{\beta}}^{(\mathrm{D})} \approx \frac{1+\lambda}{\sqrt{1+\kappa}}\boldsymbol{\beta}_0 + o(n^{-1/2}) \quad \text{and} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{D})}) \approx a^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_D^{(0)}\boldsymbol{X})^{-1} + o(n^{-1}), \tag{20}$$

where $\boldsymbol{V}_D^{(0)} = \mathrm{diag}\{h_1(1-h_1), \ldots, h_n(1-h_n)\}$ and $h_i = H\left(\{(1+\lambda)\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0\}/\{a\sqrt{1+\kappa}\}\right)$. Also, under the assumption (18), we can obtain from (17) that the solution

$$\tilde{\boldsymbol{\beta}} = (1+\lambda)\boldsymbol{\beta}_0.$$

Therefore, by Theorem 2, we have

$$E\hat{\boldsymbol{\beta}}^{(\mathrm{M})} = (1+\lambda)\boldsymbol{\beta}_0 + o(n^{-1/2}) \quad \text{and} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{M})}) = \tilde{c}(1+\kappa)(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} + o(n^{-1}), \tag{21}$$

where $\tilde{c} = 1.053$ when the default choice $c = 1.345$ is used in Huber's winsorizing function. From (19) and (21), the M estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ has the same mean but a larger variance than that of the least-squares estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$. Hence, the robust M-estimation method is not effective for dealing with the type of contaminations modeled in (18).

Based on the formulas in (19), (20) and (21), we can also compare the performance of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ with $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$. The results are summarized in the following propositions. Although this comparison is true approximately, simulation studies in Section 4 suggest that it is very accurate. Note that, in the large sample setting, the variance terms typically tend to zero at the order of $O(n^{-1})$, but the biases are typically at the order of $O(1)$ if they are not zero.

Proposition 1

Under the conditions of Lemma 1, the necessary and sufficient condition for $\left\|E\hat{\boldsymbol{\beta}}^{(\mathrm{LS})} - \boldsymbol{\beta}_0\right\| \geqslant \left\|E\hat{\boldsymbol{\beta}}^{(\mathrm{D})} - \boldsymbol{\beta}_0\right\|$ or $\left\|E\hat{\boldsymbol{\beta}}^{(\mathrm{M})} - \boldsymbol{\beta}_0\right\| \geqslant \left\|E\hat{\boldsymbol{\beta}}^{(\mathrm{D})} - \boldsymbol{\beta}_0\right\|$ is

$$\lambda \geqslant \frac{\sqrt{1+\kappa}-1}{\sqrt{1+\kappa}+1}. \tag{22}$$

The necessary and sufficient condition for $\mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}) \geqslant \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{D})})$ is

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_D^{(0)}\boldsymbol{X}/a^2 \geqslant \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/(1+\kappa). \tag{23}$$

The necessary and sufficient condition for $\mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{M})}) \geqslant \mathrm{Var}(\hat{\boldsymbol{\beta}}^{(\mathrm{D})})$ is

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_D^{(0)}\boldsymbol{X}/a^2 \geqslant \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/\{\tilde{c}(1+\kappa)\}. \tag{24}$$

In large sample situations, the bias terms dominate the comparison unless both biases are $O(n^{-1/2})$.

Let us recall that equations (18) define λ and κ as the levels of the contamination in mean and variance, proportional to the true mean and variance values, respectively. Condition (22) suggests that, whenever the magnitude of λ is greater than $1-2/(1+\sqrt{\kappa+1})$, the bias of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ is less than both the bias of $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$. Figure 1(a) and (d) depict the area of (κ, λ) values, in which the estimator from the binary regression $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ has smaller bias. Note that the boundary of the area is the curve $\lambda = (\sqrt{\kappa+1}-1)/(\sqrt{\kappa+1}+1)$. This curve increases in λ as κ increases, but it is always bounded above by the $\lambda = 1$ line. This area does not cover the case with $\lambda = 0$ when there is no bias in the error. In this case, both $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ are still unbiased or asymptotically unbiased, but $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ is biased and not consistent. However, as long as (κ, λ) falls inside the shaded area of Figure 1(a) or (d), the bias of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ is smaller than the biases of $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$. This is true even in some cases when the errors are very small (both λ and κ are very small), noting that $(\kappa, \lambda) = (0, 0)$ is a point on the boundary of the shaded area.

The conditions for the variance comparisons (23) and (24) are much more complicated than the condition (22) for the mean comparisons. They typically involve the covariates $\boldsymbol{X}$ and the true parameters $\boldsymbol{\beta}_0$, unless $\kappa \to \infty$ at which both the inequalities in (23) and (24) always hold. Thus, loosely speaking, when the magnitude of the error in variance κ gets larger and larger, the parameter estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ from the binary regression tends to have smaller variance than the least squares and M-estimators from the regular linear regression model. But the exact conditions depend on the covariates $\boldsymbol{X}$ and true parameters $\boldsymbol{\beta}_0$, as well as the magnitude of errors in the mean λ. To illustrate the conditions (23) and (24), we plot in Figure 1 (b) and (e) the variance comparisons between $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ as well as $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$, using a set of $\boldsymbol{X}$ and

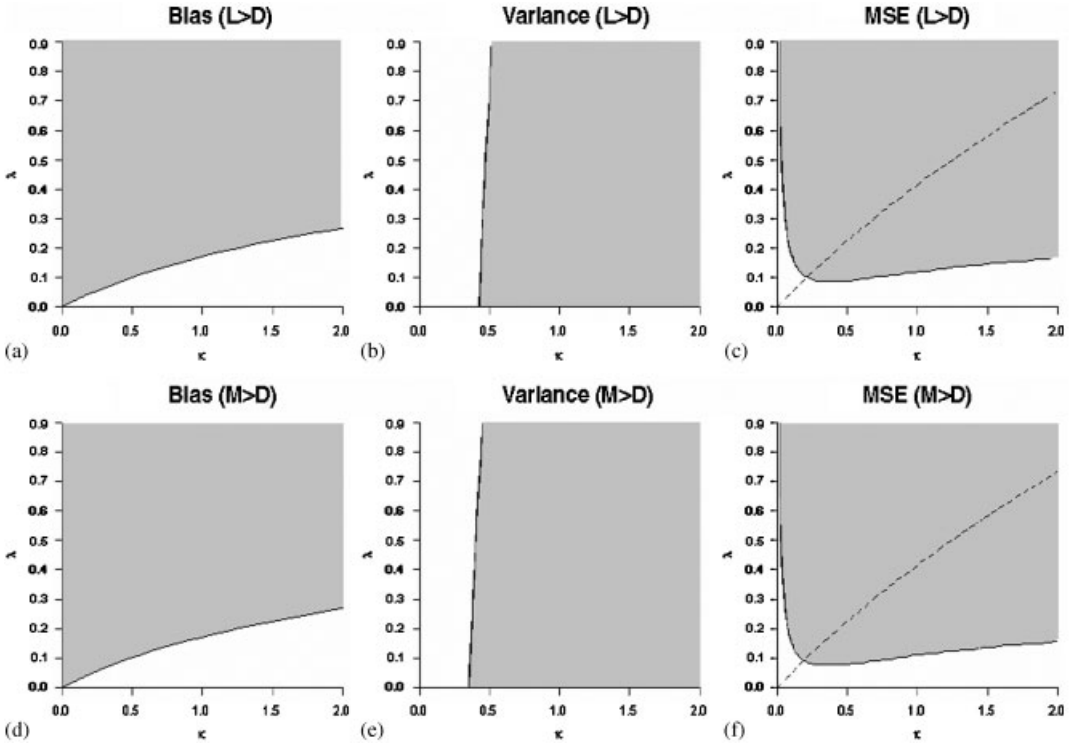

Figure 1. The plots (a)–(c) depict, for a range of κ and λ, the comparisons between the least-squares estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ and the dichotomized estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$. The plots (d)–(f) depict the comparisons between the M-estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ and the dichotomized estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$. The shaded regions in (a) and (d) correspond to the (κ, λ) values at which the parameter estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ has less bias than the least-squares estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ and the M-estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$. The shaded regions in (b) and (d) correspond to the (κ, λ) values at which $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ has a smaller variance than $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ or $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$. The shaded regions in (c) and (f) correspond to the (κ, λ) values at which $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ has a smaller MSE than $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ or $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$. At the dotted curves in (c) and (f), the bias of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ is close to zero. The variance and MSE comparisons involve the covariates $\boldsymbol{X}$ and true parameters $\boldsymbol{\beta}_0$, which are obtained from the real data setting of McGrath and Bedi [20] as described in Section 4.

$\boldsymbol{\beta}_0$ obtained from a real data setting of McGrath and Bedi [20]; see more details next in Section 4. Since $\tilde{c} \approx 1.053$, the shaded area in Figure 1 (e) is slightly larger than that in Figure 1 (b). In this example, it appears that the contamination in variance κ needs to reach certain level before resulting in a smaller variance in $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$, although the level κ could be as small as $\kappa = 50$ per cent depending on λ.

From the proposition, we can also derive the necessary and the sufficient conditions for the comparisons of mean square errors $\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}) \geqslant \mathrm{MSE}(\hat{\boldsymbol{\beta}}^{(\mathrm{D})})$ and $\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{(\mathrm{M})}) \geqslant \mathrm{MSE}(\hat{\boldsymbol{\beta}}^{(\mathrm{D})})$. Figure 1 (c) and (f) depicts the comparisons of MSE of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ with those of $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$, using the same set of $\boldsymbol{X}$ and $\boldsymbol{\beta}_0$ that produce Figure 1 (b) and (e). In particular, the shaded regions cover the upper right areas of the plots where the errors (both κ and λ) are big. Thus, dichotomization is typically preferred when the errors are large. This is not surprising, since contamination errors often have larger influence on the estimators using continuous responses than that using dichotomized responses. However, even when the contamination is small, there are cases that the dichotomization is preferred. In the example of Figure 1 (c) and (f), for instance, when the contamination error is as small as 10 per cent bias in mean and 20 per cent inflation in variance, the estimator using the dichotomized responses can still outperform the estimators using the continuous responses.

4. Data analysis: simulation studies

In this section, we provide two simulation studies to examine what might happen in actual data analysis and to validate the theoretical developments in Sections 2 and 3. In particular, we compare the performance of the numerical estimates $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$, $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ from fitting simulated data sets of $(\mathbf{Y}^*, \boldsymbol{X})$.

The simulation is performed by mimicking the real data setting of McGrath and Bedi's study [20]. The aim of their study is to 'establish normative age–gender values for the U.K. oral Health-related Quality of Life Measure (OHQoL-UK)

in Britain and to identify key factors associated with oral health-related quality of life measure in the U.K.'. Their data came from surveys. The main survey of 'OHQoL-UK' had 16 items, and they were used to produce overall OHQoL-UK scores, which ranged between 16 and 80 with standard error $\sigma=0.6856$. The overall OHQoL-UK scores were not only considered as continuous, but they were also dichotomized. McGrath and Bedi [20] used logistic models to study the dichotomized OHQoL-UK scores. Among various predictor variables, age (W_1), social class (W_2) and oral health status (W_3) were reported to be significant factors in their final model:

$$P(Z=1)=\frac{\exp(-0.75+0.28W_1+0.35W_2+0.75W_3)}{1+\exp(-0.75+0.28W_1+0.35W_2+0.75W_3)}.$$

In their study, there were $n=1838$ subjects in total, 22 per cent of which were over 65 ($W_1=1$) and 78 per cent below 65($W_1=0$). Also, 54 per cent of the subjects were from higher social class ($W_2=1$) and 47 per cent were from lower class($W_2=0$), and 70 per cent had more than 20 teeth ($W_3=1$) and 30 per cent had less than 20 teeth ($W_3=0$).

In our simulation, we use binomial models to simulate $n=1838$ covariate vectors (W_1, W_2, W_3) where there are roughly 22 per cent $W_1=1$, 54 per cent $W_2=1$ and 70 per cent $W_3=1$. To reflect the dependence of W_3 on W_1 and W_2, in the simulation the conditional mean of $E(W_3|W_1, W_2)$ is modeled by $\exp(0.8-0.3W_1+0.2W_2)/\{1+\exp(0.8-0.3W_1+0.2W_2)\}$. We generate a set of true response values Y_i using model (1) with $\sigma^2=1$ and

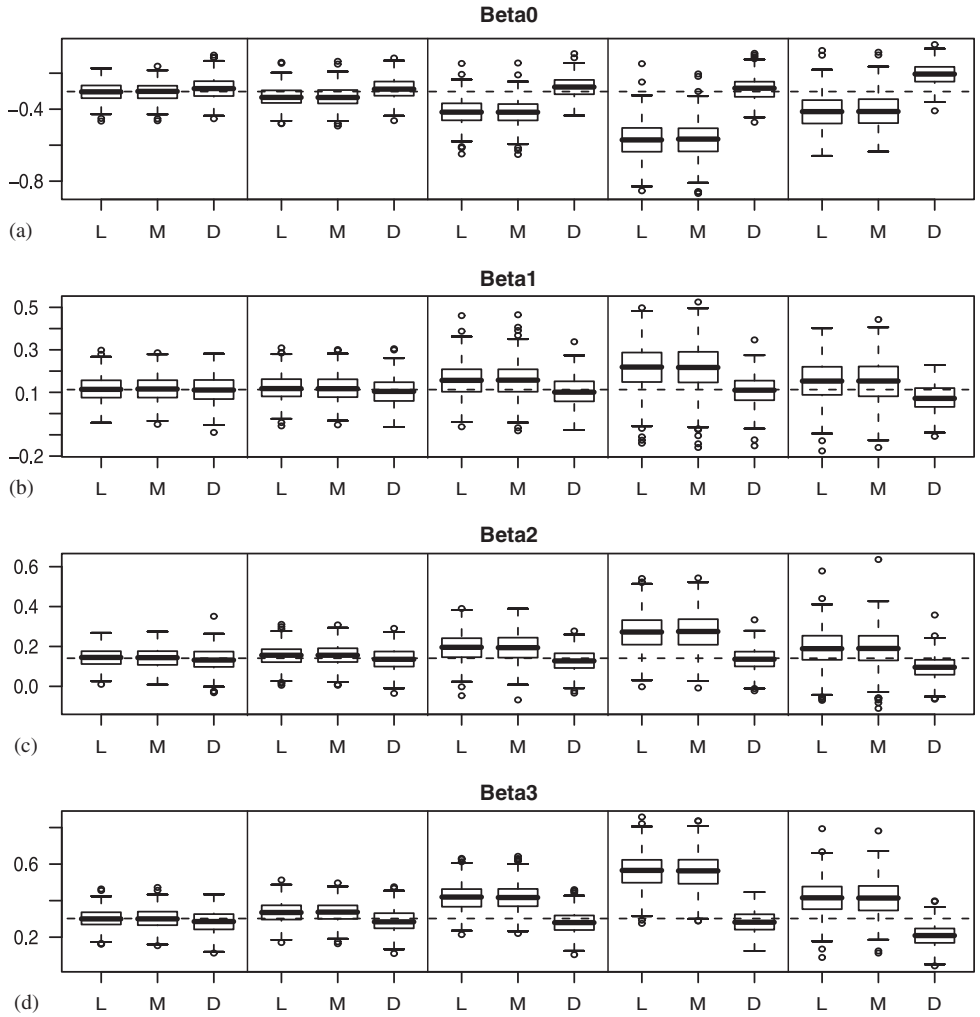

Figure 2. The figure contains 4×5 sets of side-by-side boxplots arranged in five columns, corresponding to five pairs of (κ, λ) values: (i) $(\kappa, \lambda)=(0, 0)$; (ii) $(\kappa, \lambda)=(0.2, 0.1)$; (iii) $(\kappa, \lambda)=(1.0, 0.4)$; (iv) $(\kappa, \lambda)=(2.5, 0.9)$; and (v) $(\kappa, \lambda)=(2.5, 0.4)$. The four regression parameters are: (a) intercept β_0; (b) β_1; (c) β_2; and (d) β_3. The letter "L" on the x-axes represents the least-squares estimates $\hat{\boldsymbol{\beta}}^{(LS)}$, the letter 'M' represents the M-estimates $\hat{\boldsymbol{\beta}}^{(M)}$ and the letter 'D' represents the estimates using dichotomized responses $\hat{\boldsymbol{\beta}}^{(D)}$. The four horizontal broken lines mark the true values of the four corresponding regression parameters. The plots at each setting are produced from data analysis of fitting 500 simulated data sets.

$\boldsymbol{\beta}_0 = (-0.75, 0.28, 0.35, 0.75) \times 0.6856/1.702 = (-0.3021, 0.1128, 0.1410, 0.3021)$. We then add a set of random contamination errors $e_i \sim (\lambda\mu_i, \kappa)$ with $\mu_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0$ to form contaminated responses $Y_i^* = Y_i + e_i$. We assume in our data analysis that we only know the predictor variables (W_1, W_2, W_3) and the contaminated responses Y_i^*. The dichotomized responses are $Z_i^* = I(Y_i^* > 0)$. We fit the working regression models (6) and (7) to the data using Y_i^* and Z_i^*, respectively, and study the performance of the numerical estimates $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$, $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$.

In the first simulation study, we consider five sets of (κ, λ) values: (i) $(\kappa, \lambda) = (0, 0)$, (ii) $(\kappa, \lambda) = (0.2, 0.1)$, (iii) $(\kappa, \lambda) = (1, 0.4)$, (iv) $(\kappa, \lambda) = (2.5, 0.9)$ and (v) $(\kappa, \lambda) = (2.5, 0.4)$. The first set (i) $(\kappa, \lambda) = (0, 0)$ corresponds to the case that there is no contamination error. The other four sets of (κ, λ) values are all in shaded regions of Figure 1 (c) and (f). The second set (ii) $(\kappa, \lambda) = (0.2, 0.1)$ corresponds to the case that contamination errors are relatively small, and the rest of the three sets (iii), (iv) and (v) represent different cases that contamination errors are significant. The third and fourth sets (iii) and (iv) are along the dotted curves in Figure 1 (c) and (f) at which the bias of $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ is close to zero. The fifth set (v) corresponds to large contamination errors and its (κ, λ) value is far away from the dotted curves. This simulation and data analysis are repeated 500 times for each of the five sets of (κ, λ) values.

Figure 2(a)–(d) are side-by-side boxplots of 500 estimates of the four regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ for the five sets of (κ, λ) values. Each triplet of side-by-side boxplots, marked by 'L', 'M' and 'D', are from fitting the working linear regression model (6) and the working binary logistic model (7), respectively. The four horizontal broken lines in (a)–(d) indicate, respectively, the corresponding true β_0, β_1, β_2 and β_3 values. In the first case (i) when there is

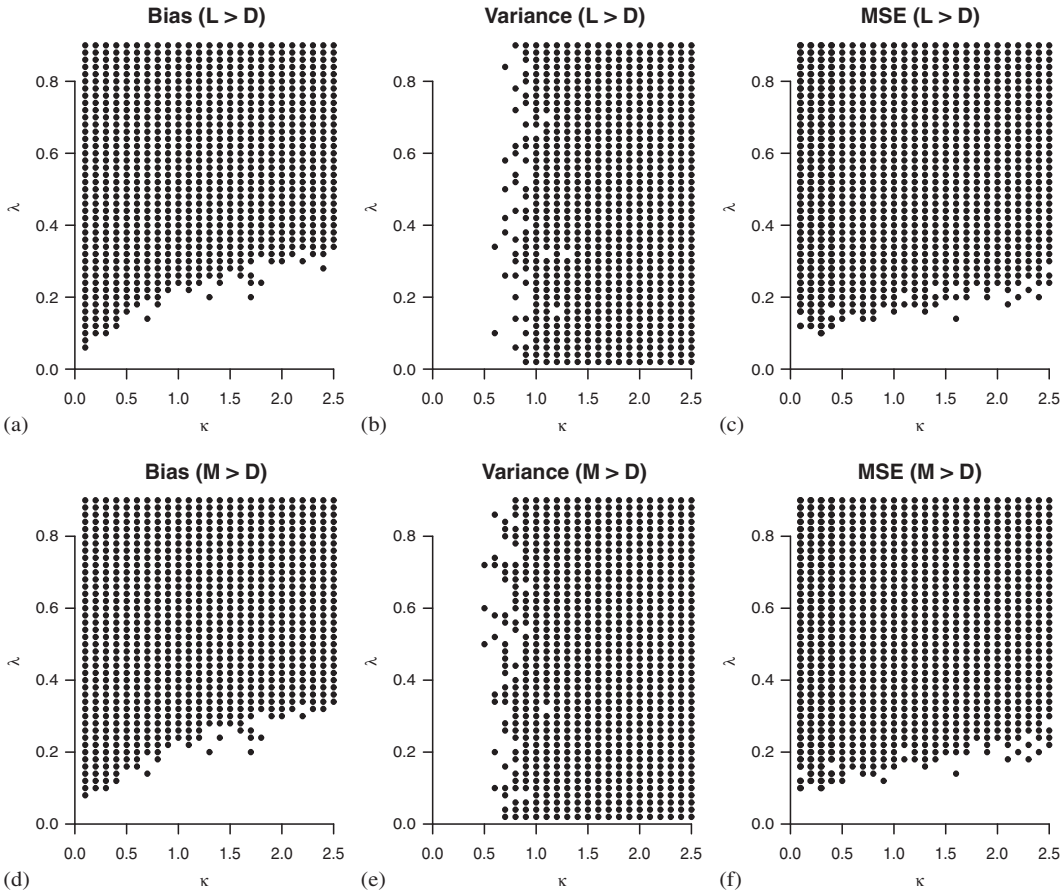

Figure 3. The six plots in (a)–(f) correspond to, respectively, the six regions in Figure 1. Unlike Figure 1, these plots are produced based on data analysis results from fitting 100 simulated data sets at each of 25×50 grids values of (κ, λ) over the range of $(0, 2.5) \times (0, 1)$. In particular, (a)–(c) plot those (κ, λ) grid values at which $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ has the smaller average bias, variance, sample MSE than those of $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$, respectively. (d)–(f) plot those (κ, λ) grid values at which $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ has the smaller average bias, variance, sample MSE than those of $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$, respectively. The shapes and margins of these plots from data analysis are consistent with those in Figure 1. The dotted curves (slightly bended) in (d) and (f) mark the (κ, λ) values at which the parameter estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ from the logistic model (7) is asymptotically unbiased.

no contamination, the estimates are all on target. The boxplots of least-squares estimates are tightest and the boxplots of the estimates from dichotomization are most spread. This is consistent with our understanding that, in the absence of contamination, the continuous responses are more informative than the dichotomized responses. In the other four cases, including the case (ii) where the contamination is very slight, we can see that the parameter estimates for the working logistic regression perform better than both the least-squares estimates and the robust M-estimates from the working linear regression model. For the type of contaminations close to the dotted curves in Figure 1 (c) and (f) (e.g. cases (iii) and (iv)), the working logistic model can provide very reasonable estimates, much better than the estimates from the working linear regression model. The benefit is more significant as the level of contamination increases. In the fifth case (v) where the contamination errors are very large and it is far away from the dotted curve, the estimates from both working models are poor. But even in this case, the estimates using the dichotomized responses are better than those using the linear regression model. All the results are consistent with the theoretical discussions in Sections 2 and 3.

In the second simulation study, we use data analysis to re-produce the regions in Figure 1 and to validate the developments in Section 3. In this simulation, the (κ, λ) region of $(0, 0.25) \times (0, 1)$ is divided into 25×50 grids. At each fixed grid of (κ, λ) values, we run the simulation 100 times and obtain 100 estimates of $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$, $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$. The average bias, variance and mean square errors are computed from each of the three sets of 100 estimates. Figure 3(a)–(c) plots in black dots those (κ, λ) grid values at which $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ has the smaller average bias, variance, sample mean square error than those of $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$, respectively. The same plots are plotted in Figure 3(d)–(f), except replacing $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ with $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$. From Figure 3, we can see that the regions of (κ, λ) obtained from the data analysis match very well, in both shapes and locations, to those of the corresponding ones in Figure 1.

We can conclude that in actual data analysis $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ could perform better than $\hat{\boldsymbol{\beta}}^{(\mathrm{LS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ in some situations when there are contaminations in the response data. The bias terms, if not zero, often dominate the comparisons. The binary model starts to show benefits even in some cases of small contaminations. When the (κ, λ) values are along the dotted lines of Figure 1 (c) or (f), the estimates obtained from binary are usually very good even when the contamination is large. If (κ, λ) values are within shaded regions of Figure 1, the estimates from the binary model are favorable compared with those from the linear model.

5. Discussion

As mentioned in [1], although 'necessary and sensible' to dichotomization in some practical settings, such simplicity is gained at a 'high cost', and may well 'create problems rather than solve them' in a research context. There is a considerable methodological literature 'examining and demonstrating negative consequences of dichotomization' and 'firmly favoring the use of regression methods on undichotomized variables' [7]. The main reason is that dichotomization of continued observations causes loss of information in statistical analysis, and this is especially true in the standard case of no contaminations. However, there is also the other side of the story in the case when the observations are contaminated or the analysis model is misspecified. This note illustrates that, in the presence of unknown additive contamination errors, dichotomization of response variable can sometimes produce a better result in statistical analysis. This surprising result can be explained by the trade-off between the loss of information and the reduction of contamination input. Our main objective of this research is to raise the awareness about the complexity of problems involving dichotomization. We should not dismiss any procedures without carefully assessing the situation in each specific case. Although our discussion is based on dichotomizing response variable in regression models, we believe that the same trade-off and a similar discussion could also apply to the practice of dichotomizing covariate variables.

When there are contaminations, a natural solution is to improve the accuracy and obtain data of a better quality, if possible. Or, if the contamination mechanism and sources are known, we may also be able to use modeling or other approaches to mitigate or even eliminate the errors. However, there are practices in which it is hard to significantly improve the quality of data under the current technology. Other times, it is just too expensive to obtain better quality data even when the technology is available. For instances, as pointed out by Mecham *et al.* [21], despite that it has been 'tremendously successful', the microarray technology has suffered greatly 'from suboptimal measurement precision'. In the high-throughput experiments in drug discovery, a large number of chemical compounds, sometime in millions or billions, are tested in a quick and cheap way [22]. Although there are more accurate ways to measure the chemical compound potencies, the costs (both in time and money) of screening a large number of compounds often hamper the use of any higher quality but more expensive tests. In these examples, dichotomization may sometimes be beneficial as a statistical methodology.

There are several possible approaches for determining the cutpoint (threshold) for dichotomization. In many situations, there are recognized thresholds that are 'widely accepted' and 'used in previous studies' [1, 7]. For instance, the threshold

to dichotomize patients' Hemoglobin A1c levels is often set to the upper limit of the reference interval in healthy individuals [1, 3]. This type of threshold, not depending on the sample data, is often considered as constant. Our development in the paper covers this type of thresholds. Other times, in the absence of a known or (clinically) meaningful threshold, a common approach is to take the sample median or sample mean, or sometimes it uses 'optimal cutpoint' which is determined by sample data using some criteria [1]. This type of thresholds typically depend on the sample data, and different studies of even the same type will use different thresholds. The theoretical development in this case depends on how the threshold is related to the data and it can be very complex. Our development cannot directly apply to dichotomizations using this type of sample-dependent thresholds, although we believe that the general trade-off between loss of information and reduction of contamination input is still intact.

Finally, we conclude the paper with a technical remark on the approximation between the standard logistic and the probit functions, which is used in the developments in Section 3 but not in Sections 2 and 4. The agreement between the simulation results in Section 4 and those developed in Section 3 provides supporting evidence that this approximation is fairly accurate and does not really affect the data analysis results. Of course, one can have the same developments using a working probit model, instead of the working logistic model (7). If the probit model is used, all the approximations '$\approx$' in the developments in Section 3 can be replace by actual equations '='. However, since the probit model is not a canonical link generalized linear model [19], the technical details of the developments will be more complicated when comparing to the logistic model. We use in the paper the logistic model because it is most commonly used in practice and also because the technical details are simpler than those if we use a probit model.

Appendix A

Sketch proof of Lemma 1

(i) Since $\boldsymbol{\beta}^*$ is a solution to (13), we have

$$\begin{aligned}
0 = E\{S_D(\boldsymbol{\beta}^*)\} &= \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\bar{\boldsymbol{\delta}}_D + \frac{1}{n}\sum_{i=1}^{n}[H(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0/a) - H(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*/a)]\mathbf{X}_i \\
&= \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\bar{\boldsymbol{\delta}}_D + \frac{1}{n}\sum_{i=1}^{n} H(\xi_i)\{1-H(\xi_i)\}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}(\boldsymbol{\beta}_0-\boldsymbol{\beta}^*)/a \\
&= \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\bar{\boldsymbol{\delta}}_D + \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_D^*\boldsymbol{X}(\boldsymbol{\beta}_0-\boldsymbol{\beta}^*)/a,
\end{aligned}$$

for some ξ_i between $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0/a$ and $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*/a$. The first conclusion of Lemma 1 is immediate from the above equation.

(ii) Under the assumed conditions, we have

$$\frac{1}{n}\sum_{i=1}^{n}\{Z_i^* - E(Z_i^*|\mathbf{X}_i)\}\mathbf{X}_i = O_p(n^{-1/2}).$$

By (11), (12) and (13), the above equation leads to

$$\begin{aligned}
O_p(n^{-1/2}) &= \frac{1}{n}\sum_{i=1}^{n}\left[H(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(\mathrm{D})}/a) - \Phi\left(\frac{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\eta_i}{\sqrt{1+\tau_i^2}}\right)\right]\mathbf{X}_i \\
&= \frac{1}{n}\sum_{i=1}^{n}[H(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(\mathrm{D})}/a) - H(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*/a)]\mathbf{X}_i \\
&= \frac{1}{n}\sum_{i=1}^{n} H(\zeta_i)\{1-H(\zeta_i)\}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}^{(\mathrm{D})}-\boldsymbol{\beta}^*)/a,
\end{aligned}$$

for some ζ_i between $\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(\mathrm{D})}/a$ and $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*/a$. Note that, under the assumed conditions, one can prove that $H(\zeta_i)$, for $i=1,2,\ldots,n$, is uniformly bounded away from 0 and 1 and $n^{-1}\sum_{i=1}^{n} H(\zeta_i)\{1-H(\zeta_i)\}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$ is positive definite. The second result of the lemma follows immediately from the above equation.

Sketch proof of Theorem 1
By the Taylor expansion, we have

$$0=S_D(\hat{\boldsymbol{\beta}}^{(\mathrm{D})})=S_D(\boldsymbol{\beta}^*)+S'_D(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}^{(\mathrm{D})}-\boldsymbol{\beta}^*)/a+o_p(n^{-1/2}).$$

Hence $\hat{\boldsymbol{\beta}}^{(\mathrm{D})}$ can be expressed in terms of $\boldsymbol{\beta}^*$:

$$\hat{\boldsymbol{\beta}}^{(\mathrm{D})}=\boldsymbol{\beta}^*-a\{S'_D(\boldsymbol{\beta}^*)\}^{-1}S_D(\boldsymbol{\beta}^*)+o_p(n^{-1/2}).$$

Note that $E\{S_D(\boldsymbol{\beta}^*)\}=0$, $\mathrm{Var}\{S_D(\boldsymbol{\beta}^*)\}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_D\boldsymbol{X}/n^2$ and $S'_D(\boldsymbol{\beta}^*)=-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_D\boldsymbol{X}/n$. The theorem follows immediately.

Sketch proof of Lemma 2

(i) Since the $\Psi_c(t)$ function does not have a derivative only at two points $t=c$ and $t=-c$, we still can use the Taylor expansion on the score function S_M. Also, write $U_i=(Y_i^*-\eta_i-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)/\sqrt{1+\tau_i^2}$. It follows that $U_i\sim\mathrm{N}(0,1)$. Since $\tilde{\boldsymbol{\beta}}$ is a solution to (17), we have

$$\begin{aligned}0=E\{S_M(\tilde{\boldsymbol{\beta}})\}&=E\{S_M(\boldsymbol{\beta}_0)\}+\frac{1}{n}\sum_{i=1}^{n}E[\Psi_c(U_i+\tilde{d}_i)-\Psi_c(U_i+d_i)]\mathbf{X}_i\\&=\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\bar{\boldsymbol{\delta}}_M+\frac{1}{n}\sum_{i=1}^{n}E\Psi_c(U_i+\xi_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}(\boldsymbol{\beta}_0-\tilde{\boldsymbol{\beta}})\\&=\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\bar{\boldsymbol{\delta}}_M+\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_M^*\boldsymbol{X}(\boldsymbol{\beta}_0-\tilde{\boldsymbol{\beta}}),\end{aligned}$$

for some ξ_i between d_i and $\tilde{d}_i$. The first conclusion of Lemma 2 is immediate from the above equation.

(ii) Under the assumed conditions, we have for an fixed $\boldsymbol{\beta}$, $S_M(\boldsymbol{\beta})-E_{\boldsymbol{\beta}_0}S_M(\boldsymbol{\beta})=O_p(n^{-1/2})$. Thus, by (16) and (17), we have in probability

$$\begin{aligned}O_p(n^{-1/2})=S_M(\hat{\boldsymbol{\beta}}^{(\mathrm{M})})-S_M(\tilde{\boldsymbol{\beta}})&=\frac{1}{n}\sum_{i=1}^{n}\{\Psi_c(U_i+\hat{d}_i)-\Psi_c(U_i+\tilde{d}_i)\}\mathbf{X}_i\\&=\frac{1}{n}\sum_{i=1}^{n}\Psi'_c(U_i+\zeta_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}^{(\mathrm{M})}-\tilde{\boldsymbol{\beta}})/\sqrt{1+\tau_i^2},\end{aligned}$$

for some ζ_i between $\hat{d}_i$ and $\tilde{d}_i$. Here, $\hat{d}_i=\{\eta_i+\mathbf{X}_i^{\mathrm{T}}(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_M)\}/\sqrt{1+\tau_i^2}$.

Note that, under the assumed conditions, $n^{-1}\sum_{i=1}^{n}\Psi'_c(U_i+\zeta_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}/\sqrt{1+\tau_i^2}$ is positive definite. The second result of the lemma follows.

Sketch proof of Theorem 2
By the Taylor expansion and in probability 1, we have

$$0=S_M(\hat{\boldsymbol{\beta}}^{(\mathrm{M})})=S_M(\boldsymbol{\beta}_0)+S'_M(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(\mathrm{M})}-\tilde{\boldsymbol{\beta}})+o_p(n^{-1/2}).$$

Hence $\hat{\boldsymbol{\beta}}^{(\mathrm{M})}$ can be expressed as

$$\hat{\boldsymbol{\beta}}^{(\mathrm{M})}=\tilde{\boldsymbol{\beta}}-\{H_M(\tilde{\boldsymbol{\beta}})\}^{-1}S_M(\tilde{\boldsymbol{\beta}})+o_p(n^{-1/2}),$$

where $H_M(\tilde{\boldsymbol{\beta}})=E\{S'_M(\tilde{\boldsymbol{\beta}})\}$. Note that $E\{S_M(\tilde{\boldsymbol{\beta}})\}=0$, $\mathrm{Var}\{S_M(\tilde{\boldsymbol{\beta}})\}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_M\boldsymbol{X}/n^2$ and $H_M(\tilde{\boldsymbol{\beta}})=-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_M\boldsymbol{X}/n$. The theorem follows immediately.

Acknowledgements

This research was partly supported by grants from NSF (DMS0915139, SES0851521), NSA (H98230-08-1-0104) and DHS (2008-DN-077-ARI012-03). The authors wish to thank Chuanwen Chen for his assistance and helpful comments. The authors also wish to thank the editor, associate editor and two referees for their instructive suggestions, which helped to improve the quality of the paper.

References

1. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 2006; **25**:127–141.
2. Xie M, Tatsuoka K, Sacks J, Young SS. Group testing with blockers and synergism. *Journal of the American Statistical Association* 2001; **96**:92–102.
3. Zhang Q, Safford M, Ottenweller J, Hawley G, Repke D, Burgess Jr JF, Dhar S, Cheng H, Naito H, Pogach LM. Performance status of health care facilities changes with risk adjustment of Hemoglobin A1c. *Diabetes Care* 2000; **23**:919–927.
4. Taylor AB, West SG, Aiken LS. Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement* 2006; **66**:228–239.
5. Long JS. *Regression Models for Categorical and Limited Dependent Variables*. Sage: Thousand Oaks, CA, 1997.
6. Cohen J. The cost of dichotomization. *Applied Psychological Measurement* 1983; **7**:249–253.
7. MacCallum RC, Zhang S, Preacher KJ, Rucker D. On the practice of dichotomization of quantitative variables. *Psychological Methods* 2002; **7**:19–40.
8. Morgan BJT. *Analysis of Quantal Response Data*. Chapman and Hall: London, 1992.
9. Demidenko E. *Mixed Models—Theory and Applications*. Wiley: NJ, 2004.
10. Silvapulle MJ. On the existence of the maximum likelihood estimators for the binomial response model. *Journal of the Royal Statistical Society B* 1981; **43**:310–313.
11. Santner T, Duffy D. *The Statistical Analysis of Discrete Data*. Springer: New York, 1989.
12. Venables WN, Ripley BD. *Modern Applied Statistics with S*. Springer: New York, 2002.
13. Uwe J, Gartner H, Rassler S. Measuring overeducation with earnings frontiers and multiply imputed censored income data. *IAB Discussion Paper 200611*, Institute for Employment Research, Nuremberg, Germany, 2006.
14. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models*: *A Modern Perspective* (2nd edn). Chapman and Hall: New York, 2006.
15. Stefanski LA. Measurement error models. *Journal of the American Statistical Association* 2000; **95**:1353–1358.
16. Buonaccorsi JP. Measurement error in the response in the general linear model. *Journal of the American Statistical Association* 1996; **91**:633–642.
17. Huber PJ. *Robust Statistics*. Wiley: New York, 1981.
18. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. Wiley: New York, 1987.
19. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman and Hall: London, 1989.
20. McGrath C, Bedi R. Population based norming of the UK oral health related quality of life measure (OHQoL-UK). *British Dental Journal* 2002; **193**:521–524.
21. Mecham BH, Wetmore DZ, Szallasi Z, Sadovsky Y, Kohane I, Mariani TJ. Increased measurement accuracy for sequence-verified microarray probes. *Physiological Genomics* 2004; **18**:308–315.
22. Abt M, Lim Y, Sacks J, Xie M, Young SS. A sequential approach for screening large chemical databases. *Statistical Sciences* 2001; **16**:154–168.