

Identifiability and bias reduction in the skew-probit model for a binary response

DongHyuk Lee & Samiran Sinha

To cite this article: DongHyuk Lee & Samiran Sinha (2019): Identifiability and bias reduction in the skew-probit model for a binary response, Journal of Statistical Computation and Simulation

To link to this article: <https://doi.org/10.1080/00949655.2019.1590579>



Published online: 14 Mar 2019.



Submit your article to this journal [↗](#)



View Crossmark data [↗](#)



Identifiability and bias reduction in the skew-probit model for a binary response

DongHyuk Lee and Samiran Sinha

Department of Statistics, Texas A&M University, College Station, TX, USA

ABSTRACT

The skew-probit link function is one of the popular choices for modelling the success probability of a binary variable with regard to covariates. This link deviates from the probit link function in terms of a flexible skewness parameter. For this flexible link, the identifiability of the parameters is investigated. Next, to reduce the bias of the maximum likelihood estimator of the skew-probit model we propose to use the penalized likelihood approach. We consider three different penalty functions, and compare them via extensive simulation studies. Based on the simulation results we make some practical recommendations. For the illustration purpose, we analyse a real dataset on heart-disease.

ARTICLE HISTORY

Received 25 September 2018
Accepted 27 February 2019


KEYWORDS

Bias; binary response; bootstrap; information matrix; penalized likelihood; skew-probit link

1. Introduction

Logistic or probit model is widely used for modelling the success probability of a binary variable in terms of covariates. Let Y and \mathbf{X} be the binary response variable and a vector of covariates, then under the logistic model $\text{pr}(Y = 1|\mathbf{X}) = H(\boldsymbol{\gamma}^T \mathbf{Z})$ with $H(u) = \exp(u)/\{1 + \exp(u)\}$, and under the probit model $\text{pr}(Y = 1|\mathbf{X}) = \Phi(\boldsymbol{\gamma}^T \mathbf{Z})$ with $\Phi(u)$ being the cumulative distribution function (CDF) of the standard normal distribution, and $\mathbf{Z} = (1, \mathbf{X}^T)^T$. Both link functions, H and Φ , are considered to be symmetric link functions as they approach to zero and one at the same rate. For a flexible regression model, practitioners may wish to use an asymmetric link that accommodates different convergence rates towards zero and one. Failure to fit a flexible model to the data may result in biased estimates of regression parameters, odds ratios, or risk differences. Towards that goal, Chen et al. [1] introduced a class of asymmetric link functions for modelling binary data and discussed a Bayesian inference in that context. A special case of their link is the *skew-probit* link, where the success probability of the binary response is modelled via the cumulative distribution function (CDF) of the standard skew-normal distribution. In this paper, we consider the skew probit link function that is used to model the success probability of Y

CONTACT Samiran Sinha  sinha@stat.tamu.edu  Department of Statistics, Texas A&M University, 77843-3143 College Station, TX, USA

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/00949655.2019.1590579>

given \mathbf{X} as

$$\text{pr}(Y = 1|\mathbf{X}) = F(\eta, \delta) = \int_{-\infty}^{\eta} 2\phi(u)\Phi(\delta u)du, \quad (1)$$

where $\eta = \mathbf{Z}^T \boldsymbol{\beta} = \beta_0 + \mathbf{X}^T \boldsymbol{\beta}_1$ with $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T)^T$, and $\phi(u) = d\Phi(u)/du$. Note that the integrand in (1) represents the density of the standard skew-normal distribution with the skewness parameter δ , that is denoted by Skew-Normal($\xi = 0, \omega = 1, \delta$). To be clearer, the probability density function of Skew-Normal(ξ, ω, δ) is $2\omega^{-1}\phi\{\omega^{-1}(z - \xi)\}\Phi\{\delta\omega^{-1}(z - \xi)\}$. In Equation (1), F denotes the CDF of Skew-Normal($\xi = 0, \omega = 1, \delta$).

The skew-normal distribution and its properties are well studied in the literature [2–4]. Now we point out some differences of the papers that used the skew-probit link function or its variations. Chen et al. [1] considered the CDF of Skew-Normal($0, 1 + \delta^2, -\delta$) to model the success probability of binary Y . On the other hand, Bazán et al. [5] used the CDF of Skew-Normal($0, 1, \delta$) to model the success probability in the context of item response theory. Bazán et al. [6] considered a unified skew-probit link function that yields the link functions of Chen et al. [1] and Bazán et al. [5] as special cases. Stingo et al. [7] considered an extended family of skew-probit link functions that contains Equation (1) as a special case. Bazán et al. [8] introduced two classes of asymmetrical link functions. One class is based on the CDF of the power-normal distribution that has the density function $f_1(s) = \lambda[\Phi(s)]^{\lambda-1}\phi(s)$, and this link becomes our skew-probit link (1) with $\delta = 1$ when $\lambda = 2$. The other class is based on the CDF of the reciprocal power-normal distribution with the density function $f_2(s) = \lambda[\Phi(-s)]^{\lambda-1}\phi(-s)$, and this link becomes our skew-probit link (1) with $\delta = -1$ when $\lambda = 2$. Kim et al. [9] proposed a generalized t-link function to model binary response variables, and when their parameters $\nu_1 = \nu_2 = \nu \rightarrow \infty$ their link function reduces to the CDF of Skew-Normal($0, 1 + \delta^2, -\delta$).

In this paper, we address two important issues, identifiability of the model parameters and the bias of the maximum likelihood estimator (MLE) of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \delta)^T$ of model (1). A clear knowledge on the identifiability of parameters is necessary for proposing any method of estimation. Secondly, biased estimates may lead to incorrect inference regarding the model parameters, the association between the response and covariates and the marginal effect of the covariate. Although these issues are important for model formulations and deciding on the appropriate method of analysis, to the best of our knowledge, these issues have not been investigated till date.

Now we briefly mention some existing literature on these issues. Genton and Zhang [10] investigated identifiability for some non-Gaussian spatial random fields that include multivariate skew-normal distributions. Castro et al. [11] studied parameter identifiability for multivariate skew-normal distributions. Otiniano et al. [12] investigated parameter identifiability for a finite mixture of skew-normal distributions and a finite mixture of skew- t distributions. Although these approaches considered the important case of a continuous response variable, parameter identifiability has not been investigated for a binary response variable that follows the skew-probit link.

The bias in the MLE of the skew-normal model where the response Y is continuous and follows Skew-Normal(μ, ω, δ), is a well-researched topic. Following Firth [13]'s general recommendation to reducing finite sample bias, Sartori [14] proposed to estimate the skewness parameter δ of the Skew-Normal($\mu = 0, \omega = 1, \delta$) model by maximizing the penalized log-likelihood, $\ell + 0.5\log\{\text{determinant}(\mathcal{I})\}$, where ℓ stands for the log-likelihood

while \mathcal{I} stands for the Fisher information matrix. Sartori [14] also considered estimation of δ in the presence of unknown μ and ω , where only δ was estimated by maximizing a penalized profiled log-likelihood function and the other parameters were estimated by maximizing the likelihood function for a given δ . Later on, Azzalini and Arellano-Valle [15] applied the penalized likelihood idea in the general case of three-parameter Skew Normal(μ, ω, δ) model, where all three parameters were estimated by maximizing the penalized log-likelihood function. To reduce the finite sample bias, researchers considered Bayesian inference of the skew-normal model under various priors including default and proper priors [16,17].

In this paper, we are considering model (1) for a binary response variable Y . Hence, our model is distinct from the models of the papers discussed in the previous two paragraphs where the response Y was considered to be a continuous variable. Furthermore, we are considering the issue in the presence of a regressor variable X that no one has considered before even when a continuous Y followed a skew-normal distribution. As a general strategy to reduce the first order bias in the MLE of β and δ , one may consider the bootstrap bias correction approach or the bias correction approach of Cox and Snell [18]. These two approaches require the MLE to be finite that may not happen in small samples. Therefore, as an alternative, we consider estimation of the parameters by maximizing a penalized likelihood function. In this penalized likelihood method, first we apply Firth [13]'s method to prevent the bias where the likelihood function is penalized by the Jeffrey's prior. Additionally, we consider two more penalization approaches one by using the generalized information matrix prior [19] and two by using the Cauchy prior [20]. Finally, all these methods are compared through extensive simulation studies.

Instead of maximizing an objective function to estimate parameters, one may use a Bayesian approach by specifying a prior distribution for each of the parameters. For instance, Bazán et al. [6] proposed a Bayesian inference for fitting skewed link functions for binary regression, and incorporation of any prior knowledge on parameters is quite useful to reduce bias especially for small to moderate sample sizes. We would like to point out that there is a close connection between the penalized methods we consider and the Bayesian inference. In the penalized likelihood method, first we multiply the likelihood function by the Jeffrey's prior [13], generalized Jeffrey's prior [19] or the Cauchy prior [20], and then we maximize the resultant likelihood. These estimators can be viewed as the mode of the posterior distribution of the parameters (the so-called MAP estimator). Also, the penalized estimator with Jeffrey's prior can be seen as an objective Bayes estimator as it is derived under the Jeffrey's prior, a well-known objective prior [21].

This research was partly motivated by a dataset on heart-disease [22], where the interest is in finding association between the occurrence of artery blockage and several clinical variables. A standard probit analysis of this data indicates a lack-of-fit at the 5% level of significance and that led us to consider the skew-probit model. However, for this skew-probit model, the maximum likelihood estimation (MLE) and Bayesian estimation with flexible prior on the regression parameters and the skewness parameter led to somewhat different parameter estimates. The estimate (95% CI) of δ under the MLE was 1.54(−0.35, 3.43), while the posterior means (95% credible interval) for δ were 0.18(−0.85, 1.59) and 0.54(0.02, 1.93) for the Bayesian method with uniform(−5, 5) prior and uniform(0, 5) prior on δ , respectively, and normal(0, 5²) prior on all regression parameters. These differences in the results indicate that an objective method is needed to

estimate the model parameters. In Section 5 we analyse this dataset using different methods and compare their results.

Before concluding this section we would like to highlight the novelties of this paper. To the best of our knowledge, this is the first paper that investigates parameter identifiability and the bias in the MLE of the binary model with the skew-probit link function. To reduce finite sample bias, we apply general bias reduction strategies to this particular problem, and compare and assess the effectiveness of the approaches through simulation studies.

2. Parameter identifiability

In general, model parameters are identifiable if the parameter values uniquely identify the underlying probability model. Now, following Rothenberg [23]'s general concept of identifiability, we present a formal definition of identifiability in our context of the skew-probit model.

Identifiability. The parameter set $\theta = (\beta^T, \delta)^T$ is said to be identifiable if $F(\mathbf{Z}^T \beta, \delta) = F(\mathbf{Z}^T \beta', \delta')$ for every \mathbf{Z} implies $(\beta', \delta') = (\beta, \delta)$. A parameter set θ is said to be locally identifiable if within a neighbourhood \mathcal{N} there does not exist a $(\beta', \delta') \in \mathcal{N} \setminus \{(\beta, \delta)\}$ such that $F(\mathbf{Z}^T \beta, \delta) = F(\mathbf{Z}^T \beta', \delta')$ for every \mathbf{Z} . A necessary and sufficient condition for local identifiability is the non-singularity of the Fisher information matrix [23]. Now, we investigate identifiability of three different cases, no covariate, a binary covariate and a continuous covariate, and these cases are presented in the following propositions. Based on these investigations, we conclude that the model parameters are not globally identifiable, however, for a continuous X and when $\delta \neq 0$ the parameters are identifiable.

Proposition 2.1: *In the absence of any covariate, the intercept β_0 and the skewness parameter δ are not identifiable in the skew-probit model $\text{pr}(Y = 1) = F(\beta_0, \delta) = \int_{-\infty}^{\beta_0} 2\phi(u)\Phi(\delta u)du$. In other words, for a given value of (β_0, δ) , there exist $\beta'_0 \neq \beta_0$ and $\delta' \neq \delta$ such that $F(\beta_0, \delta) = F(\beta'_0, \delta')$.*

Heuristic proof: In the absence of any covariate, the intercept β_0 and the skewness parameter δ are not identifiable in the skew-probit model $\text{pr}(Y = 1) = F(\beta_0, \delta) = \int_{-\infty}^{\beta_0} 2\phi(u)\Phi(\delta u)du$. In other words, for a given value of (β_0, δ) we can find another (β'_0, δ') such that $F(\beta_0, \delta) = F(\beta'_0, \delta')$. This fact is illustrated in Figure 1. This figure contains two CDFs for Skew-Normal($\mu = 0, \omega = 1, \delta$) and Skew-Normal($\mu = 0, \omega = 1, \delta'$) distributions. At the abscissa β_0 , the height of the dotted vertical line up to the CDF for the Skew-Normal($\mu = 0, \omega = 1, \delta$) distribution is $F(\beta_0, \delta)$. For the same value of the CDF, $F(\beta_0, \delta)$, there is another β'_0 and δ' , such that $F(\beta_0, \delta) = F(\beta'_0, \delta')$. Particularly, the abscissa of the point where the horizontal line at $F(\beta_0, \delta)$ hits the CDF for the Skew-Normal($\mu = 0, \omega = 1, \delta'$) distribution is β'_0 . This signifies that the CDF of the Skew-Normal($\mu = 0, \omega = 1, \delta'$) distribution at β'_0 is the same as $F(\beta_0, \delta)$. If ℓ stands for the log-likelihood, then analytical calculations show that $E(\partial^2 \ell / \partial \beta_0 \partial \beta_0) = -4\phi^2(\beta_0)\Phi^2(\beta_0\delta)/F(\beta_0, \delta)\{1 - F(\beta_0, \delta)\}$, $E(\partial^2 \ell / \partial \delta \partial \delta) = -\exp\{-\beta_0^2(1 + \delta^2)\}/\pi^2(1 + \delta^2)^2 F(\beta_0, \delta)\{1 - F(\beta_0, \delta)\}$, $E(\partial^2 \ell / \partial \beta_0 \partial \delta) = 2\phi(\beta_0)\Phi(\beta_0\delta) \exp\{-\beta_0^2(1 + \delta^2)/2\}/\pi(1 + \delta^2)F(\beta_0, \delta)\{1 - F(\beta_0, \delta)\}$, and the determinant of the Fisher information matrix $E(\partial^2 \ell / \partial \beta_0 \partial \beta_0)E(\partial^2 \ell / \partial \delta \partial \delta) - E^2(\partial^2 \ell / \partial \beta_0 \partial \delta) = 0$.

■

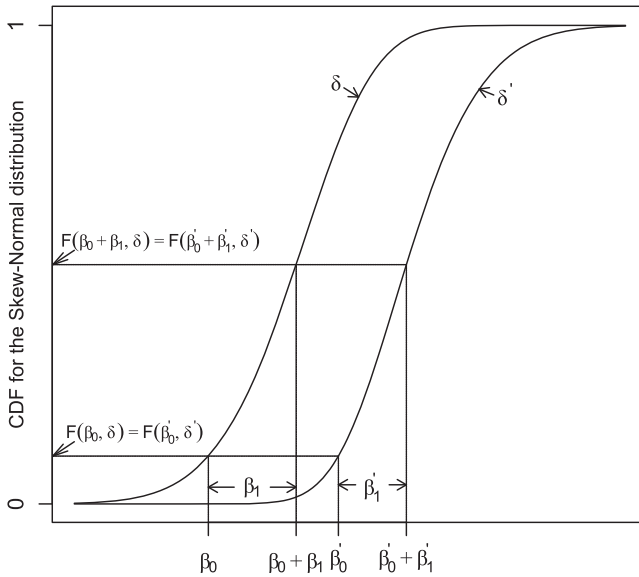


Figure 1. Illustration of parameter identifiability in the skew-probit model.

Analytical proof: Now, suppose that $(U_1, U_2)^T \sim \text{BVN}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (0, 0)^T$ and $\Sigma = \begin{pmatrix} 1 & -\lambda \\ -\lambda & 1 \end{pmatrix}$, where $\lambda = \lambda(\delta) = \delta/\sqrt{1 + \delta^2}$. Then following the marginal-conditional factorization approach [5,24] the CDF of the bivariate normal variable can be written as

$$\begin{aligned} 2\Phi_2((\eta, 0)^T; \boldsymbol{\mu}, \Sigma) &= 2 \int_{-\infty}^{\eta} \int_{-\infty}^0 f_{U_1}(u) f_{U_2|U_1}(v|u) dv du \\ &= 2 \int_{-\infty}^{\eta} \phi(u) \int_{-\infty}^0 \frac{1}{\sqrt{1 - \lambda^2}} \phi\left(\frac{v + \lambda u}{\sqrt{1 - \lambda^2}}\right) dv du \\ &= 2 \int_{-\infty}^{\eta} \phi(u) \Phi\left(\frac{\lambda u}{\sqrt{1 - \lambda^2}}\right) du = \int_{-\infty}^{\eta} 2\phi(u) \Phi(\delta u) du. \end{aligned} \quad (2)$$

Equation (2) shows the equivalence between the CDF of the bivariate normal variable and the skew-probit link. Because of this equivalence, when $\lambda = \delta/\sqrt{1 + \delta^2}$, the identifiability of $(\beta_0, \beta_1, \lambda)^T$ of the CDF of the bivariate normal variable is the same as the identifiability of $(\beta_0, \beta_1, \delta)^T$ of the skew-probit link function given in Equation (1).

In the no covariate case, for showing non-identifiability of $\eta = \beta_0$ and λ , we have to show that for a given (η, λ) where $\lambda \neq 0$, we can find $(\eta^*, \lambda^*) \neq (\eta, \lambda)$ such that

$$2 \int_{-\infty}^{\eta^*} \phi(u) \Phi\left(\frac{\lambda^* u}{\sqrt{1 - \lambda^{*2}}}\right) du = 2 \int_{-\infty}^{\eta} \phi(u) \Phi\left(\frac{\lambda u}{\sqrt{1 - \lambda^2}}\right) du. \quad (3)$$

Note that for an arbitrary choice of $\lambda^* \in (-1, 1)$, the left-hand side (LHS) of Equation (3) is a continuous and increasing function in η^* , and the LHS $\rightarrow 0$ and 1 as $\eta^* \rightarrow -\infty$ and ∞ , respectively. On the other hand, for given (η, λ) the right-hand side (RHS) of Equation (3)

is a known fraction r . Therefore, for an arbitrary choice of λ^* , Equation (3) yields a unique solution for η^* that depends on λ^* and r . Although it is difficult or impossible to get a closed form solution, η^* can be easily obtained numerically for an arbitrary choice of λ^* and known r . ■

Proposition 2.2: *In the presence of a binary covariate X , parameters $(\beta_0, \beta_1, \delta)$ of the skew-probit model $\text{pr}(Y = 1|X) = F(\beta_0 + \beta_1 X, \delta)$ are not identifiable. In other words, for a given value of $(\beta_0, \beta_1, \delta)$, there exist $(\beta'_0, \beta'_1, \delta') \neq (\beta_0, \beta_1, \delta)$ such that $F(\beta_0 + \beta_1 X, \delta) = F(\beta'_0 + \beta_1 X, \delta')$ for all $X = 0, 1$.*

Heuristic proof: Now suppose that there is a binary covariate X , and the model is $\text{pr}(Y = 1|X) = F(\beta_0 + \beta_1 X, \delta)$. If the parameter $(\beta_0, \beta_1, \delta)$ is non-identifiable, then we can find a $(\beta'_0, \beta'_1, \delta') \neq (\beta_0, \beta_1, \delta)$, such that $F(\beta_0 + \beta_1 X, \delta) = F(\beta'_0 + \beta_1 X, \delta')$ for every X . Now consider the two probabilities, $\text{pr}(Y = 1|X = 1) = F(\beta_0 + \beta_1, \delta)$ and $\text{pr}(Y = 1|X = 0) = F(\beta_0, \delta)$. From the discussion in the previous paragraph, we know that for a given (β_0, δ) we can find a $(\beta'_0, \delta') \neq (\beta_0, \delta)$ such that $F(\beta_0, \delta) = F(\beta'_0, \delta')$. Now, it turns out that given these two sets, (β_0, δ) and (β'_0, δ') , for every β_1 we can find a β'_1 , such that $F(\beta_0 + \beta_1, \delta) = F(\beta'_0 + \beta_1, \delta')$. In Figure 1, at the abscissa $(\beta_0 + \beta_1)$ the height of the dotted vertical line up to the CDF for the Skew-Normal($\mu = 0, \omega = 1, \delta$) distribution is $F(\beta_0 + \beta_1, \delta)$. Now, the abscissa of the intersection point of the horizontal line at $F(\beta_0 + \beta_1, \delta)$ with the CDF for the Skew-Normal($\mu = 0, \omega = 1, \delta'$) distribution is $\beta'_0 + \beta'_1$. That means, $F(\beta_0 + \beta_1, \delta) = F(\beta'_0 + \beta'_1, \delta')$. Hence, the model parameters are not identifiable. Using similar arguments we conclude that for a categorical covariate X , the model parameters of a skew-probit model are not identifiable. ■

Analytical proof: Due to relation (2), identifiability of $(\beta_0, \beta_1, \delta)$ and $(\beta_0, \beta_1, \lambda)$ are equivalent. We show that for a binary covariate X , $\beta_0, \beta_1, \lambda$ are not identifiable. Model parameters are not identifiable when for every given $(\beta_0, \beta_1, \lambda)$ we can find $(\beta_0^*, \beta_1^*, \lambda^*)$ such that

$$2 \int_{-\infty}^{\beta_0^* + \beta_1^* X} \phi(u) \Phi\left(\frac{\lambda^* u}{\sqrt{1 - \lambda^{*2}}}\right) du = 2 \int_{-\infty}^{\beta_0 + \beta_1 X} \phi(u) \Phi\left(\frac{\lambda u}{\sqrt{1 - \lambda^2}}\right) du \quad (4)$$

for every X . Using $X = 0$ and $X = 1$ in (4) we obtain

$$2 \int_{-\infty}^{\beta_0^*} \phi(u) \Phi\left(\frac{\lambda^* u}{\sqrt{1 - \lambda^{*2}}}\right) du = 2 \int_{-\infty}^{\beta_0} \phi(u) \Phi\left(\frac{\lambda u}{\sqrt{1 - \lambda^2}}\right) du \quad (5)$$

and

$$2 \int_{-\infty}^{\beta_0^* + \beta_1^*} \phi(u) \Phi\left(\frac{\lambda^* u}{\sqrt{1 - \lambda^{*2}}}\right) du = 2 \int_{-\infty}^{\beta_0 + \beta_1} \phi(u) \Phi\left(\frac{\lambda u}{\sqrt{1 - \lambda^2}}\right) du. \quad (6)$$

Like the analytical proof of Proposition 2.1, for given (β_0, λ) and an arbitrary choice of λ^* , Equation (5) yields a unique solution for β_0^* that is a function of (β_0, λ) and λ^* . Next consider Equation (6) whose RHS is a known fraction. After plugging in the solution of $\beta_0^* = \beta_0^*(\beta_0, \lambda, \lambda^*)$ into (6) we obtain a non-linear equation in β_1^* that is continuous, bounded and increasing. Importantly, the LHS of (6) converges to 0 or 1 as β_1^* converges to $-\infty$ or ∞ , respectively. Hence, (6) produces a unique solution for β_1^* for given $\beta_0, \beta_1, \lambda, \beta_0^*, \lambda^*$. ■

Proposition 2.3: *In the presence of a continuous covariate X , parameters $(\beta_0, \beta_1, \delta)$ of the skew-probit model $pr(Y = 1|X) = F(\beta_0 + \beta_1 X, \delta)$ are identifiable. In other words, for a given value of $(\beta_0, \beta_1, \delta)$, there does not exist another $(\beta'_0, \beta'_1, \delta') \neq (\beta_0, \beta_1, \delta)$ such that $F(\beta_0 + \beta_1 X, \delta) = F(\beta'_0 + \beta'_1 X, \delta')$ for all X .*

Proof: We assume that β_1 is non-zero, otherwise it will be the same as the case where there is no covariate. First, we show that for a fixed δ , (β_0, β_1) is identifiable. That is, $F(\beta_0 + \beta_1 X, \delta) = F(\beta'_0 + \beta'_1 X, \delta)$ for all X implies $(\beta_0, \beta_1) = (\beta'_0, \beta'_1)$. Note that $\int_{-\infty}^{\beta_0 + \beta_1 X} 2\phi(u)\Phi(\delta u)du = \int_{-\infty}^{\beta'_0 + \beta'_1 X} 2\phi(u)\Phi(\delta u)du$ implies $\beta_0 + \beta_1 X = \beta'_0 + \beta'_1 X$ for all X . Therefore, $\beta_0 = \beta'_0$ and $\beta_1 = \beta'_1$. Thus, if δ is fully specified, the remaining parameters are identifiable. However, identifiability of all three parameters $(\beta_0, \beta_1, \delta)$ is nontrivial and somewhat tricky, and this is what we consider next.

Suppose that $\theta = (\beta_0, \beta_1, \delta)^T$ involved in the skew-probit model is not identifiable. In the following discussion we shall be using the fact that for a fixed δ , $F(\cdot, \delta)$ is a strictly increasing function so that its inverse function, denoted by $F_\delta^{-1}(\cdot)$, exists. If the parameters are not identifiable, then there exists a $\theta' = (\beta'_0, \beta'_1, \delta') \neq \theta$, where $\delta' \neq \delta$ such that

$$F(\beta_0 + \beta_1 X, \delta) = F(\beta'_0 + \beta'_1 X, \delta') \text{ for all } X, \tag{7}$$

and particularly for $X = 0$, non-identifiability implies

$$F(\beta_0, \delta) = F(\beta'_0, \delta'). \tag{8}$$

Now, using the inverse operation on (8) and (7) we obtain

$$F_\delta^{-1}\{F(\beta_0, \delta)\} = \beta_0 = F_\delta^{-1}\{F(\beta'_0, \delta')\}, \tag{9}$$

$$\beta_0 + \beta_1 X = F_\delta^{-1}\{F(\beta'_0 + \beta'_1 X, \delta')\}. \tag{10}$$

Subtracting (9) from (10) we obtain

$$\beta_1 X = F_\delta^{-1}\{F(\beta'_0 + \beta'_1 X, \delta')\} - F_\delta^{-1}\{F(\beta'_0, \delta')\} \tag{11}$$

for $X \neq 0$. Differentiating both sides of Equation (11) with respect to X we get

$$\beta_1 = \frac{\phi(\beta'_0 + \beta'_1 X)\Phi\{\delta'(\beta'_0 + \beta'_1 X)\}\beta'_1}{\phi[F_\delta^{-1}\{F(\beta'_0 + \beta'_1 X, \delta')\}]\Phi[\delta F_\delta^{-1}\{F(\beta'_0 + \beta'_1 X, \delta')\}]}. \tag{12}$$

Since $\delta' \neq \delta$, $F_\delta^{-1}\{F(\beta'_0 + \beta'_1 X, \delta')\} \neq \beta'_0 + \beta'_1 X$ for all X , which means that the right-hand side of (12) is a non-linear function of X while the left-hand side is a constant. Therefore, our assumption that θ is not identifiable is wrong and it completes the proof.

To investigate this matter little further, we provide the following proof to show that for a continuous X and for any given $(\beta_0, \beta_1, \delta)$, where $\beta_1 \neq 0$ is fixed and $\delta \neq 0$, there does not exist another $(\beta'_0, \beta'_1, \delta')$ such that $F(\beta_0 + \beta_1 X, \delta) = F(\beta'_0 + \beta'_1 X, \delta')$ for all X .

Suppose that for a fixed $\beta_1 \neq 0$ and $\delta \neq 0$, there exist $\beta'_0 \neq \beta_0$ and $\delta' \neq \delta$ such that

$$F(\beta_0 + \beta_1 X, \delta) = F(\beta'_0 + \beta_1 X, \delta') \quad (13)$$

for all X . Taking derivative of (13) with respect to X , we obtain

$$\phi(\beta_0 + \beta_1 X)\Phi\{\delta(\beta_0 + \beta_1 X)\} = \phi(\beta'_0 + \beta_1 X)\Phi\{\delta'(\beta'_0 + \beta_1 X)\}. \quad (14)$$

Plugging in $X = -\beta_0/\beta_1$ and $-\beta'_0/\beta_1$ in (14), we obtain

$$\begin{aligned} \phi(0)\Phi(0) &= \phi(\beta'_0 - \beta_0)\Phi\{\delta'(\beta'_0 - \beta_0)\}, \quad \text{and} \\ \phi(0)\Phi(0) &= \phi(\beta_0 - \beta'_0)\Phi\{\delta(\beta_0 - \beta'_0)\}. \end{aligned} \quad (15)$$

Relations in (15) together imply $\Phi\{\delta'(\beta'_0 - \beta_0)\} = \Phi\{\delta(\beta_0 - \beta'_0)\}$ which implies

$$\delta(\beta_0 - \beta'_0) = \delta'(\beta'_0 - \beta_0). \quad (16)$$

Equation (16) holds if either 1) $\beta'_0 = \beta_0$ or 2) $\delta' = -\delta$. Suppose that 1) $\beta'_0 = \beta_0$ holds. Then from (14), we obtain $\phi(\beta_0)\Phi(\delta\beta_0) = \phi(\beta_0)\Phi(\delta'\beta_0)$ when $X = 0$, this implies $\delta = \delta'$. Hence $(\beta'_0, \beta_1, \delta') = (\beta_0, \beta_1, \delta)$. Next, suppose that 2) $\delta' = -\delta$ holds. Then from (14), we obtain

$$\frac{\phi(\beta_0 + \beta_1 X)}{\phi(\beta'_0 + \beta_1 X)} = \frac{\Phi\{-\delta(\beta'_0 + \beta_1 X)\}}{\Phi\{\delta(\beta_0 + \beta_1 X)\}} \quad (17)$$

for all X . Without loss of generality assume that $\delta > 0$. Define $\mathcal{X} = \{X : X > \max(-\beta_0/\beta_1, -\beta'_0/\beta_1)\}$. Now, for any $X \in \mathcal{X}$ the right-hand side of (17) lies in $(0, 1)$ while the left-hand side of (17) is not guaranteed to lie in $(0, 1)$. Hence we obtain a contradiction. \blacksquare

3. Bias reduction

3.1. Maximum likelihood and bootstrap

Suppose that the observed data $\mathbf{D} = (D_1, \dots, D_n)$ with $D_i = (Y_i, \mathbf{X}_i)$, $i = 1, \dots, n$ are collected from n subjects that are randomly drawn from the underlying population. At least one component of the covariate vector is assumed to be continuous. We want to fit the regression model 1 to the data. The logarithm of the likelihood is

$$\ell = \sum_{i=1}^n Y_i \log\{F(\eta_i, \delta)\} + (1 - Y_i) \log\{1 - F(\eta_i, \delta)\},$$

where $\eta_i = \mathbf{Z}_i^T \boldsymbol{\beta}$ and $\mathbf{Z}_i = (1, \mathbf{X}_i^T)^T$. The maximum likelihood estimators (MLE) of $\boldsymbol{\beta}$ and δ are obtained by solving $\partial \ell / \partial \boldsymbol{\theta} = (\partial \ell / \partial \boldsymbol{\beta}^T, \partial \ell / \partial \delta)^T = \mathbf{0}$, where

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\beta}} &= 2 \sum_{i=1}^n \left\{ \frac{Y_i}{F(\eta_i, \delta)} - \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)} \right\} \phi(\eta_i) \Phi(\delta \eta_i) \mathbf{Z}_i, \\ \frac{\partial \ell}{\partial \delta} &= \sum_{i=1}^n \left\{ -\frac{Y_i}{F(\eta_i, \delta)} + \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)} \right\} \frac{\exp\{-\eta_i^2(1 + \delta^2)/2\}}{\pi(1 + \delta^2)}. \end{aligned}$$

In principle, the parameter estimates can be obtained by solving the above equations using the scoring method. Let $\boldsymbol{\theta}^{(t)}$ be the parameter value at the t th iteration of the scoring

method. Then at the $(t + 1)$ th iteration we obtain

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathcal{I}^{-1}(\boldsymbol{\theta}^{(t)}) \left(\frac{\partial \ell}{\partial \boldsymbol{\theta}} \right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}},$$

where the information matrix $\mathcal{I}(\boldsymbol{\theta}) = -E(\partial^2 \ell / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T)$ with

$$\begin{aligned} E \left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \boldsymbol{\beta}^T} \right) &= -4 \sum_{i=1}^n \frac{\phi^2(\eta_i) \Phi^2(\delta \eta_i)}{F(\eta_i, \delta) \{1 - F(\eta_i, \delta)\}} \mathbf{Z}_i \mathbf{Z}_i^T, \\ E \left(\frac{\partial^2 \ell}{\partial \delta^2} \right) &= - \sum_{i=1}^n \frac{\exp\{-\eta_i^2(1 + \delta^2)\}}{\pi^2(1 + \delta^2)^2 F(\eta_i, \delta) \{1 - F(\eta_i, \delta)\}}, \\ E \left(\frac{\partial^2 \ell}{\partial \delta \partial \boldsymbol{\beta}} \right) &= 2 \sum_{i=1}^n \frac{\phi(\eta_i) \Phi(\delta \eta_i) \exp\{-\eta_i^2(1 + \delta^2)/2\}}{\pi(1 + \delta^2) F(\eta_i, \delta) \{1 - F(\eta_i, \delta)\}} \mathbf{Z}_i. \end{aligned}$$

We note that the information matrix can be written as $\mathcal{I}(\boldsymbol{\theta}) = \mathbf{W}(\boldsymbol{\theta})^T \mathbf{A}(\boldsymbol{\theta}) \mathbf{W}(\boldsymbol{\theta})$, where $\mathbf{A}(\boldsymbol{\theta}) = \text{diag}[F(\eta_i, \delta) \{1 - F(\eta_i, \delta)\}]^{-1}$, $\mathbf{W}(\boldsymbol{\theta})^T = [\mathbf{W}_1(\boldsymbol{\theta}), \dots, \mathbf{W}_n(\boldsymbol{\theta})]$, $\mathbf{W}_i^T(\boldsymbol{\theta}) = [2\phi(\eta_i) \Phi(\delta \eta_i) \mathbf{Z}_i^T, -\exp\{\eta_i = \mathbf{Z}_i^T \boldsymbol{\beta} \mathbf{W}_i^T(\boldsymbol{\theta}) \mathbf{Z}_i \mathbf{W}(\boldsymbol{\theta}) \boldsymbol{\theta}^{(t+1)} = \{\mathbf{W}(\boldsymbol{\theta}^{(t)})^T \mathbf{A}(\boldsymbol{\theta}^{(t)}) \mathbf{W}(\boldsymbol{\theta}^{(t)})\}^{-1} \mathbf{W}(\boldsymbol{\theta}^{(t)})^T \mathbf{A}(\boldsymbol{\theta}^{(t)}) \mathbf{Y}^*(\boldsymbol{\theta}^{(t)}) \mathbf{Y}^*(\boldsymbol{\theta}^{(t)}) = \mathbf{W}(\boldsymbol{\theta}^{(t)}) \boldsymbol{\theta}^{(t)} + \{\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}^{(t)})\}]$, $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\boldsymbol{\mu}(\boldsymbol{\theta}^{(t)}) = (F(\eta_1^{(t)}, \delta^{(t)}), \dots, F(\eta_n^{(t)}, \delta^{(t)}))^T$, $\eta_i^{(t)} = \mathbf{Z}_i^T \boldsymbol{\beta}^{(t)}$.

For larger values of δ , the curvature $E(-\partial^2 \ell / \partial \delta^2)$ tends to be small, resulting in highly biased MLE of δ . Additionally, if there is no covariate, and the model for Y is $\text{pr}(Y = 1) = 2 \int_{-\infty}^0 \phi(u) \Phi(\delta u) du$ that involves with only one parameter δ , the probability that the MLE of δ diverges to $+\infty$ or $-\infty$ is $p_n(\delta) = \text{pr}(Y_1 = \dots = Y_n = 0) + \text{pr}(Y_1 = \dots = Y_n = 1) = \{\pi + 2 \tan^{-1}(\delta)/(2\pi)\}^n + \{\pi - 2 \tan^{-1}(\delta)/(2\pi)\}^n$. Although this probability goes to zero as $n \rightarrow \infty$, this may not be negligible for a moderate value of n . This $p_n(\delta)$ is also the probability of diverging MLE of δ when a continuous response follows skew-normal $(\mu = 0, \omega = 1, \delta)$ [15].

In order to reduce the finite sample bias of the MLE that is of the order $O(n^{-1})$, we consider the following strategies. First, we apply the bootstrap method to reduce the bias of the MLE. Suppose that $b(\widehat{\boldsymbol{\theta}}_{MLE})$ denotes the bias of $\widehat{\boldsymbol{\theta}}_{MLE}$, the MLE of $\boldsymbol{\theta}$. Based on B bootstrap samples, we estimate $b(\widehat{\boldsymbol{\theta}}_{MLE})$, and denote this estimator of bias by $\widehat{b}_{\text{boot}}(\widehat{\boldsymbol{\theta}}_{MLE})$. The bias corrected estimator is then defined as $\widehat{\boldsymbol{\theta}}_{MLE} - \widehat{b}_{\text{boot}}(\widehat{\boldsymbol{\theta}}_{MLE})$. This approach is referred to as method B.

3.2. Penalized maximum likelihood

Next, we propose to estimate the parameters by maximizing a penalized likelihood,

$$\ell_p = \ell + M(\boldsymbol{\theta}),$$

where $M(\boldsymbol{\theta})$ is the penalty function. The estimator obtained by maximizing ℓ_p can be seen as the posterior mode where the prior distribution $\pi(\boldsymbol{\theta}) \propto \exp\{M(\boldsymbol{\theta})\}$. Unlike the other bias correction approaches that require the estimator to be finite, this approach does not require the MLE to be finite. Rather penalization helps to add a curvature in a otherwise flat likelihood surface, and thereby the penalized likelihood method prevents the estimate

to be infinite or unrealistically large and also reduces finite sample bias. Following the general strategy of Firth [13], we replace $M(\boldsymbol{\theta})$ by $0.5\log[\det\{\mathcal{I}(\boldsymbol{\theta})\}]$, where \det stands for matrix determinant. Thus, the maximum penalized likelihood estimator, denoted by $\widehat{\boldsymbol{\theta}}_{pj}$, is obtained by solving

$$\begin{aligned}\frac{\partial \ell_p}{\partial \boldsymbol{\beta}} &= 2 \sum_{i=1}^n \left\{ \frac{Y_i}{F(\eta_i, \delta)} - \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)} \right\} \phi(\eta_i) \Phi(\delta \eta_i) \mathbf{Z}_i + \frac{1}{2} \text{trace} \left\{ \mathcal{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathcal{I}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right\} = \mathbf{0}, \\ \frac{\partial \ell_p}{\partial \delta} &= \sum_{i=1}^n \left\{ -\frac{Y_i}{F(\eta_i, \delta)} + \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)} \right\} \frac{\exp\{-\eta_i^2(1 + \delta^2)/2\}}{\pi(1 + \delta^2)} \\ &\quad + \frac{1}{2} \text{trace} \left\{ \mathcal{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathcal{I}(\boldsymbol{\theta})}{\partial \delta} \right\} = 0.\end{aligned}$$

This approach is referred to as method J. This estimator can be seen as the posterior mode when the Jeffrey's prior is used on the parameters as $e^{M(\boldsymbol{\theta})} = \det\{\mathcal{I}(\boldsymbol{\theta})\}^{1/2}$. Although this approach of bias reduction has been extensively used in various contexts including when a continuous response follows skew-normal ($\mu = 0, \omega = 1, \delta$) [15], the approach has never been applied to the case where the binary response variable Y is modelled via the skew-probit link.

Next, we consider a generalization of the Jeffrey's prior [19], where the prior $\pi_{GI}(\boldsymbol{\theta}) \propto |\det\{\mathcal{I}(\boldsymbol{\theta})\}|^{1/2} \exp\{-(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathcal{I}(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\theta}_0)/2c_0\}$. For a large c_0 , $\pi_{GI}(\boldsymbol{\theta})$ converges $|\det\{\mathcal{I}(\boldsymbol{\theta})\}|^{1/2}$, that is Jeffrey's prior. Gupta and Ibrahim [19] showed that under a logistic model, π_{GI} has lower mass around the centre and heavier tail than the normal distribution resulting in a relatively non-informative prior. Adopting their prior distribution with $c_0 = 1$ and $\boldsymbol{\theta}_0 = \mathbf{1}$, and setting $M(\boldsymbol{\theta}) = \log\{\pi_{GI}(\boldsymbol{\theta})\}$ in our penalized likelihood ℓ_p , we obtain the following estimating equations to estimate $(\boldsymbol{\beta}^T, \delta)^T$:

$$\begin{aligned}\frac{\partial \ell_p}{\partial \boldsymbol{\beta}} &= 2 \sum_{i=1}^n \left\{ \frac{Y_i}{F(\eta_i, \delta)} - \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)} \right\} \phi(\eta_i) \Phi(\delta \eta_i) \mathbf{Z}_i + \frac{1}{2} \text{trace} \left\{ \mathcal{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathcal{I}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right\} \\ &\quad - \frac{1}{2} \frac{\partial \boldsymbol{\theta}^T \mathcal{I}(\boldsymbol{\theta}) \boldsymbol{\theta}}{\partial \boldsymbol{\beta}} = \mathbf{0}, \\ \frac{\partial \ell_p}{\partial \delta} &= \sum_{i=1}^n \left\{ -\frac{Y_i}{F(\eta_i, \delta)} + \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)} \right\} \frac{\exp\{-\eta_i^2(1 + \delta^2)/2\}}{\pi(1 + \delta^2)} + \frac{1}{2} \text{trace} \left\{ \mathcal{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathcal{I}(\boldsymbol{\theta})}{\partial \delta} \right\} \\ &\quad - \frac{1}{2} \frac{\partial \boldsymbol{\theta}^T \mathcal{I}(\boldsymbol{\theta}) \boldsymbol{\theta}}{\partial \delta} = 0.\end{aligned}$$

This method is referred to as method G.

Gelman et al. [20] pointed out use of Jeffrey's prior distribution might produce unreliable computation and be difficult to interpret in the Bayesian context. To avoid these potential issues, they proposed weakly informative Cauchy distribution prior for estimating logistic model parameters which results in stable and regularized estimates. Adopting their recommendation in our setup we consider $e^{M(\boldsymbol{\theta})} = \prod_k \{\pi(1 + \theta_k^2/2.5^2)\}^{-1}$, i.e. $M(\boldsymbol{\theta}) = -\sum_k \log(1 + \theta_k^2/2.5^2)$. This implies independent Cauchy(0, 2.5) prior for each

component of θ . Corresponding estimators are obtained by solving

$$\begin{aligned} \frac{\partial \ell_p}{\partial \beta} &= 2 \sum_{i=1}^n \left\{ \frac{Y_i}{F(\eta_i, \delta)} - \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)} \right\} \phi(\eta_i) \Phi(\delta \eta_i) \mathbf{Z}_i \\ &\quad - \mathbf{1}^T \text{Diag} \left(\frac{2\beta_0}{2.5^2 + \beta_0^2}, \frac{2\beta_1}{2.5^2 + \beta_1^2}, \dots, \frac{2\beta_q}{2.5^2 + \beta_q^2} \right) = \mathbf{0}, \\ \frac{\partial \ell_p}{\partial \delta} &= \sum_{i=1}^n \left\{ -\frac{Y_i}{F(\eta_i, \delta)} + \frac{(1 - Y_i)}{1 - F(\eta_i, \delta)} \right\} \frac{\exp\{-\eta_i^2(1 + \delta^2)/2\}}{\pi(1 + \delta^2)} - \frac{2\delta}{2.5^2 + \delta^2} = 0, \end{aligned}$$

where q is the number of covariates. This approach is referred to as method C.

Penalized estimators are obtained by solving the modified score equation $(\partial \ell_p / \partial \theta) = \mathbf{0}$. Note that the penalty function $M(\theta)$ is a $O_p(1)$ order term while the log-likelihood ℓ is $O_p(n)$ order term. Therefore, the asymptotic standard error calculation using the Fisher information matrix is still valid. That is, under certain regularity conditions, we may apply the standard likelihood theory to test hypotheses regarding parameters. However, the expected information matrix is singular when $\delta = 0$. Therefore, one can use Fisher’s information matrix for statistical inference as long as $\delta \neq 0$. Alternative parametrization (Pearson’s skewness parameter) is required to handle $\delta = 0$ case [25, p.66].

We consider two other penalized estimators. First, where the Jeffrey’s prior for δ is constructed assuming $\beta_0 = 0$ and $\beta_1 = \mathbf{0}$, and the logarithm of the prior density is used as the penalty function $M(\theta)$. Second, we take $M(\theta)$ to be the logarithm of the density function of the t distribution with degrees of freedom 2, location 0 and scale parameter 0.5 on the skewness parameter δ . This t density for δ arises due to a non-informative prior on κ when a standard skew-normal variable U with the skewness parameter δ is expressed as $U = \sqrt{1 - \kappa^2}Z + \kappa Z^*$, with $Z \sim \text{Normal}(0, 1)$, and Z^* follows a half-normal density with the density function $f(Z^*) = 2(2\pi)^{-1/2} \exp\{-(Z^*)^2/2\}, Z^* > 0$ [26]. However, in our initial numerical studies the performance of these penalized estimators is much worse than the other penalized estimators, so we have omitted them from further consideration.

4. Simulation study

Design: We simulated datasets of different sizes, $n = 200, 500, 1000, 2000$ and 5000 . We considered cases with a scalar covariate X_1 (scenarios 1–12) and cases with multiple covariates, X_1 and X_2 (scenarios 13–16), and for each dataset the response Y was a binary variable. We set $\beta_1 = 1, \beta_2 = -0.7, \delta = 4$ or 8 and defined $p_m = \text{pr}(Y = 1) = \int \text{pr}(Y = 1|x)g(x)dx$ as the marginal success probability. Depending on δ, p_m and the distribution of X_1 and X_2, β_0 was determined. Given X_1 (or X_1 and X_2), Y was generated using the Bernoulli distribution with success probability $\text{pr}(Y = 1|X) = F(\beta_0 + \beta_1 X_1, \delta)$ (or $F(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \delta)$). The simulation scenarios are listed below while a summary of the simulation design is presented in Table 1. In all scenarios, the variance of each covariate X_1, X_2 remains the same, and we considered small (12%) and moderate (40%) values for p_m . Importantly, we have used three different distributions for the covariate, uniform (U), normal (N), and a two-component mixture of normals. For the scalar and multiple

Table 1. A summary of the simulation designs. Here $p_m = E\{\text{pr}(Y = 1|X)\}$ denotes the marginal success probability of Y .

Scenario	X_1	X_2	δ	β_0	p_m
1	Uniform(-2, 2)	NA	4	-0.87	12%
2				0.37	40%
3	Normal(0, $(\sqrt{4/3})^2$)	NA	8	-0.85	12%
4				0.38	40%
5			4	-0.77	12%
6				0.42	40%
7	8	-0.73	12%		
8		0.44	40%		
9	0.5Normal(-1, $(\sqrt{1/3})^2$)	NA	4	-0.85	12%
10				0.35	40%
11	+0.5Normal(1, $(\sqrt{1/3})^2$)	NA	8	-0.82	12%
12				0.37	40%
13	Uniform(-2, 2)	Normal(0, $(\sqrt{4/3})^2$)	4	-1.09	12%
14				0.34	40%
15	NA	NA	8	-1.07	12%
16				0.36	40%

covariate cases we estimated $\theta = (\beta_0, \beta_1, \delta)^T$ and $\theta = (\beta_0, \beta_1, \beta_2, \delta)^T$, respectively, using the five methods (N, B, J, G, C) discussed in the previous section.

- (S1) $X_1 \sim U(-2, 2), \beta_1 = 1, \delta = 4, \beta_0 = -0.87, p_m = 12\%$;
- (S2) $X_1 \sim U(-2, 2), \beta_1 = 1, \delta = 4, \beta_0 = 0.37, p_m = 40\%$;
- (S3) $X_1 \sim U(-2, 2), \beta_1 = 1, \delta = 8, \beta_0 = -0.85, p_m = 12\%$;
- (S4) $X_1 \sim U(-2, 2), \beta_1 = 1, \delta = 8, \beta_0 = 0.38, p_m = 40\%$;
- (S5) $X_1 \sim N(0, (\sqrt{4/3})^2), \beta_1 = 1, \delta = 4, \beta_0 = -0.77, p_m = 12\%$;
- (S6) $X_1 \sim N(0, (\sqrt{4/3})^2), \beta_1 = 1, \delta = 4, \beta_0 = 0.42, p_m = 40\%$;
- (S7) $X_1 \sim N(0, (\sqrt{4/3})^2), \beta_1 = 1, \delta = 8, \beta_0 = -0.73, p_m = 12\%$;
- (S8) $X_1 \sim N(0, (\sqrt{4/3})^2), \beta_1 = 1, \delta = 8, \beta_0 = 0.44, p_m = 40\%$;
- (S9) $X_1 \sim 0.5N(-1, (\sqrt{1/3})^2) + 0.5N(1, (\sqrt{1/3})^2), \beta_1 = 1, \delta = 4, \beta_0 = -0.85, p_m = 12\%$;
- (S10) $X_1 \sim 0.5N(-1, (\sqrt{1/3})^2) + 0.5N(1, (\sqrt{1/3})^2), \beta_1 = 1, \delta = 4, \beta_0 = 0.35, p_m = 40\%$;
- (S11) $X_1 \sim 0.5N(-1, (\sqrt{1/3})^2) + 0.5N(1, (\sqrt{1/3})^2), \beta_1 = 1, \delta = 8, \beta_0 = -0.82, p_m = 12\%$;
- (S12) $X_1 \sim 0.5N(-1, (\sqrt{1/3})^2) + 0.5N(1, (\sqrt{1/3})^2), \beta_1 = 1, \delta = 8, \beta_0 = 0.37, p_m = 40\%$;
- (S13) $X_1 \sim U(-2, 2), X_2 \sim N(0, (\sqrt{4/3})^2), \beta_1 = 1, \beta_2 = -0.7, \delta = 4, \beta_0 = -1.09, p_m = 12\%$;
- (S14) $X_1 \sim U(-2, 2), X_2 \sim N(0, (\sqrt{4/3})^2), \beta_1 = 1, \beta_2 = -0.7, \delta = 4, \beta_0 = 0.34, p_m = 40\%$;
- (S15) $X_1 \sim U(-2, 2), X_2 \sim N(0, (\sqrt{4/3})^2), \beta_1 = 1, \beta_2 = -0.7, \delta = 8, \beta_0 = -1.07, p_m = 12\%$;
- (S16) $X_1 \sim U(-2, 2), X_2 \sim N(0, (\sqrt{4/3})^2), \beta_1 = 1, \beta_2 = -0.7, \delta = 8, \beta_0 = 0.36, p_m = 40\%$;

Implementation: For all penalized methods we set the initial value of β -parameter to the probit regression estimates while the initial value of δ was set to 3. We also experimented by setting the initial value of δ to random numbers generated from uniform(0, 10), however, the parameter estimates do not change with the initial value as long as the initial value of δ has the same sign as of the true δ . However, as the sample size increases, the sign of the initial value seems to have less impact on the final estimate. To obtain estimates for all five methods, we used the R function `ucminf` in package `ucminf` that makes use of the quasi-Newton optimization with BFGS updating for the inverse Hessian matrix [27]. There was no convergence issue in using this optimization method. However, one may face convergence issues in using the `nlm` function or the L-BFGS-B algorithm used in `optim`, and a detailed comparison of different optimization algorithms is given in Section S.2 of the supplementary materials. Particularly, in Table S.8 of the supplementary materials we compare different algorithms in terms of the number of datasets associated with non-convergent estimates. All simulations were conducted in a supermicro server with 28 core Intel Xeon CPU E5-2680 v4 @ 2.40 GHz and 64 GB of 2400 MHz DDR4 RAM.

Results: We present boxplots of estimates for each parameter ($\beta_0 \equiv$ intercept, slope(s), $\delta \equiv$ skewness) with the empirical coverage probability for the 95% nominal level of significance. We note that the scales of the y -axis might be different so that a direct comparison needs to be done with caution. All results are based on 1000 replications. The empirical coverage probability was calculated using Wald-type confidence intervals, where the standard errors were calculated by inverting the Fisher information matrix. For the bootstrap approach (method B), we have used 200 bootstrap samples. Performances of estimates for scenarios 1–4 are presented in Figures 2–5, respectively. In addition, mean and standard deviation of the computation time for scenarios 1–4 are provided in Table 2. Corresponding figures and tables for scenarios 5–12 are given in Section S.1 of the supplementary materials. We observed that the resulting figures and the pattern of computation times are similar across the scenarios we consider.

Regarding the intercept and slope estimator, and considering all sample sizes, the performance of method J is the best among these five approaches in terms of bias, variability and coverage probability, and its bias and variability decrease with the sample size.

The performance of method G is poor as its bias does not decrease with the sample size. The bias, variability and coverage probability of methods B and C are poor when the sample size is small, while they get better as the sample size increases. Particularly, for $n = 5000$, method C performs as well as method J. To save space we present the bias and standard deviation of estimators only for scenario 6 in the supplementary materials (See Tables S.4–S.6). A similar comparative performance of bias and standard deviation is observed for other scenarios.

For the skewness parameter (δ) estimation, again method J outperforms all methods across all the scenarios. Under small sample sizes, boxplots corresponding to method N do not fit in the extended y -axis scale. The bias of method C is larger than that of method J for small sample sizes, but they become closer for larger sample sizes. On the contrary, methods B and G seem unreliable for δ estimation.

Under the two most promising methods, J and C, coverage probabilities for the regression parameters get close to the nominal level for moderate sample size ≥ 500 . However, for the skewness parameter, the coverage probability converges to the nominal level at a slower



Figure 2. Simulation results based on 1000 replications when $X \sim \text{Uniform}(-2, 2)$, $\delta = 4$, $\beta_0 = -0.87$, $\beta_1 = 1$ and $p_m = 12\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey’s prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

rate than that of the regression parameters. This rate seems to be faster for $\delta = 4$ than $\delta = 8$ across different sample sizes. Overall as the sample size increases, the performance of methods N, J and C improves, and method C tends to perform as well as method J.

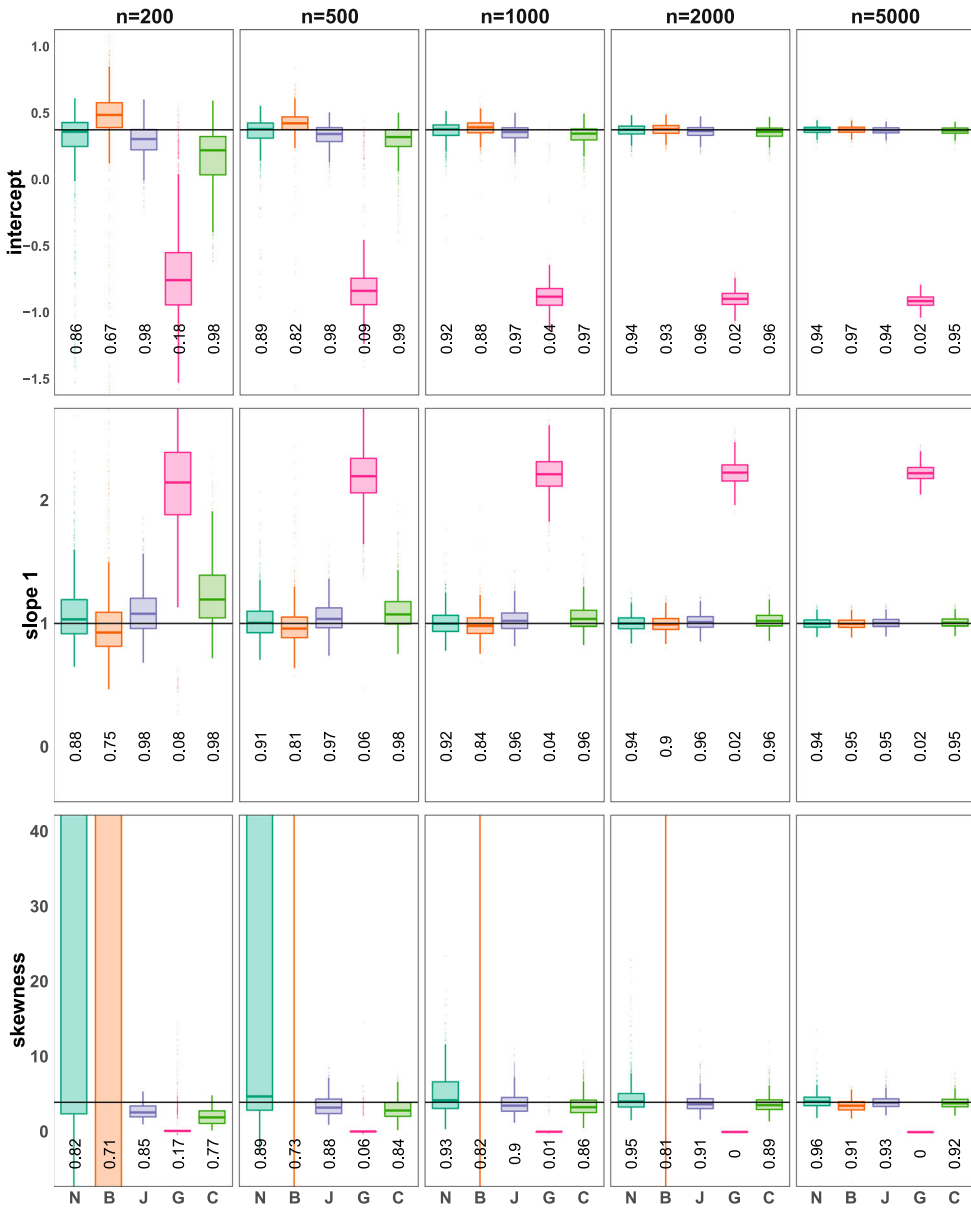


Figure 3. Simulation results based on 1000 replications when $X \sim \text{Uniform}(-2, 2)$, $\delta = 4$, $\beta_0 = 0.37$, $\beta_1 = 1$ and $p_m = 40\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey's prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

Following referees' suggestions, in Figures 6–9 we present the simulation results for scenarios 13–16 involving two covariates. Based on the operating characteristics of the estimators, as before, method J is superior than any other method.

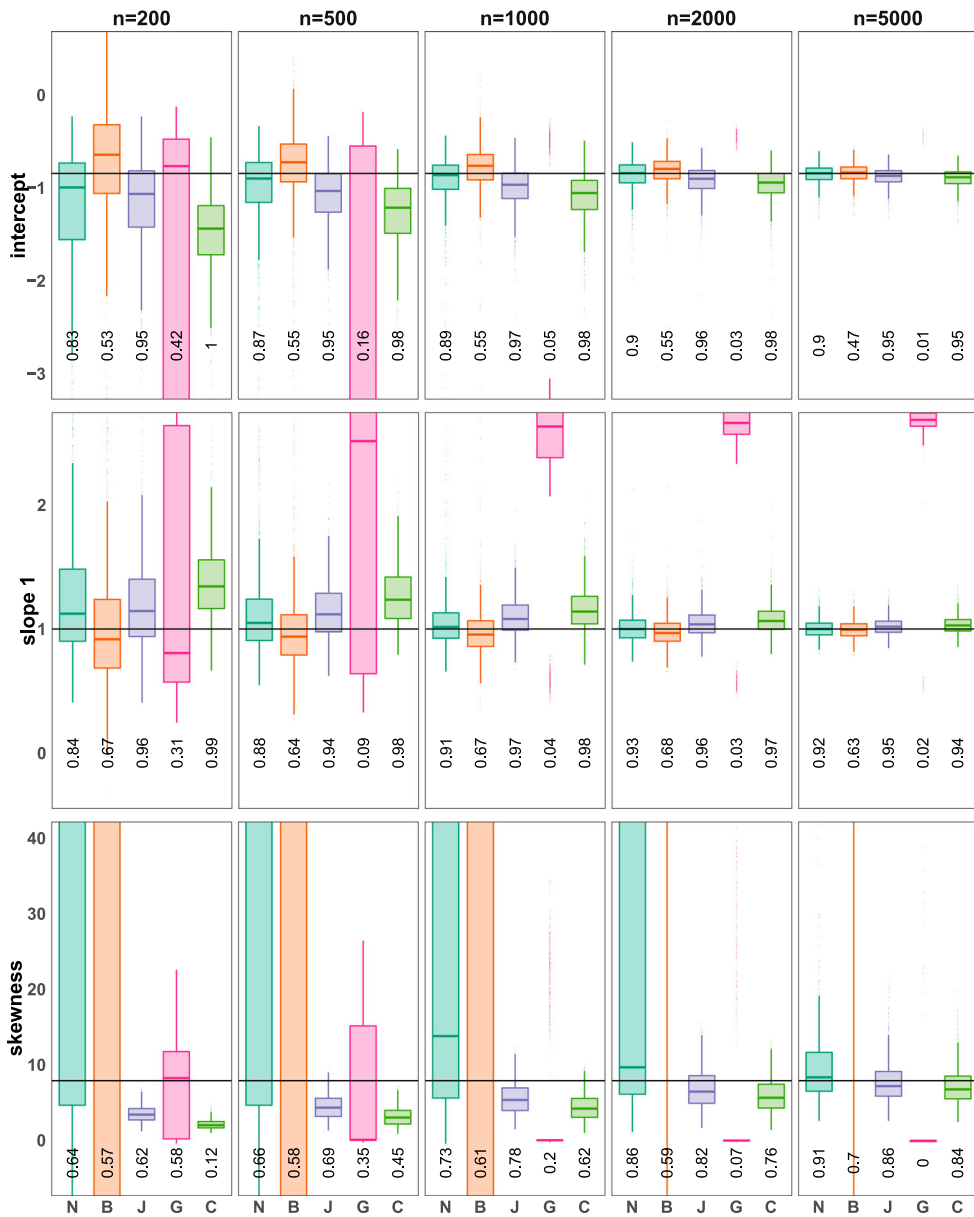


Figure 4. Simulation results based on 1000 replications when $X \sim \text{Uniform}(-2, 2)$, $\delta = 8$, $\beta_0 = -0.85$, $\beta_1 = 1$ and $p_m = 12\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey's prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

With respect to the computation time, except some handful of scenarios method C is the fastest across the scenarios (Table 2). The computation time of the best performing method J is almost twice the computation of the fastest method C. Of course, method J is faster than methods B and G, and it is usually faster than method N for $n < 5000$.

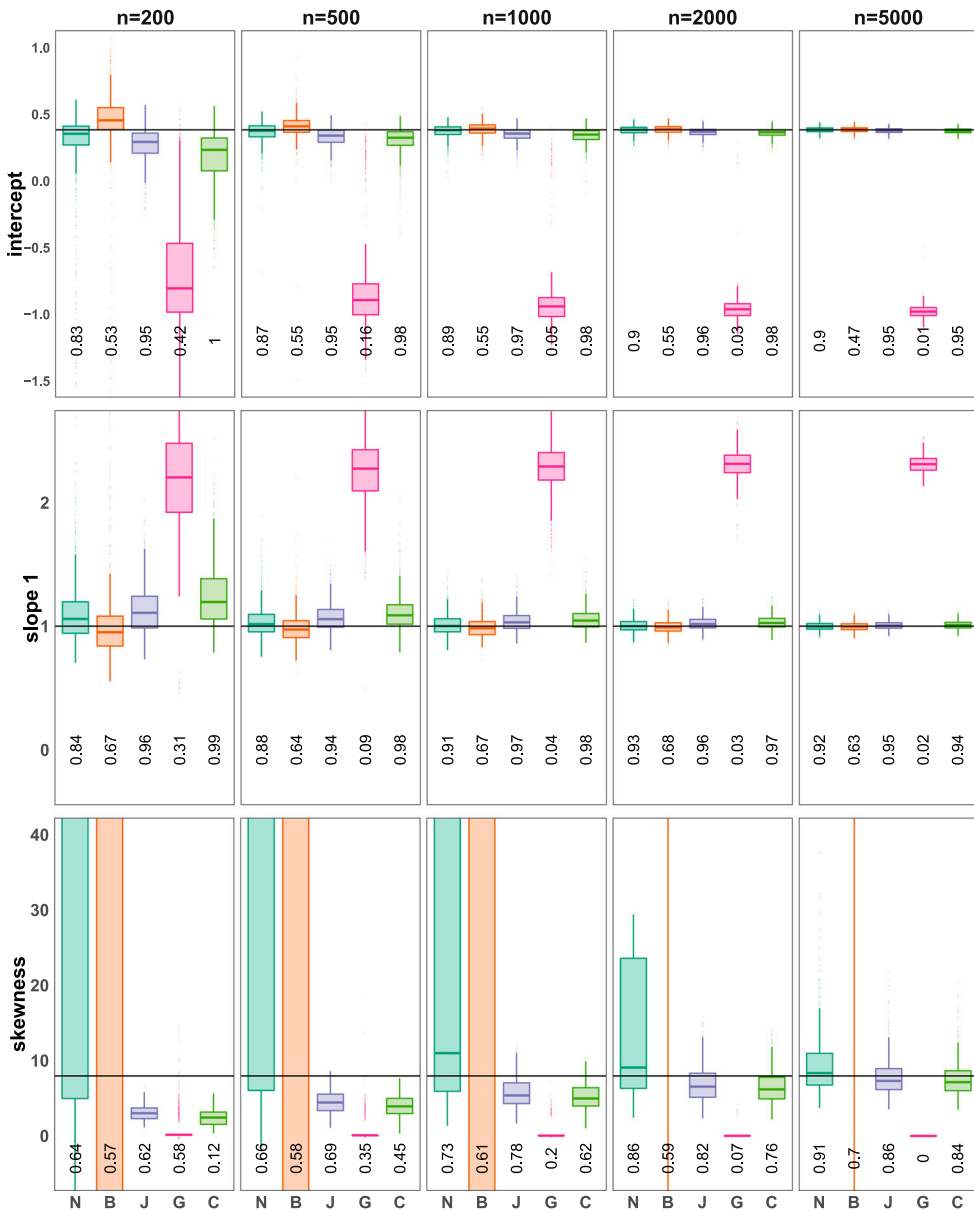


Figure 5. Simulation results based on 1000 replications when $X \sim \text{Uniform}(-2, 2)$, $\delta = 8$, $\beta_0 = 0.38$, $\beta_1 = 1$ and $p_m = 40\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey’s prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

In summary we can make the following conclusions. The maximum likelihood estimator has a skewed distribution, especially for small to moderate sample sizes. In general, the bootstrap bias corrected MLE (method B) does not show any better performance than method N. Rather, in some cases method B was worse than method N. From our simulation

Table 2. This table contains the mean (standard deviation) of the computation time in seconds for simulation scenarios 1–4 of Section 4.

Scenario	n	Method				
		N	B	J	G	C
1	200	4.021	802.474	3.236	3.189	1.346
		(1.992)	(164.685)	(0.780)	(1.418)	(0.319)
	500	8.373	1801.018	8.856	6.753	4.095
		(4.850)	(532.554)	(1.958)	(3.628)	(0.766)
	1000	13.338	3060.539	18.272	12.123	8.652
	(7.766)	(1015.910)	(3.510)	(6.782)	(1.488)	
2	200	22.918	5195.306	39.436	22.026	18.951
		(9.966)	(1621.352)	(6.580)	(11.791)	(3.199)
	500	52.876	10808.723	101.636	45.616	50.374
		(6.457)	(1403.345)	(14.585)	(18.034)	(7.529)
	1000	2.875	630.499	1.541	2.125	0.721
	(1.949)	(173.924)	(0.695)	(0.626)	(0.322)	
3	200	5.459	1263.778	4.861	5.088	2.280
		(4.376)	(551.474)	(1.983)	(1.157)	(0.912)
	500	8.089	1951.812	10.744	9.938	5.160
		(6.320)	(992.649)	(3.787)	(1.845)	(1.850)
	1000	13.119	2934.751	23.197	19.926	11.401
	(6.712)	(1208.295)	(6.597)	(3.440)	(3.397)	
4	200	32.080	6313.242	62.780	48.839	31.165
		(5.119)	(890.488)	(12.567)	(8.993)	(6.358)
	500	4.364	847.642	3.189	3.283	1.413
		(1.842)	(144.126)	(0.742)	(1.334)	(0.283)
	1000	10.608	2107.168	9.224	7.115	4.324
(4.982)		(500.728)	(1.753)	(3.601)	(0.797)	
2000	18.903	3972.760	18.562	13.055	8.985	
	(9.967)	(1118.155)	(3.269)	(7.124)	(1.589)	
5	200	32.304	7283.712	40.877	22.564	19.873
		(17.859)	(2318.334)	(6.666)	(12.602)	(3.150)
	500	64.590	14428.782	110.381	46.191	55.077
		(26.485)	(4144.437)	(18.285)	(16.888)	(7.924)
	200	3.545	704.645	1.850	2.209	0.826
	(1.804)	(146.620)	(0.736)	(0.661)	(0.343)	
6	200	8.466	1717.645	6.109	5.430	2.943
		(4.571)	(498.850)	(1.637)	(1.513)	(0.840)
	500	14.916	3177.247	13.387	10.593	6.594
		(9.376)	(1120.031)	(2.712)	(2.542)	(1.378)
	1000	24.143	5603.115	28.390	21.097	13.961
	(16.311)	(2181.588)	(5.270)	(3.879)	(2.519)	
5000	43.449	10221.305	73.587	49.665	37.014	
	(22.715)	(3669.447)	(12.971)	(9.169)	(6.473)	

Notes: N: naive MLE, B: bootstrap bias correction, J: penalized likelihood estimation with jeffrey's prior, G: penalized likelihood estimation with generalized information matrix, C: penalized likelihood estimation with cauchy distribution.

studies, method J seems to be the best performing method for reducing bias and variability of the MLE for all parameters regardless of the marginal success probability. In addition to smaller bias, the regression parameter estimators in method J have a lot less variability than any other method. Method C seems to be the next best method after method J. In terms of the computational time method C beats method J.

Following a comment from a reviewer, we conducted a simulation study to assess the performance of cross-validation statistics in choosing the best estimation method. We



Figure 6. Simulation results based on 1000 replications when $X_1 \sim \text{Uniform}(-2, 2)$, $X_2 \sim \text{Normal}(0, (\sqrt{4/3})^2)$, $\delta = 4$, $\beta_0 = -1.09$, $\beta_1 = 1$, $\beta_2 = -0.7$ and $p_m = 12\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey's prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

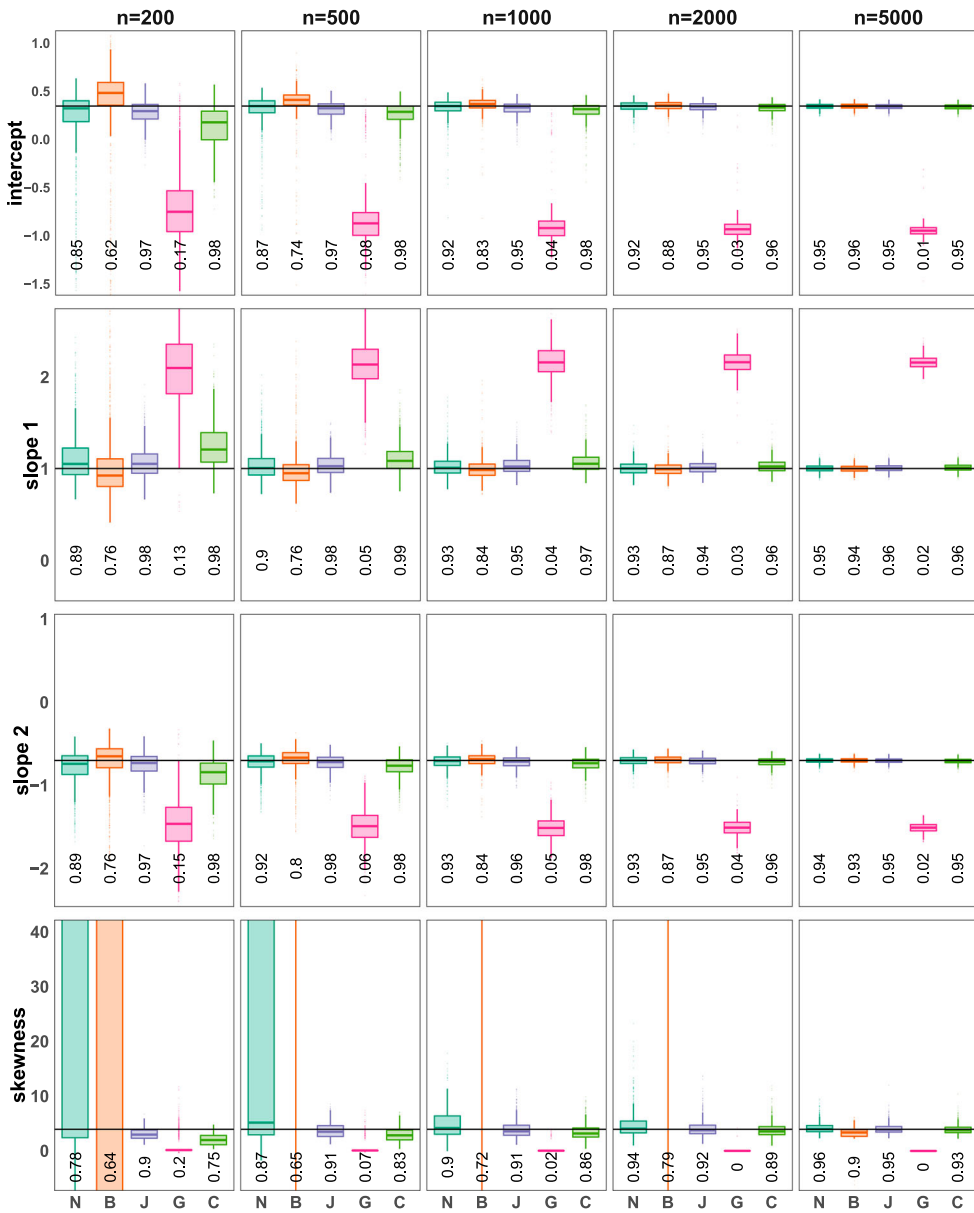


Figure 7. Simulation results based on 1000 replications when $X_1 \sim \text{Uniform}(-2, 2)$, $X_2 \sim \text{Normal}(0, (\sqrt{4/3})^2)$, $\delta = 4$, $\beta_0 = 0.34$, $\beta_1 = 1$, $\beta_2 = -0.7$ and $p_m = 40\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey's prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

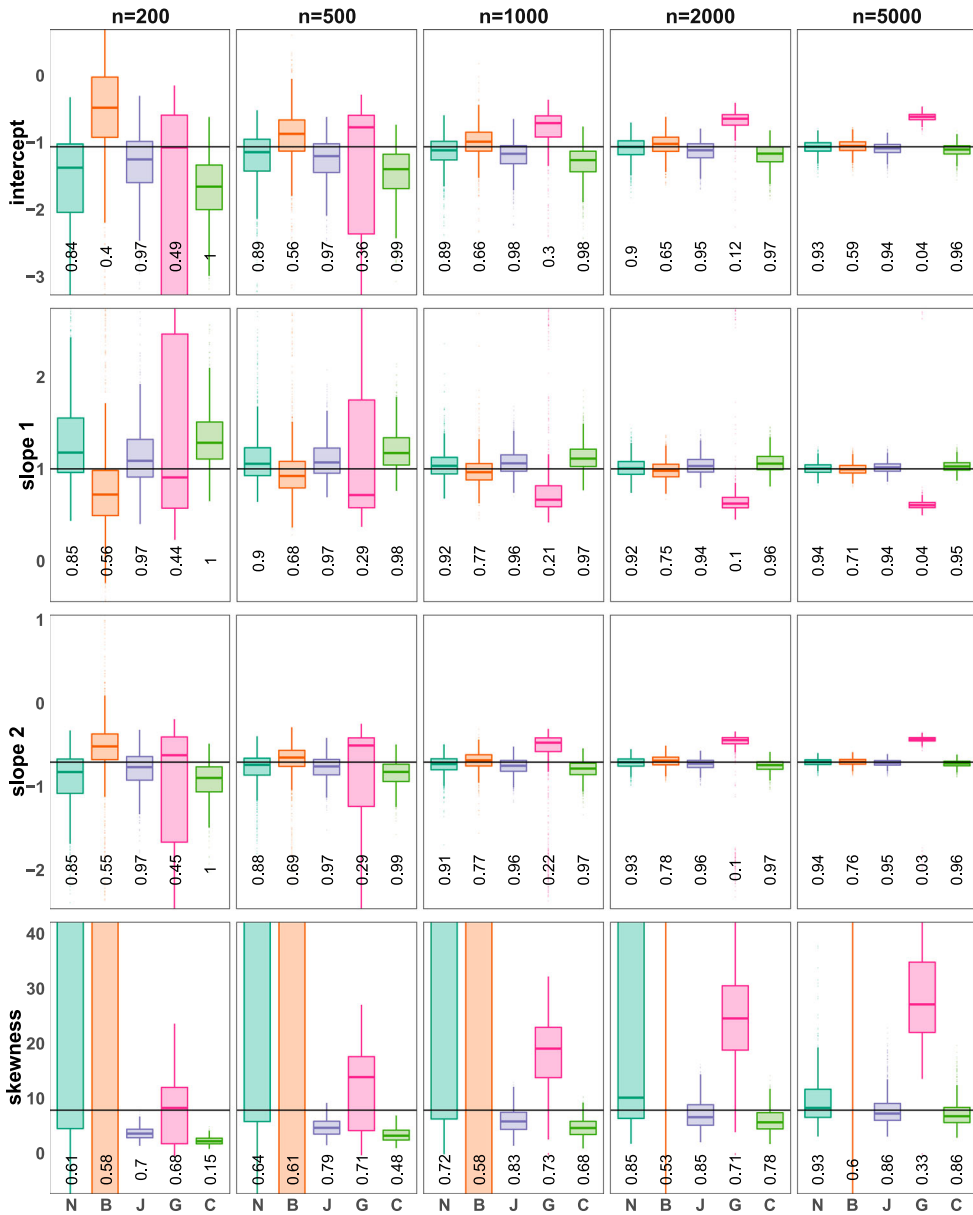


Figure 8. Simulation results based on 1000 replications when $X_1 \sim \text{Uniform}(-2, 2)$, $X_2 \sim \text{Normal}(0, (\sqrt{4/3})^2)$, $\delta = 8$, $\beta_0 = -1.07$, $\beta_1 = 1$, $\beta_2 = -0.7$ and $p_m = 12\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey's prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

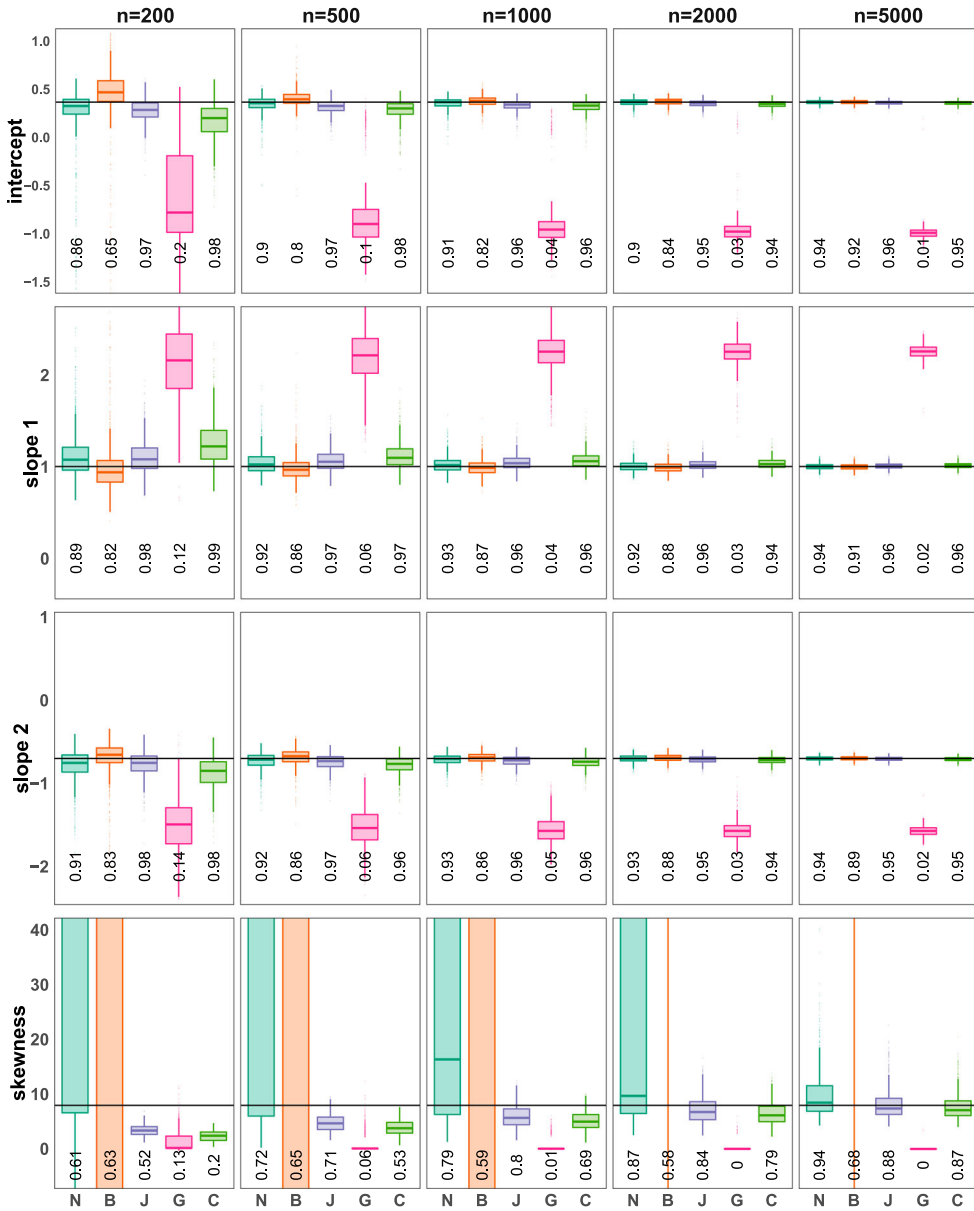


Figure 9. Simulation results based on 1000 replications when $X_1 \sim \text{Uniform}(-2, 2)$, $X_2 \sim \text{Normal}(0, (\sqrt{4/3})^2)$, $\delta = 8$, $\beta_0 = 0.36$, $\beta_1 = 1$, $\beta_2 = -0.7$ and $p_m = 40\%$. The numbers in the boxplots are the empirical coverage probabilities for the nominal level 0.95 based on the standard error derived from the Fisher information matrix. The horizontal line in each figure indicates the true value of the parameter. N: Naive MLE, B: Bootstrap bias correction, J: Penalized likelihood estimation with Jeffrey’s prior, G: Penalized likelihood estimation with generalized information matrix, C: Penalized likelihood estimation with Cauchy distribution.

Table 3. The table contains mean, median, standard deviation (SD) and robust standard deviation based on the IQR (ISD) of the deviance and RSS statistics.

n	Cross-validation measure		Method					
			P	N	B	J	G	C
300	Deviance	Mean	307.490	418.045	3162.402	321.672	322.647	304.325
		Median	306.927	357.303	2940.603	315.700	322.896	304.262
		SD	24.195	242.636	1740.110	49.961	17.120	20.629
	RSS	ISD	21.443	58.079	1515.310	26.272	16.349	21.281
		Mean	50.407	50.484	102.407	50.143	53.329	50.160
		Median	50.485	50.554	101.032	50.222	53.425	50.187
600	Deviance	SD	3.872	3.841	24.151	3.778	3.538	3.855
		ISD	3.991	3.946	26.680	3.866	3.487	4.026
		Mean	591.730	646.601	7753.755	595.498	637.807	589.779
	RSS	Median	592.959	616.880	8126.433	594.830	639.159	590.650
		SD	25.741	155.213	3453.938	31.172	22.139	26.076
		ISD	24.225	42.726	2928.648	27.137	20.352	25.046
600	Deviance	Mean	98.037	97.909	230.725	97.695	105.482	97.712
		Median	98.159	98.046	241.134	97.760	105.752	97.772
		SD	5.036	5.039	51.762	4.981	4.670	5.024
	RSS	ISD	4.815	4.755	57.809	4.753	4.310	4.855

Table 4. The number of times each method is selected as the best method based on the cross-validation criteria.

n	Cross-validation measure	Method						Total counts
		P	N	B	J	G	C	
300	Deviance	344	11	4	199	4	438	1000
	RSS	202	82	3	469	1	243	1000
600	Deviance	353	11	3	172	0	461	1000
	RSS	258	75	3	429	0	235	1000

considered two K-fold cross validation statistics [[28, Chapter 6.9], [29]]:

$$\begin{aligned}
 \text{Deviance} &\stackrel{\text{def}}{=} -2 \sum_{t=1}^K \sum_{i=1}^{n_t} [\{y_i^t \log(\hat{\pi}_i^t) + (1 - y_i^t) \log(1 - \hat{\pi}_i^t)\}], \\
 \text{RSS} &\stackrel{\text{def}}{=} \sum_{t=1}^K \sum_{i=1}^{n_t} (y_i^t - \hat{\pi}_i^t)^2,
 \end{aligned}$$

where K is the number of non-overlapping exhaustive subsets of a given dataset, n_t is the size of the t th subset, $t = 1, \dots, K$, y_i^t is the i th observation from the t th subset, $\hat{\pi}_i^t$ is the estimated probability for the i th observation from the t th subset based on the estimated coefficients using the data from the $(K - 1)$ subsets excluding the t th subset. For this assessment, we simulated datasets mimicking the real dataset with multiple covariates, where the marginal success probability was 46%. We considered two different sample sizes $n = 300$ and $n = 600$, while $n = 300$ was close to the sample size 297 of the real dataset analysed in this paper. For generating response variable Y , we used the skew-probit link function and set the true values of the regression parameters β and the skewness parameter δ to the real data estimates obtained under method J (see Section 5).

Next, we fitted the skew-probit model to the simulated datasets using all five different approaches, also we fitted the probit model to the simulated datasets. Under each of six

different approaches, we calculated deviance and RSS using $K = 5$ and $n_t = 60$ for $n = 300$ and $n_t = 120$ for $n = 600$, for all $t = 1, \dots, 5$.

In Table 3 we present the mean, median, standard deviation (SD) and robust standard deviation based on the IQR (ISD) of the deviance and RSS statistics. Here ISD is defined as $(Q3 - Q1)/1.349$, where $Q1$ and $Q3$ denote the first and the third quartile of the statistic (Deviance or RSS) based on 1000 replications. Results indicate that based on the mean or median value of deviance, the minimum occurs for method C. Based on the mean value of RSS, for $n = 300$ the minimum occurs for method J. To get a better picture of this simulation study in Table 4 we show how many times each approach possesses the smallest statistics (deviance and RSS) out of the six competing approaches. Based on the minimum RSS value, method J was selected maximum number of times for both sample sizes.

5. Application to heart-disease data

For the illustration purpose, we analyse the heart-disease data from the Cleveland database [22]. The dataset can be found in UCI database (<http://archive.ics.uci.edu/ml/datasets/Heart+Disease>). The goal of this analysis is to fit a model that explains the association between Y , the occurrence of a $> 50\%$ diameter narrowing in an angiography, and other clinical and test variables. In our analysis we consider subjects who have complete observations without any missing values. With this definition we have a total of 297 subjects out of 303 subjects in our analysis, and 137 (46.13%) of them experienced the primary event. Among 13 available covariates, we choose the following 6 covariates which are statistically significant at the 5% level from a probit model: gender (Gender), chest pain type (CP), resting blood pressure (BP), the slope of the peak exercise ST segment (Slope), number of major vessels coloured by flourosopy (CF), and thallium heart scan results (Thal). We create relevant dummy variables for the categorical covariates, Gender = 1 for male and 0 for female; CP_{TA} , CP_{AA} and CP_{NA} are dummies for chest pain types, typical angina, atypical angina and non-anginal pain, respectively with asymptomatic being the reference; $Slope_U$ and $Slope_D$ are dummies for upsloping and downsloping of ST segment with flatness as the reference; and $Thal_F$ and $Thal_R$ are dummies for fixed detect and reversible detect while normal is considered as the reference category for Thal. Here BP and CF are continuous.

We analysed this data using the skew-probit model and estimate the model parameters using methods N, B, J, G and C. In addition, we also fitted the probit model for comparison purpose. For the analysis, we set initial values of β parameters to the probit regression estimates. We took different initial values for δ with some positive and some negative values. For each set of initial values, we obtained the parameter estimate using the `ucminf` function. The reported parameter estimates correspond to the minimum value of the negative of the log-likelihood function. In Table 5 we provide the estimates and 95% Wald-type confidence interval for each parameter based on the standard error calculated from the Fisher information matrix (CI).

The 95% CI for δ based on method J indicates that δ is significantly different from 0 at the 5% level ($\hat{\delta}$: 2.73 and 95% CI: 0.566, 4.893). On the other hand, we note that the 95% CIs for δ based on the other approaches indicate δ is not statistically significant. Also we note that parameter estimates of methods N and C are close to each other including the skewness

Table 5. Results of the analysis of the heart-disease data.

Covariate		Method					
		P	N	B	J	G	C
δ	Est	—	1.540	0.954	2.730	0.139	1.468
	CI	—	(-0.353, 3.433)	(-2.833, 4.741)	(0.566, 4.893)	(-0.002, 0.279)	(-0.166, 3.103)
Intercept	Est	-0.356	0.382	0.569	0.481	-0.305	0.364
	CI	(-0.816, 0.104)	(-0.062, 0.827)	(-0.569, 1.708)	(0.191, 0.771)	(-0.801, 0.190)	(-0.073, 0.801)
Gender	Est	0.815	0.608	0.515	0.501	1.101	0.597
	CI	(0.315, 1.315)	(0.197, 1.018)	(-0.107, 1.137)	(0.197, 0.806)	(0.507, 1.695)	(0.200, 0.993)
CP _{TA}	Est	-1.355	-0.985	-0.791	-0.794	-1.561	-0.959
	CI	(-2.055, -0.654)	(-1.604, -0.366)	(-1.643, 0.060)	(-1.237, -0.350)	(-2.306, -0.816)	(-1.541, -0.378)
CP _{AA}	Est	-0.917	-0.680	-0.597	-0.582	-0.604	-0.673
	CI	(-1.481, -0.353)	(-1.132, -0.228)	(-1.276, 0.083)	(-0.905, -0.259)	(-1.206, -0.003)	(-1.115, -0.230)
CP _{NA}	Est	-1.272	-0.911	-0.804	-0.728	-1.542	-0.904
	CI	(-1.754, -0.790)	(-1.413, -0.409)	(-1.630, 0.021)	(-1.068, -0.389)	(-2.113, -0.972)	(-1.373, -0.435)
BP	Est	1.959	1.420	1.167	1.154	2.148	1.316
	CI	(0.458, 3.459)	(0.210, 2.631)	(-0.391, 2.725)	(0.212, 2.095)	(0.573, 3.723)	(0.177, 2.455)
Slope _U	Est	-0.963	-0.697	-0.604	-0.551	-1.206	-0.695
	CI	(-1.398, -0.528)	(-1.114, -0.281)	(-1.246, 0.039)	(-0.854, -0.248)	(-1.680, -0.731)	(-1.089, -0.301)
Slope _D	Est	-0.230	-0.204	-0.184	-0.190	-0.379	-0.192
	CI	(-0.976, 0.515)	(-0.775, 0.367)	(-0.795, 0.428)	(-0.665, 0.285)	(-1.311, 0.554)	(-0.760, 0.376)
CF	Est	0.666	0.514	0.445	0.433	0.839	0.516
	CI	(0.416, 0.917)	(0.283, 0.746)	(0.041, 0.848)	(0.259, 0.607)	(0.595, 1.084)	(0.295, 0.738)
Thal _F	Est	0.051	0.009	0.026	-0.029	-0.105	0.024
	CI	(-0.752, 0.855)	(-0.602, 0.620)	(-0.635, 0.686)	(-0.546, 0.488)	(-0.878, 0.668)	(-0.582, 0.630)
Thal _R	Est	0.820	0.602	0.526	0.492	0.791	0.613
	CI	(0.383, 1.257)	(0.210, 0.993)	(-0.034, 1.086)	(0.196, 0.788)	(0.344, 1.237)	(0.230, 0.995)

Notes: Est: estimate, CI: the wald confidence interval where standard errors are calculated by inverting the fisher information matrix, P: probit model with MLE, N: skew-probit model with MLE, B: skew-probit model with bootstrap bias correction, J, G, C: skew-probit model with jeffrey's prior, generalized information matrix and cauchy prior penalization, respectively.

parameter δ . In terms of estimates for other covariates, male subjects have higher risk for heart-disease than female subjects while any kind of chest pain is associated with a lower probability of heart-disease compared to the asymptomatic pain. Also, based on method J, BP, CF and Thal_F turn out to be positively associated with the probability of $Y = 1$. Although, the statistical significance of regression parameters β (except the intercept) do not change across methods P, N, J, G and C, method J yields narrower confidence intervals compared to other methods.

In order to get a sense of which of the methods provides the best fit for this dataset, we provide the deviance residual plot (Figure 10) and conduct a cross-validation study. Based on the $1.5 \times$ IQR criteria we have identified the outlying residual values that are denoted by the + notation in the residual plots. Two very competing approaches J and C show similar residual plots. In method J we found six outlying residuals while in method C we found eight outlying residuals including the 6 that were identified in method J. For the cross-validation study we have used the deviance and RSS statistics described in the simulation section with $K = 5$. Table 6 contains the value of the two statistics for all these methods. The results indicate that method J yields the minimum deviance statistic while method C yields the minimum RSS statistic. However, we want to point out RSS values for methods C, J and P are quite close. Although, in light of the simulation results concerning the cross-validation methods either method J or method C provides the best fit to the dataset, considering a relatively small sample size and overall superior performance of method J in the simulation study, we are more confident with the method J estimates.

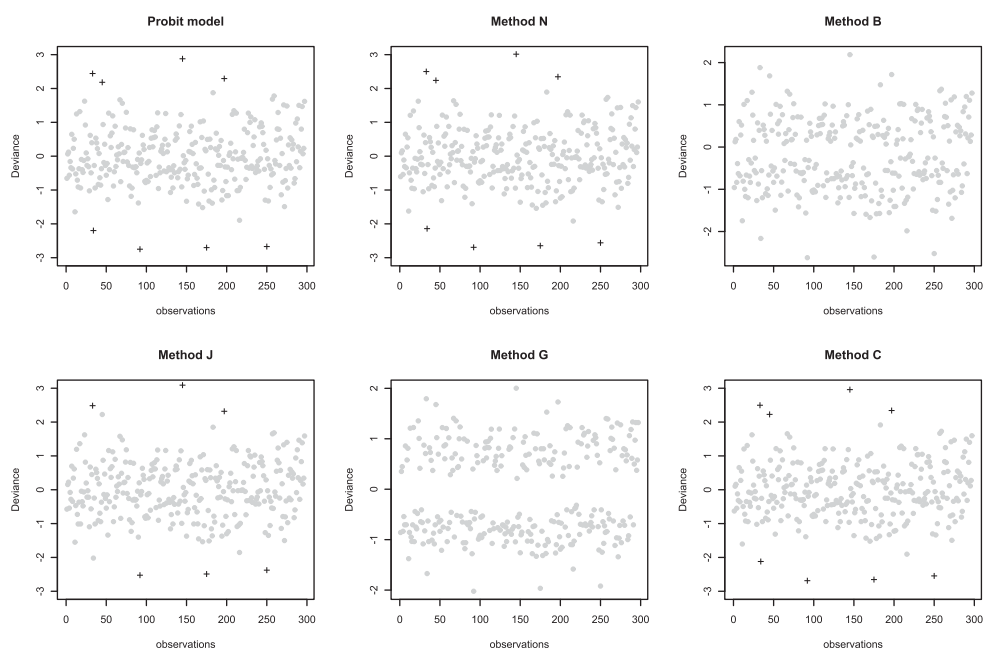


Figure 10. Deviance residuals for the real data analysis.

Table 6. Cross-validated goodness-of-fit statistic for the heart-disease data.

CV measure	P	N	B	J	G	C
Deviance	244.83	367.21	3901.25	233.72	303.24	243.32
RSS	35.54	35.17	122.50	35.19	39.89	35.04

6. Conclusion

We have investigated parameter identifiability and several bias reduction approaches for the MLE of the skew-probit model for a binary response variable. The identifiability results will guide researchers to craft their model more carefully for the skew-probit link function. Several bias reduction strategies have been considered, and through simulation studies we have compared the performance of different approaches. Overall, method J is the best performing for small to large sample sizes, for small to moderate marginal success probabilities of the response. The next competing method is method C whose performance becomes similar to that of method J for a large sample size. Variability of the intercept and slope parameter estimator under method J is always smaller than that of method C, and this difference is somewhat significant when the sample size is less than 1000. The variability of the skewness parameter estimator is comparable between methods J and C. Simulation results also indicate that even with the best performing approach, one needs moderate to large sample size to estimate the skewness parameter of the skew-probit model reasonably well. Finally, we have applied the proposed strategies to analyse a real dataset on heart-disease, and the results show that methods with a proper bias correction provide a better fit than the regular MLE. Overall this research and the simulation results will help to develop a

unique and robust method of analyses for the skew-probit model. We believe that a similar study can be done for other link functions [9,30].

Acknowledgments

The authors would like to thank associate editor and two reviewers for constructive suggestions that led to a substantive improvement in the quality of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- [1] Chen MH, Dey DK, Shao QM. A new skewed link model for dichotomous quantal response data. *J Am Stat Assoc.* 1999;94:1172–1186.
- [2] Azzalini A. A class of distributions which includes the normal ones. *Scand J Stat.* 1985;12:171–178.
- [3] Genton MG, He L, Liu X. Moments of skew-normal random vectors and their quadratic forms. *Stat Probab Lett.* 2001;51:319–325.
- [4] Ma Y, Genton MG. Flexible class of skew-symmetric distributions. *Scand J Stat.* 2004;31:459–468.
- [5] Bazán JL, Branco MD, Bolfarine H. A skew item response model. *Bayesian Anal.* 2006;1:861–892.
- [6] Bazán JL, Bolfarine H, Branco MD. A framework for skew-probit links in binary regression. *Commun Stat Theory Methods.* 2010;39:678–697.
- [7] Stingo FC, Stanghellini E, Capobianco R. On the estimation of a binary response model in a selected population. *J Stat Plan Inference.* 2011;141:3293–3303.
- [8] Bazán JL, Romeo JS, Rodrigues J. Bayesian skew-probit regression for binary response data. *Braz J Probab Stat.* 2014;28:467–482.
- [9] Kim S, Chen MH, Dey DK. Flexible generalized t-link models for binary response data. *Biometrika.* 2008;95:93–106.
- [10] Genton MG, Zhang H. Identifiability problems in some non-Gaussian spatial random fields. *Chil J Stat.* 2012;3:171–179.
- [11] Castro LM, Martín ES, Arellano-Valle RB. A note on the parameterization of multivariate skewed-normal distributions. *Braz J Probab Stat.* 2013;27:110–115.
- [12] Otiniano CEG, Rathie PN, Ozelim LCSM. On the identifiability of finite mixture of skew-normal and skew-t distributions. *Stat Probab Lett.* 2015;106:103–108.
- [13] Firth D. Bias reduction of maximum likelihood estimates. *Biometrika.* 1993;80:27–38.
- [14] Sartori N. Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *J Stat Plan Inference.* 2006;136:4259–4275.
- [15] Azzalini A, Arellano-Valle RB. Maximum penalized likelihood estimation for skew-normal and skew-t distributions. *J Stat Plan Inference.* 2013;143:419–433.
- [16] Liseo B, Loperfido N. A note on reference priors for the scalar skew-normal distribution. *J Stat Plan Inference.* 2006;136:373–389.
- [17] Bayes CL, Branco MDE. Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Braz J Probab Stat.* 2007;21:141–163.
- [18] Cox DR, Snell EJ. A general definition of residuals. *J R Stat Soc Ser B.* 1968;30:248–275.
- [19] Gupta M, Ibrahim JG. An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data. *Stat Sin.* 2009;19:1641–1663.
- [20] Gelman A, Jakulin A, Pittau MG, et al. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat.* 2008;2:1360–1383.
- [21] Berger JO, Bernardo JM, Sun D. Overall objective priors. *Bayesian Anal.* 2015;10:189–221.

- [22] Detrano R, Janosi A, Steinbrunn W, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. *Am J Cardiol.* [1989](#);64:304–310.
- [23] Rothenberg TJ. Identification in parametric models. *Econometrica.* [1971](#);39:577–591.
- [24] Parrish R, Bargmann R. A method for evaluation of cumulative probabilities of bivariate distributions using the Pearson family. *Stat Distrib Sci Work.* [1981](#);5:241–257.
- [25] Azzalini A, Capitanio A. *The skew-normal and related families.* Cambridge (UK): Cambridge University Press; [2014](#).
- [26] Henze N. A probabilistic representation of the ‘skew-normal’ distribution. *Scand J Stat.* [1986](#);13:271–275.
- [27] Nielsen HB. UCMINF – an algorithm for unconstrained, nonlinear optimization. Department of mathematical modelling, Technical University of Denmark; 2000. (Report IMM-REP-2000-18).
- [28] Hastie TJ, Tibshirani RJ. *Generalized additive models.* 1st ed. Boca Raton (FL): CRC Press; [1990](#).
- [29] Kim KI, Simon R. Overfitting, generalization, and MSE in class probability estimation with high dimensional data. *Biom J.* [2014](#);56:256–269.
- [30] Li D, Wang X, Song S, et al. Flexible link functions in a joint model of binary and longitudinal data. *Stat.* [2015](#);4:320–330.