https://communications.science.ankara.edu.tr

Commun.Fac.Sci.Univ.Ank.Ser. A1 Math. Stat.
Volume 69, Number 2, Pages 1083–1103 (2020)
DOI: 10.31801/cfsuasmas.614492
ISSN 1303–5991 E-ISSN 2618–6470

Received by the editors: September 02, 2019; Accepted: May 29, 2020

A COMPARATIVE STUDY ON THE PERFORMANCE OF FREQUENTIST AND BAYESIAN ESTIMATION METHODS UNDER SEPARATION IN LOGISTIC REGRESSION

Yasin ALTINISIK

Department of Statistics, Sinop University, 57000 Sinop, TURKEY

Abstract. Separation is one of the most commonly encountered estimation problems in the context of logistic regression, which often occurs with small and medium sample sizes. The method of maximum likelihood (MLE; [8]) provides spuriously high parameter estimates and their standard errors under separation in logistic regression. Many researchers in social sciences utilize simple but ad-hoc solutions to overcome this issue, such as "doing nothing strategy", removing variable(s) from the model, and combining the levels of the categorical variable in the data causing separation etc. The limitations of these basic solutions have motivated researchers to use more appropriate and innovative estimation techniques to deal with the problem. However, the performance and comparison of these techniques have not been fully investigated yet. The main goal of this paper is to close this research gap by comparing the performance of frequentist and Bayesian estimation methods for coping with separation. A simulation study is performed to investigate the performance of asymptotic, bootstrap-based, and Bayesian estimation techniques with respect to bias, precision, and accuracy measures under separation. In line with the simulation study, a real-data example is used to illustrate how to utilize these methods to solve separation in logistic regression.

1. Introduction

The logistic regression is a well-founded analysis technique that can be utilized to determine the relationship between a dichotomous outcome and a set of categorical and/or continuous predictors. Although researchers in social sciences often do not encounter challenges in applying this technique to their data sets, complications may arise when a linear combination of predictors allocate the values of

2020 *Mathematics Subject Classification.* Primary 05C38, 15A15; Secondary 05A15, 15A18.

Keywords and phrases. Logistic regression, separation problem, frequentist and Bayesian estimation, bias, precision, and accuracy measures.

yaltinisik@sinop.edu.tr

0000-0001-9375-2276.

outcome, which is called the separation problem [1]. To illustrate the separation problem in logistic regression, consider the simplest scenario in which a dichotomous response is predicted by a continuous predictor. Suppose that the outcome has the values of $\mathrm{R} = \{0, 0, 0, 0, 0, 1, 1, 1, 1, 1\}$ and the predictor has the values of $P = \{2, 7, 3, 5, 6, 9, 14, 10, 12, 16\}$. In this case, the values of response are zero when the values of predictor are smaller than 8 and the values of response are 1 for the values of predictor greater than 8. This implies that the probability of observing zero or one is perfectly predicted (known as complete separation) and there is nothing left to be estimated. When separation occurs, the method of maximum likelihood (MLE; [8]) does not provide a reliable set of parameter estimates and their standard errors, which in turn cause to obtain undependable test statistics. Many researchers benefit from basic (but ad-hoc) solutions to overcome separation in logistic regression.

Since separation does not necessarily have a negative influence on all parameters in the model, some researchers do not pay special attention to this issue by simply and only reporting their results with respect to chi-square test statistics; although these statistics are only correct for non-problematic variables in the data. However, these variables often interact with problematic ones, and thus, the estimates and standard errors of these interactions should not be trusted either. Moreover, if the variable causing separation is categorical, then the estimates obtained for other variables in the model are not interpretable, since they are determined on the basis of the reference level of this categorical variable. Some researchers avoid these issues by removing the problematic variable(s) from the model. However, this approach is subject to two main drawbacks. First, discarding an important variable may end up with an inappropriate model specification, and consequently, a set of bias estimates for model parameters, which is known as the omitted variable bias [24]. Second, even if a predictor causing separation has an insignificant (or weakly significant) effect on the outcome, caution should be taken when eliminating this variable from the model, since it can be a confounder. That is, the relationship between this variable and the outcome may influence the outcome's associations with other variables in the model. Another common way of coping with this issue is combining the levels of variable causing separation, which is only applicable when this variable is categorical. This approach is also not recommended not only because collapsing categories alter the research question at hand, but also because it may cause the loss of information obtained from the data [1].

In response to these challenges, many researchers focus on more complicated but powerful data analysis techniques to deal with separation in logistic regression. Heinze and Schemper [14] compare the performance of Firth's penalized maximum likelihood estimation (PMLE; [7]) against the method of maximum likelihood [8], an imputation method using Bayesian logistic regression [3], and exact logistic regression [22]. This study is limited in the sense that it investigates the performance of only these four methods with respect to (only) bias measures. In the discussion

of their study, they suggest the use of Firth's method to cope with separation in logistic regression. Moreover, they state that the separation problem may not only occur in the original sample, but it may also occur in bootstrap samples. However, they do not inspect the performance of Firth's method in the context of bootstrapping. Ohkura and Kamakura [28] utilized nonparametric bootstrapping in conjunction with Firth's method to compare the performance of their bootstrap-base test against Wald and Firth's tests under separation. However, the performance of Firth's method with nonparametric bootstrapping has not been compared against any Bayesian estimation method and the usual Firth's method with respect to bias, precision and accuracy measures. This study aims at filling this gap by investigating and comparing the performance of frequentist and Bayesian estimation methods with respect to bias, precision, and accuracy measures, respectively. Here, frequentist way of coping with separation is performed using Firth's method [7] and its counterpart with nonparametric bootstrapping [6]. The choice of prior distribution is a crucial point to solve separation in logistic regression using Bayesian methods. Thus, the Markov Chain Monte Carlo (MCMC) algorithms are utilized as Bayesian solutions to separation using seven different priors.

The outline of the paper is as follows. In Sections 2 and 3, the logistic regression and the separation problem in logistic regression are elaborated, respectively. In Section 4, three methods used to obtain the estimates of model parameters and their standard errors under separation are described. In Section 5, a simulation study is performed to investigate and compare the performance of these methods with respect to bias, precision, and accuracy measures. In Section 6, a real life example is presented to exemplify how to deal with separation using these estimation techniques in logistic regression. The paper will be concluded with a brief discussion.

2. Logistic Regression Modeling

The logistic regression is one of the most commonly used analysis techniques to predict a binary outcome (containing zeros and ones) in the context of generalized linear models [21]. The logistic regression model is defined as:

$$f(\pi_i) = x_i^T \beta \quad , \quad i = 1, 2, ..., N, \tag{1}$$

where $\pi_i = E(y_i)$ is the expected value of the binary outcome for the ith observation, $\beta = (\beta_0, \beta_1, ..., \beta_{P-1})^T \in \mathbb{R}^{\mathrm{P}\times 1}$ is the vector of model parameters and $x_i^T = (1, x_{i1}, x_{i2}, ..., x_{i(P-1)}) \in \mathbb{R}^{\mathrm{N}\times \mathrm{P}}$ is the design matrix containing ones in the first column as the coefficients of the intercept, β_0, and the values of the explanatory variables in the data, respectively. The logit link function, $f(\pi_i)$, relates the expected values of the outcome to the linear predictor, $x_i^T\beta$:

$$f(\pi_i) = log(\frac{\pi_i}{1-\pi_i}), \tag{2}$$

where $\pi_i = \frac{exp(x_i^T\beta)}{1+exp(x_i^T\beta)}$, which is also known as the conditional probability of success.

Since the outcome containing 0's and 1's has a Bernoulli distribution with the probability of success π_i for the ith observation, the likelihood function of the data can be defined as follows:

$$L(\beta \mid y_1, y_2, ..., y_N) = \prod_{i=1}^{N} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \tag{3}$$

where $y_i \in \{0, 1\}$ for $i = 1, 2, ..., N$. The likelihood function above is not easy to differentiate, and thus, it is transformed from the original scale into the log scale:

$$log\ L(\beta \mid y_1, y_2, ..., y_N) = \sum_{i=1}^{N} y_i\ log\ (\pi_i) + (1 - y_i)\ log\ (\ 1 - \pi_i\ ). \tag{4}$$

The β's are estimated by maximizing the log likelihood function above using the method of maximum likelihood [8], so that the data at hand have the highest probability of being observed. This is done by differentiating the log likelihood function above with respect to the β's, setting the resulting functions to zeros and solving the equations for each of β's, respectively.

Since the maximum likelihood estimates of model parameters, the $\hat{\beta}$'s, and their standard errors do not involve closed-form solutions, they are obtained numerically. This can be achieved quickly and conveniently by utilizing computer-intensive iterative methods such as the Newton-Raphson algorithm [27]. However, there may be certain situations in which even the numerical methods fail to provide parameter estimates and their standard errors. In the next section, one of these situations called the separation problem will be elaborated.

3. Separation Problem

The logistic regression cannot always be easily used to predict a dichotomous outcome containing zeros and ones. One common issue that arises when estimating model parameters and their standard errors in the context of logistic regression causing (nearly) perfect allocation of the values of an outcome in the data at hand is called the (quasi) complete separation problem [1]. In a regular situation in which there is no problem of (quasi) complete separation, the expected probabilities of an outcome for a logistic regression model can take values between the numbers 0 and 1. In complete separation, since a linear function of predictor(s) perfectly predicts the outcome, the expected probabilities are either 0 or 1 (and not between these values). Similarly, in quasi complete separation, since the values of an outcome almost perfectly predicted, almost all expected probabilities (but not all of them) are either 0 or 1.

Figure 1 is created based on two empirical data sets given in the study of [33, p. 276], which shows the scatter plot of the values of an outcome against that of a linear predictor in the presence of complete and quasi-complete separation. As can be seen on the left panel of the figure for the first data, the values of the linear predictor perfectly separate the values of the outcome. Thus, only by observing the

plot, we can make a perfect inference about the predicted values of the outcome. That is, the predicted values of the outcome take the value of zero when the linear predictor is smaller than zero and take the value of one when the linear predictor is larger than one. Similarly, as can be seen on the right panel of the figure for the second data, the values of the linear predictor nearly perfectly separate the values of the outcome, which is a sign of quasi-complete separation. In this case, the predicted values of the outcome take the value of zero, a value between zero and one (only for three observations) and the value of one, when the linear predictor is smaller than zero, equal to zero, and larger than zero, respectively. Next, it will be

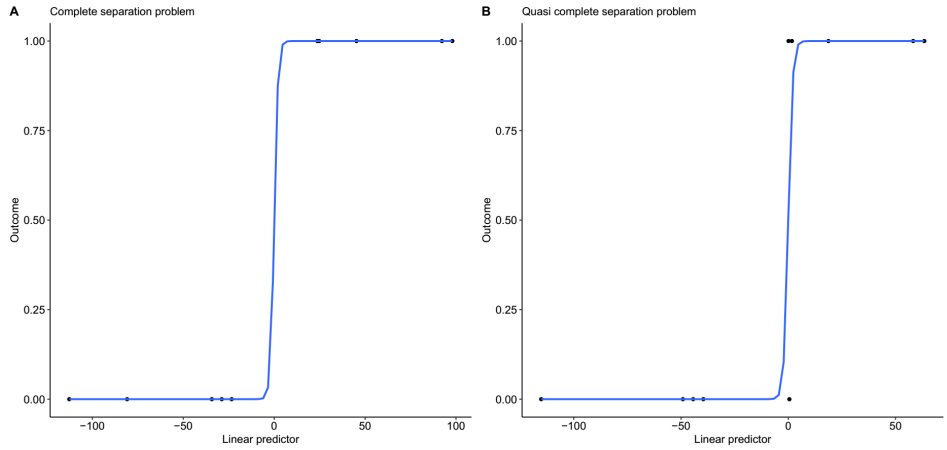

FIGURE 1. Illustrations of the (quasi) complete separation problem

elaborated how to remedy the adverse impacts of separation in estimating model parameters and their standard errors using three different estimation methods.

4. Estimation Methods

The separation [1] often occurs with small and medium sample sizes when estimating model parameters and their standard errors in logistic regression. The Newton-Raphson algorithm used to obtain MLEs does not converge for (some of) model parameters when the data suffer from separation. This nonconvergence causes spuriously high parameter estimates and standard errors [33, pp. 282-283] and results in unreliable test statistics and hypothesis testing. In response to this challenge, researchers have been paying attention to more appropriate estimation techniques than MLE to overcome separation in logistic regression. In the sequel, three of such advanced estimation methods will be elaborated, respectively.

Firth's method: Firth [7] proposed a method to improve the parameter estimates in logistic regression by reducing the bias occurs with small samples when using the

method of maximum likelihood for estimation. Since Firth's method incorporates a penalizing factor into the log likelihood in (4), it is also known as the method of penalized maximum likelihood estimation. Firth's penalized log likelihood function is defined as:

$$L^*(\beta \mid y_1, y_2, ..., y_N) = L(\beta \mid y_1, y_2, ..., y_N) + \frac{1}{2}\, log\, |I(\beta)|, \tag{5}$$

where $I(\beta) = x_i^T W x_i$ is the information matrix and $W = \text{diag}[\pi_\text{i}(1 - \pi_\text{i})]$ [35, p. 164]. Heinze and Schemper [14] have adopted the penalized log likelihood function above to overcome separation in the analysis of two cancer studies. Firth's method is flexible in the sense that it can be incorporated into nonparametric resampling techniques when estimating model parameters and their standard errors.

Firth's method with nonparametric bootstrapping: Nonparametric bootstrapping [6] is a resampling (with replacement) technique that can be used as an alternative to the method of maximum likelihood to obtain MLEs and their standard errors, when model assumptions are not satisfied (see [34], [15, p. 44]). Nonparametric bootstrapping uses the information given in the original sample to generate, for example, $B = 1000$ bootstrap samples, in each of which model parameters are estimated using the method of maximum likelihood. Subsequently, it calculates the averages and standard deviations of the bootstrap estimates across these samples to obtain the overall parameter estimates and their standard errors.

The usual nonparametric bootstrapping using the method of maximum likelihood for estimation in each bootstrap sample assumes that the original sample adequately represents the population of interest, which is often not a reasonable assumption for small samples. Thus, since separation usually occurs with small and medium samples, it is not recommended to use nonparametric bootstrapping in conjunction with MLEs under separation. Nonparametric bootstrapping can still be used for a small or medium sample in the context of logistic regression when the data suffer from separation. This can be done by replacing MLEs with PMLEs obtained using Firth's method in each bootstrap sample. The method of maximum likelihood and nonparametric bootstrapping with MLEs produce bias estimates with small samples [15], and thus, they should not be used to overcome separation in logistic regression. Bayesian methods are good alternatives to Firth's method and nonparametric bootstrapping with PMLEs to deal with separation in logistic regression.

Bayesian approach using MCMC algorithms: Bayesian estimation using Markov chain Monte Carlo (MCMC) algorithms benefits from prior knowledge on the distribution of model parameters and information in the data at hand to generate posterior samples, which are, in turn, utilized to obtain parameter estimates and their standard errors. The Metropolis Hastings [13, 23], Gibbs sampling [10], and Hamiltonian Monte Carlo (HMC; [2, 5, 26]) are three of the best known MCMC algorithms that can be used to obtain the estimates of model parameters and their standard errors for small samples in logistic regression. The HMC (also known

as Hybrid Monte Carlo) and Gibbs sampling algorithms are used for Bayesian estimation in this paper using the R packages "rstanarm" [12], "runjags" [4], and "bayesreg" [19].

Rainey [29] suggests to utilize two priors when estimating model parameters using Bayesian approaches under separation in logistic regression, which are Jeffrey's invariant prior [16], [35] and a weakly informative Cauchy(0, 2.5) prior [9]. Bayesian approach using Jeffrey's prior is the same with Firth's penalized maximum likelihood estimation method, since the penalty part of the log likelihood function in (5), $\frac{1}{2}\ log\ |I(\beta)|$, is equal to the log of Jeffrey's prior in logistic regression [29]. Moreover, using weakly informative Cauchy(0, 2.5) prior to cope with separation in logistic regression is highly controversial. Ghosh, Li and Mitra [11] state that using a Cauchy(0, 2.5) prior imposes too much insufficient information into the analysis to overcome separation in logistic regression. They show that using Cauchy(0, 2.5) prior may cause spuriously high posterior means for parameters in the presence of separation in logistic regression and may not even enable researchers to obtain these means. Their results suggest to use weakly informative priors with lighter tails than that of Cauchy(0, 2.5) prior such as Normal and Student-t (df = 7) priors. Thus, in addition to Cauchy(0, 2.5) prior, a weakly informative Normal(0, 2.5) prior (the default prior for regression coefficients in rstanarm) and Student-t(0, 2.5, df = 7) prior will be utilized to obtain parameter estimates and their standard errors.

Mansournia, Geroldinger, Greenland, and Heinze [20] utilize Firth's method [7], Ridge logistic regression [31], lasso logistic regression [17], [30], and Bayesian estimation using weakly informative priors. The difference between the current study and the study in Mansournia et al. [20] is threefold. First, Mansournia et al. [20] utilize Bayesian estimation using only Cauchy(0, 2.5) and Log-F(1, 1) priors. As will be shown later in this paper, Bayesian estimation using these priors does not necessarily perform well in logistic regression under separation problem. Thus, the current study also uses Bayesian estimation via Normal(0, 2.5), Student-t(0, 2.5, df = 7), and Log-F(2, 2) priors. Second, Mansournia et al. [20] do not perform a simulation study to inspect the performance of methods used in their study, while the current study compares the performance of both frequentist and Bayesian estimation methods with respect to bias, precision, and accuracy measures. Third, Mansournia et al. [20] investigate the frequentist Ridge and Lasso logistic regressions to cope with separation. Researchers often need to determine the value of a penalizing parameter ($\lambda \geq 0$; also called the tuning or shrinkage parameter utilized on all the regression coefficients besides the intercept in the model) using, for example, cross-validation in order to employ these techniques to solve the problem. However, obtaining the tuning parameter λ is often a complicated and cumbersome task in logistic regression under separation. In many cases where the data suffer from the separation problem the tuning parameter can be estimated as very close to zero, which means that the penalized estimates are very close to the usual MLEs.

To remedy this, the current study does not inspect the usual Ridge and Lasso logistic regressions to solve the separation problem in logistic regression, but instead it utilizes their Bayesian counterparts, that is, Bayesian Ridge and Bayesian Lasso logistic regressions. Note that the tuning parameter λ is set to 1 in Bayesian Ridge logistic regression and $\lambda^2 \sim \text{Exp}(1)$ in Bayesian Lasso logistic regression for each regression coefficients in the model (see [19, p. 7]).

5. Simulation study

5.1. **Simulation Steps.** In this section, the performance of the methods on estimating model parameters will be compared to each other for the data sets containing separation in the context of logistic regression. The model used in the simulation is:

$$f(\pi_i) = \beta_0 + \beta_1 I_i + \beta_2 x_{i1} + \beta_3 x_{i2}, \tag{6}$$

where $f(\pi_i)$ is the logit link function in (2), β_0 is the intercept, β_1 is the coefficient of a dummy variable I_i and β_2 and β_3 are the coefficients of two continuous variables x_{i1} and x_{i2}, respectively, for $i = 1, 2, ..., N$. The simulation comprises the following steps:

(1) Set the entries in the vector of model parameters, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$, equal to 1.
(2) Choose the sample size in the simulation as $N = 20, 50$, and 100.
(3) Generate the values of dummy variable I_i of size N, such that the probability of observing a success is 0.25.
(4) Generate the values of continuous variables x_{i1} and x_{i2} of size N from the standard normal distribution, such that their values are independent from each other and the values of dummy variable.
(5) By multiplying the values of the design matrix $x_i^T = (1, I_i, x_{i1}, x_{i2}) \in \mathbb{R}^{\mathrm{N}\times\mathrm{P}}$ and parameter vector $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T \in \mathbb{R}^{\mathrm{P}\times 1}$, calculate the linear predictor part of the model, $x_i^T\beta$, where $N = 20, 50$, or 100 and $P = 4$.
(6) Calculate the probability of success for each observation, $\pi_i = \frac{exp(x_i^T\beta)}{1+exp(x_i^T\beta)}$ for $i = 1, 2, ..., N$.
(7) Generate the values of the response using the success probabilities, that is, $y_i \sim \text{Bernoulli}(\pi_i)$ for $i = 1, 2, ..., N$.
(8) Check the model fit to detect separation in the data using the R package "brglm2".
    (a) If there is no separation problem in the data, return to Step 3.
    (b) If there is a separation problem in the data, obtain the estimates of model parameters using each estimation method elaborated in the previous section.
(9) Repeat Steps 3-8 until having a set of parameter estimates for S = 1000 samples, each of which containing separation problem.

(10) Calculate the values of the bias, precision, and accuracy measures for each method using the estimates obtained for these samples.

Note that the measures of bias, precision, and accuracy need to be calculated for each method, which are the method of maximum likelihood, Firth's method (with and without nonparametric bootstrapping), and Bayesian approach using Normal(0, 2.5), Cauchy(0, 2.5), Student-t(0, 2.5, df = 7), Log-F(1, 1), Log-F(2, 2), Ridge and Lasso priors.

5.2. **Bias, precision, and accuracy measures for evaluating performance.** The performance of the methods will be compared to each other using the measures of bias, precision, and accuracy given in Walther and Moore [32]. These measures are defined as:

$$\begin{aligned}
\text{Bias}_{\text{p}} &= \frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sp} - \beta_p),\\
\text{Precision}_{\text{p}} &= \frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sp} - \bar{\beta}_p)^2,\\
\text{Accuracy}_{\text{p}} &= \frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_{sp} - \beta_p)^2,
\end{aligned} \tag{7}$$

where $\bar{\beta}_p = \frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_{sp}$ and $\beta_j = 1$ for $s = 1, 2, ..., 1000$ and $p = 0, 1, 2, 3$. The Bias_{p} is the mean of the differences between parameter β_p and its estimate across $S = 1000$ samples. Similarly, $\text{Precision}_{\text{p}}$ is the mean of the squared differences between an estimate and its expected value (i.e., $\bar{\beta}_p$) in $S = 1000$ samples, which is calculated for each parameter, separately. The measure of accuracy for the pth parameter, $\text{Accuracy}_{\text{p}}$, is the mean of the squared differences between parameter β_p and its estimates across $S = 1000$ samples, which is a combination of Bias_{p} and $\text{Precision}_{\text{p}}$. Note that the term "bias" is directly related and the terms "precision" and "accuracy" are inversely related to their corresponding equations in (7). That is, a small value of Bias_{p} means a low bias, while small values of $\text{Precision}_{\text{p}}$ and $\text{Accuracy}_{\text{p}}$ imply high precision and accuracy when estimating model parameters.

Another accuracy measure that can be used to investigate the performance of methods on estimating model parameters is the mean squared error (MSE), representing the estimation error for each sample in the simulation. The MSE is the total mean squared error between all parameters and their estimates:

$$\text{MSE} = \frac{1}{P}\sum_{p=0}^{P-1}(\beta_p - \hat{\beta}_p)^2, \tag{8}$$

where $P = 4$ is the number of parameters in the model. The mean of MSE values across $S = 1000$ simulation samples can be used to compare the overall performance

of methods on estimating model parameters. A small value of MSE means a high overall accuracy when estimating model parameters.

5.3. **Simulation Results.** Table 1 displays Bias_p, $\text{Precision}_\text{p}$ and Accuracy_p values obtained from 1000 simulated data sets, each of which contains the separation problem. This table shows that the estimate of parameter β_1 often has a higher bias and a lower precision and accuracy than that of parameters β_2 and β_3, since dummy variables are more prone to suffer from separation than continuous variables. Because of the same reason, although increasing the sample size increases the precision when estimating each parameter, this reduces the bias and improves the accuracy only for parameters β_0, β_2, and β_3, but not for parameter β_1. It seems that Firth's penalized maximum likelihood estimation and Bayesian estimation using Log-F(2, 2) prior provide smaller biases and higher precision and accuracy measures when compared to other estimation methods. Similarly, these methods have smaller MSE values (higher overall accuracy measures) when compared to other methods (see Table 2). Moreover, both tables show that Bayesian estimation may not perform well with Ridge prior, since the corresponding estimates may have spuriously high precision and accuracy values (indicating low precision and accuracy for these estimates). However, the values in these tables are point estimates, and thus, a set of graphical visualizations are designed to facilitate the interpretation of the simulation results.

TABLE 1. Bias, precision, and accuracy measures for performance evaluation.

		PMLE				PMLE via NB				MCMC Normal(0, 2.5)			
Measure	N	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Bias_p	20	0.12	0.28	0.09	0.07	0.30	0.09	0.32	0.29	0.89	0.74	0.53	0.50
	50	0.01	1.57	0.02	0.04	0.19	1.61	0.24	0.26	0.28	1.93	0.19	0.19
	100	0.01	2.21	-0.02	0.01	0.07	2.23	0.06	0.09	0.14	2.47	0.06	0.09
$\text{Precision}_\text{p}$	20	0.81	2.49	0.99	1.13	0.90	2.89	1.16	1.19	1.77	1.26	1.37	1.35
	50	0.31	0.98	0.39	0.54	0.44	1.10	0.55	0.64	0.38	0.32	0.34	0.32
	100	0.10	0.32	0.12	0.11	0.12	0.36	0.15	0.14	0.12	0.16	0.14	0.12
Accuracy_p	20	0.82	2.57	1.00	1.13	0.99	2.90	1.26	1.27	2.57	1.81	1.65	1.60
	50	0.31	3.46	0.39	0.54	0.48	3.70	0.61	0.71	0.46	4.03	0.38	0.36
	100	0.10	5.22	0.13	0.11	0.13	5.35	0.16	0.15	0.14	6.24	0.14	0.13
		MCMC Cauchy(0, 2.5)				MCMC Student-t(0, 2.5, df = 7)				MCMC Log-F(1, 1)			
Measure	N	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Bias_p	20	1.23	2.02	0.97	0.93	0.95	0.91	0.61	0.57	0.63	1.23	0.51	0.47
	50	0.25	3.89	0.21	0.22	0.27	2.20	0.20	0.20	0.22	2.56	0.18	0.18
	100	0.12	4.85	0.06	0.09	0.14	2.80	0.07	0.09	0.12	3.23	0.06	0.09
$\text{Precision}_\text{p}$	20	4.09	6.08	4.38	4.90	2.07	1.59	1.67	1.66	1.13	1.85	1.35	1.35
	50	0.49	1.88	0.47	0.55	0.40	0.47	0.36	0.35	0.32	0.60	0.34	0.32
	100	0.12	0.88	0.14	0.13	0.12	0.24	0.14	0.12	0.12	0.38	0.14	0.12
Accuracy_p	20	5.60	10.15	5.33	5.76	2.98	2.41	2.04	1.98	1.53	3.35	1.62	1.57
	50	0.56	16.99	0.51	0.60	0.48	5.30	0.40	0.39	0.37	7.15	0.37	0.36
	100	0.14	24.39	0.15	0.14	0.14	8.08	0.15	0.13	0.13	10.81	0.14	0.13
		MCMC Log-F(2, 2)				Bayesian Ridge LR				Bayesian Lasso LR			
Measure	N	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Bias_p	20	0.29	0.36	0.10	0.08	4.32	2.10	3.47	3.86	0.45	-0.38	-0.45	-0.43
	50	0.15	1.40	0.05	0.04	0.28	1.85	-0.01	0.05	0.18	0.96	-0.28	-0.27
	100	0.10	2.05	0.01	0.03	0.10	2.83	-0.03	-0.01	0.07	2.45	-0.13	-0.10
$\text{Precision}_\text{p}$	20	0.51	0.61	0.57	0.59	607.5	135.9	382.9	602.4	1.00	1.09	1.03	1.33
	50	0.23	0.32	0.22	0.21	8.32	3.65	5.29	13.38	1.07	4.34	1.54	2.04
	100	0.10	0.21	0.11	0.10	0.11	0.80	0.14	0.12	0.10	0.89	0.14	0.13
Accuracy_p	20	0.59	0.73	0.59	0.60	626.2	136.3	394.9	617.3	1.20	1.24	1.23	1.52
	50	0.25	2.27	0.23	0.21	8.40	7.07	5.29	13.38	1.10	5.25	1.62	2.11
	100	0.11	4.41	0.11	0.10	0.12	8.78	0.14	0.12	0.11	6.89	0.16	0.14

TABLE 2. The overall accuracy measure (MSE) for performance evaluation.

N	PMLE $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$	PMLE via NB $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$	MCMC Normal(0, 2.5) $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$
20	1.38	1.61	1.91
50	1.18	1.37	1.31
100	1.39	1.44	1.66
N	MCMC Cauchy(0, 2.5) $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$	MCMC Student-t(0, 2.5, df = 7) $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$	MCMC Log-F(1, 1) $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$
20	6.71	2.36	2.02
50	4.66	1.64	2.06
100	6.20	2.13	2.80
N	MCMC Log-F(2, 2) $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$	Bayesian Ridge LR $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$	Bayesian Lasso LR $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$
20	0.63	750.4	1.30
50	0.74	8.53	2.52
100	1.18	2.29	1.82

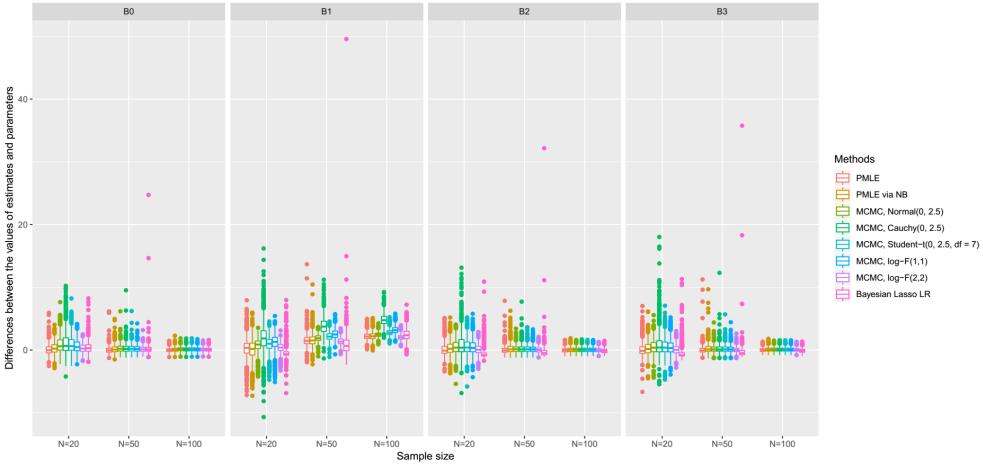

FIGURE 2. Boxplots used to interpret bias measures

Figures 2 and 3 display the differences between the values of estimates and parameters and the squared differences between the values of estimates and their expected values across the simulated data sets using varying sample sizes, which

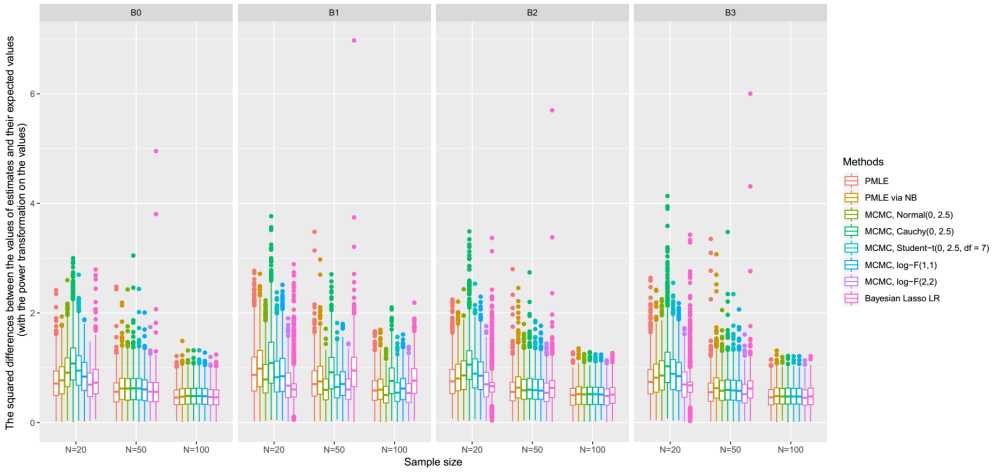

FIGURE 3. Boxplots used to interpret precision measures

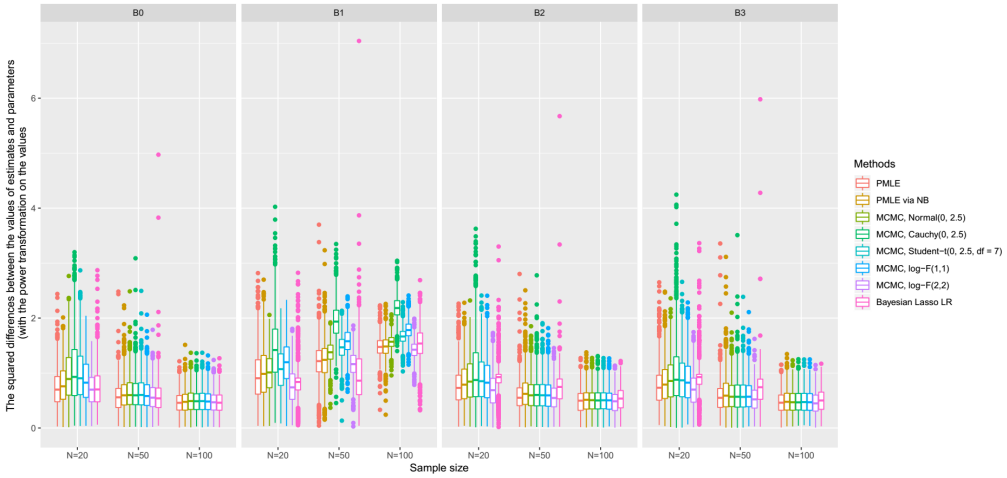

FIGURE 4. Boxplots used to interpret accuracy measures

are used to obtain the values of Bias_p and $\text{Precision}_\text{p}$ for each estimation method, respectively. [1] It seems that most of the methods perform well in terms of Bias_p and $\text{Precision}_\text{p}$ measures. However, Bias_p and $\text{Precision}_\text{p}$ measures of Bayesian

[1]The $\text{Precision}_\text{p}$, Accuracy_p and MSE values are always positive and they spread over large scales. Thus, the $y = x^{\frac{1}{4}}$ transformation is utilized on these values to better visualize and compare the performance of the methods (see [18, p. 12]).

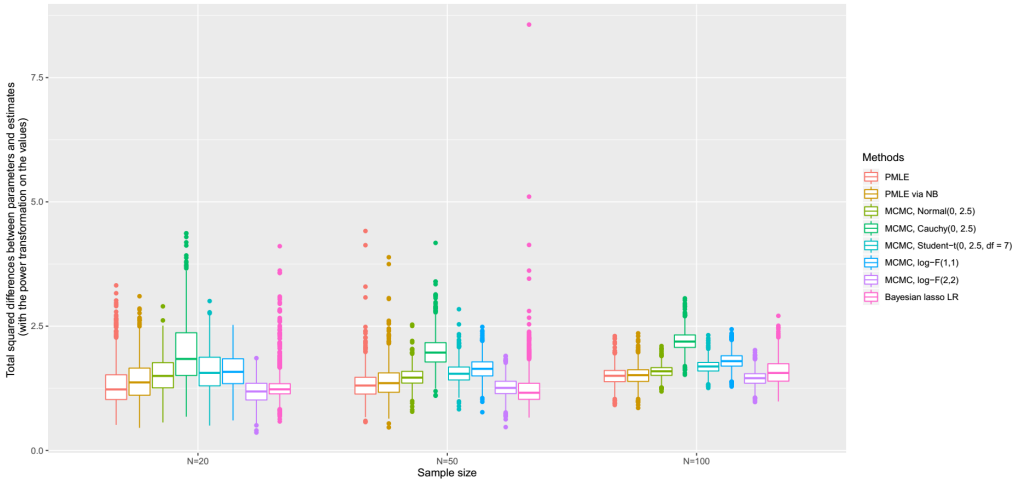

FIGURE 5. Boxplots used to interpret MSE values

estimation using Cauchy(0, 2.5) and Lasso priors have higher standard errors when compared to that of other methods under investigation. It seems that Bayesian estimation using log-F(2, 2) prior involves smaller amount of bias and have higher precision in estimating model parameters when compared to other methods. Note that the figures in the paper do not show the results for Bayesian estimation using Ridge prior, since this method produces spuriously high parameter estimates and their standard errors.

Figures 4 and 5 show the squared differences and the sums of squared differences between the values of estimates and parameters using varying sample sizes, which are utilized to obtain Accuracy_p and MSE values, respectively. Increasing the sample size improves the accuracy for each parameter, and thus, the total accuracy when estimating model parameters using each method. The estimates obtained by using Bayesian estimation with Log-F(2, 2) prior often have higher (total) accuracy measures, and thus, lower Accuracy_p and MSE values, when compared to other methods. Since nonparametric bootstrapping assumes an original sample that adequately represents the population of interest, the performance of Firth's method and Firth's method with nonparametric bootstrapping better resemble each other for large sample sizes (e.g., when $N = 100$). It seems that Bayesian approach with weakly informative Normal(0, 2.5) prior performs better than that with Student-t(0, 2.5, df = 7) or Log-F(1, 1) prior which in turn performs better than that with Cauchy(0, 2.5) prior. This result is in line with the suggestions made in Ghosh et al. [11], which state that Cauchy(0, 2.5) prior provides too much deficient information, and thus, instead of using this prior, Normal(0, 2.5) and Student-t(0, 2.5, df = 7) priors should be used when dealing with separation in logistic regression.

6. An example: Endometrial cancer data

A study in Heinze and Schemper [14] is used to illustrate how to analyze the data at hand under separation in logistic regression. In the study, the dichotomous outcome histology (HG: 0 = grade 0-II, 1 = grade III-IV) represents the histology of the endometrium by commonly accepted risk factors for endometrial cancer patients ($N = 79$). This outcome is predicted by the categorical variable neovasculization (NV: 0 = absent, 1 = present) and two continuous variables pulsatility index of arteria uterina (PI) and endometrium high (EH). The logistic regression model used to analyze the endometrial cancer data is:

$$f(\pi_i) = \beta_0 + \beta_1 \mathrm{NV}_i + \beta_2 \mathrm{PI}_i + \beta_3 \mathrm{EH}_i, \tag{9}$$

where $f(.)$ is the logit link function, β_0 is the intercept and β_1, β_2, and β_3 are the regression coefficients of variables NV, PI and EH, respectively, for $i = 1, 2, ..., 79$.

Since there is no observation in the endometrial cancer data for NV = 1 and HG = 0, the data suffer from quasi-complete separation, which has a detrimental effect on the estimate of parameter β_1 and its standard error when the estimation process is performed using the usual method of maximum likelihood. Therefore, Firth's method, Firth's method with nonparametric bootstrapping, Bayesian approach using Normal(0, 2.5), Cauchy(0, 2.5), Student-t(0, 2.5, df= 7), Log-F(1, 1), Log-F(2, 2), Ridge and Lasso priors are used to obtain parameter estimates and their standard errors (see Table 3).[2]

The estimates of parameters β_2 and β_3 across the methods are reasonably close to each other, while the estimates of parameters β_0 and β_1 across the methods may differ from each other. Figure 6 shows that the predicted probabilities of the outcome histology for some of the observations in the data are exactly equal to 1 (in the upper right corner of the plot), when using the method of maximum likelihood for estimation, which is a sign of the quasi-complete separation problem. Bayesian approach using the MCMC algorithm with Cauchy(0, 2.5) prior does not provide a convincing solution to the separation for endometrial cancer data, since some of the predicted probabilities of the outcome are (almost) equal to 1. The plots for other methods more closely resemble the regular logistic regression plot in which predicted probabilities are between the numbers 0 and 1.

Here, several diagnostics are introduced to inspect whether the MCMC algorithm produces adequate posterior samples for parameters when using weakly informative Normal(0, 2.5), Cauchy(0, 2.5), and Student-t(0, 2.5, df = 7) priors. The potential scale reducing factor ($\hat{R}$) and effective sample size (ESS) statistics for each parameter are used to determine whether the MCMC algorithm converges properly with high estimation accuracy. These statistics are obtained by inspecting multiple chains and dissimilarities between them (default number of chains is often 4). The $\hat{R}$ statistic shows whether the chains converge to the same area by exploring the

[2]For more details on obtaining the estimates of model parameters and their standard errors using R code for each estimation method see Supplementary material.

TABLE 3. Estimates and standard errors of the coefficients for the logistic regression.

	PMLE		PMLE via NB		MCMC Normal(0, 2.5)	
β	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}$	$SE(\hat{\beta})$
β_0	3.77	1.49	4.69	2.45	4.52	1.55
β_1	2.93	1.55	3.25	1.10	3.33	1.39
β_2	-0.03	0.04	-0.05	0.07	-0.04	0.04
β_3	-2.60	0.78	-3.11	1.29	-3.09	0.82
	MCMC Cauchy(0, 2.5)		MCMC Student-t(0, 2.5, df =7)		MCMC Log-F(1, 1)	
β	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}$	$SE(\hat{\beta})$
β_0	4.40	1.57	4.47	1.59	3.23	1.20
β_1	5.53	4.07	3.53	1.73	3.95	1.63
β_2	-0.04	0.04	-0.04	0.04	-0.02	0.04
β_3	-3.01	0.82	-3.08	0.85	-2.45	0.66
	MCMC Log-F(2, 2)		Bayesian Ridge LR		Bayesian Lasso LR	
β	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}$	$SE(\hat{\beta})$
β_0	2.40	1.04	3.86	1.59	3.53	1.49
β_1	3.10	1.31	4.64	4.59	3.60	3.15
β_2	-0.01	0.03	-0.03	0.04	-0.02	0.03
β_3	-2.06	0.59	-2.71	0.84	-2.58	0.82

ratio of their within and between variances. A value of $\hat{R} < 1.1$ indicates good convergence of the chains for the corresponding parameter. A high value of the ESS statistic indicates low autocorrelation and high estimation accuracy within the chains, where ESS > 1000 is often considered to be an adequate sample size statistic for many social scientists [25]. Table 4 displays the values of $\hat{R}$ and ESS statistics obtained for each parameter, where the MCMC algorithm is used with Normal(0, 2.5), Cauchy(0, 2.5), and Student-t(0, 2.5, df = 7) priors, respectively. The use of MCMC algorithm with Normal(0, 2.5) and Student-t(0, 2.5, df = 2.5) priors results in good convergence of the chains (i.e., $\hat{R} = 1$ for each parameter) with low autocorrelation, and consequently, high estimation accuracy (i.e., ESS > 1000 for each parameter). Although the MCMC algorithm with Cauchy(0, 2.5) prior produces good convergence of the chains for each parameter, there is a high autocorrelation and a low estimation accuracy within parameter samples, especially when looking at the relationship between the outcome and dichotomous predictor NV (i.e., ESS = 103 for parameter β_1). Thus, the focus from now on will be particularly on parameter β_1 to visually inspect the difference between the MCMC

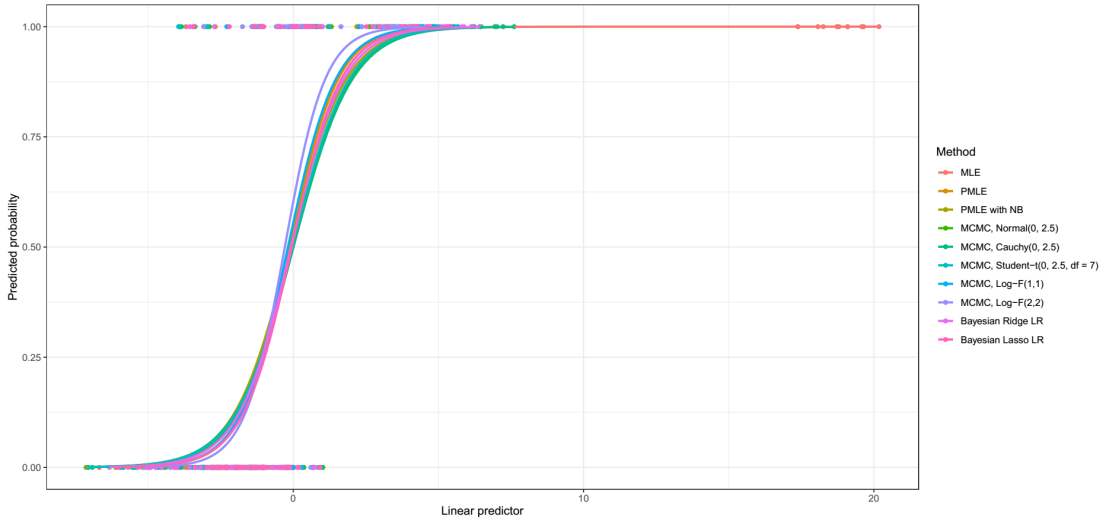

FIGURE 6. The values of linear predictor against predicted probabilities

algorithm with weakly informative Normal(0, 2.5), Cauchy(0, 2.5), and Student-t(0, 2.5, df = 7) priors.

TABLE 4. The $\hat{R}$ and ESS statistics for each parameter under Normal(0, 2.5), Cauchy(0, 2.5), and Student-t(0, 2.5, df = 7) priors.

	HMC Normal(0, 2.5)		HMC Cauchy(0, 2.5)		HMC Student-t(0, 2.5, df = 7)	
	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS
β_0	1.0	2620	1.0	1800	1.0	2303
β_1	1.0	2064	1.0	848	1.0	1752
β_2	1.0	3342	1.0	1988	1.0	3355
β_3	1.0	2280	1.0	1874	1.0	1933

Figure 7 shows the histograms of marginal posterior distribution, trace plot (chains separate), autocorrelation plot (combined chains) and log posterior for parameter β_1 under the three priors, respectively. A marginal posterior distribution is obtained for one single parameter by not taking other parameters in the model into account. The histograms show that the marginal posterior distribution of parameter β_1 is normal when using the normal prior and is close to be normal when using the Student-t prior with df = 7 degrees of freedom, for which the mean (solid

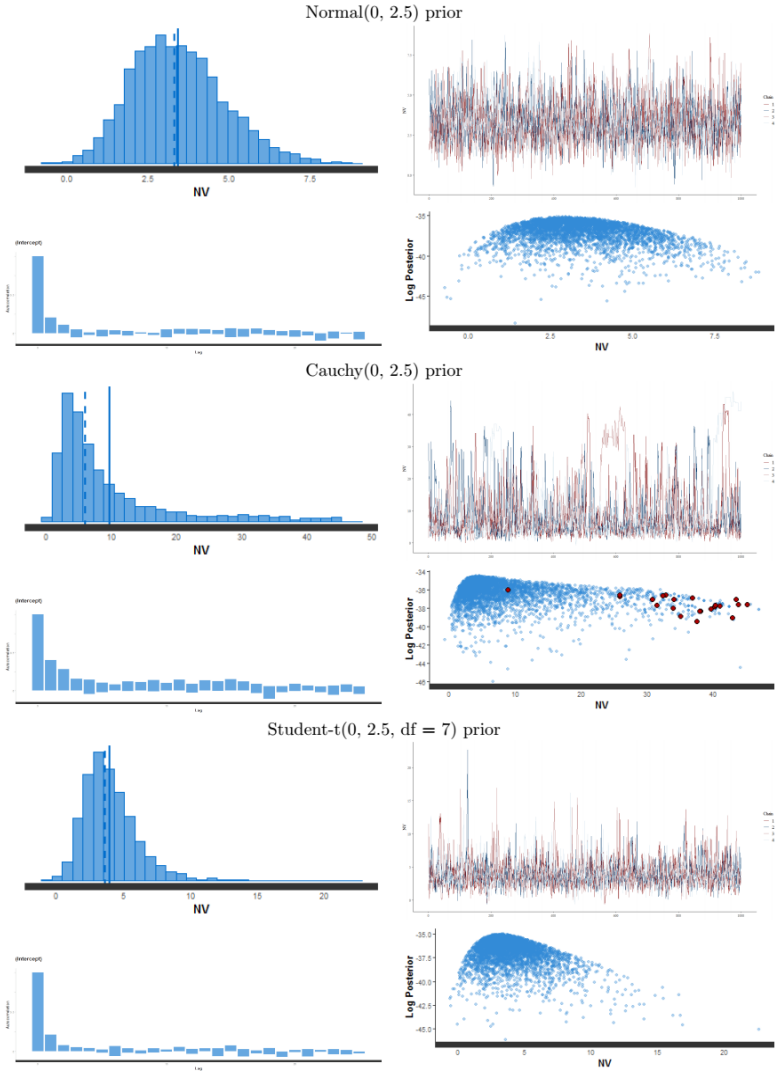

FIGURE 7. Marginal posterior distributions, trace and autocorrelation plots and log posteriors for parameter β_1 under weakly informative Normal(0, 2.5), Cauchy(0, 2.5), and Student-t(0, 2.5, df = 7) priors

line) and the median (dashed line) are (almost) equal to each other. The marginal posterior of parameter β_1 using the Cauchy prior has a right skewed (i.e., the

mean to the right of the median) distribution. By default, the MCMC algorithm in rstanarm utilizes 2000 posterior samples of parameter β_1 for each chain (i.e., 8000 samples in total), half of which are used in a warm-up phase and discarded later on before showing diagnostics and making inference. Thus, each of the four trace plots above under the three priors is created by using 1000 posterior samples of parameter β_1. Based on these plots, the chains display adequate mixing under Normal(0, 2.5) and Student-t(0, 2.5, df = 7) priors, but they may exhibit consecutive periods in positive direction under Cauchy(0, 2.5) prior. Based on the autocorrelation plots, independently from the prior distribution of parameter β_1, the correlation between variable NV and its value at lag zero is one, since the latter represents the variable itself. The height of spike at lag zero is quickly reduced to zero (and fluctuated around zero afterwards) with increasing values of lags under Normal(0, 2.5) and Student-t(0, 2.5, df = 7) priors for parameter β_1, respectively, which is a sign against autocorrelation. However, when using Cauchy(0, 2.5) prior for parameter β_1, the decrease in the height of spike at lag zero is relatively slow (and does not fluctuate considerably around zero) compared to that using Normal(0, 2.5) and Student-t(0, 2.5, df = 7) priors, which is a sign of positive autocorrelation.

The marginal posterior distribution for β_1 is highly curved when using the MCMC algorithm with Cauchy(0, 2.5) prior. This causes many divergent transitions in the MCMC algorithm, which are shown by the red points in the log posterior scatter plot above. This is evidence of too large step size in the MCMC algorithm under Cauchy(0, 2.5) prior. In this case, the results of MCMC algorithm should not be trusted. The MCMC algorithm needs a smaller step size to avoid divergent transitions and to draw plausible samples from the marginal posterior distribution of β_1, which can easily be adjusted by increasing the default value of δ parameter in rstanarm (e.g., from 0.95 to 0.99). Table 5 shows the estimates of parameters and their standard errors and the values of $\hat{R}$ and ESS statistics, when using Cauchy(0, 2.5) prior with divergent ($\delta = 0.95$) and non-divergent ($\delta = 0.99$) transitions, respectively. Based on this table, decreasing the step size in the MCMC algorithm by increasing the value of δ from 0.95 to 0.99 does not have much influence on parameter estimates and their standard errors. Moreover, increasing the value of δ results in a non-convergence (i.e., $\hat{R} = 1.1$ for parameter β_1) and a decrease in estimation accuracy (i.e., ESS is only 35 for parameter β_1). Therefore, it is not recommended to use this prior to overcome separation in the endometrial cancer data.

7. Discussion

Researchers in social sciences commonly use simple data manipulation techniques to overcome separation in logistic regression. These solutions are often unsatisfactory and do not meet the expectations of researchers. Thus, many researchers have been paying attention to more convenient approaches for estimation, such as

TABLE 5. Estimates and standard errors and the $\hat{R}$ and ESS statistics under Cauchy(0, 2.5) prior with divergent and non-divergent transitions.

	Divergent transitions $\delta = 0.95$				Non-divergent transitions $\delta = 0.99$			
β	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{R}$	ESS	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{R}$	ESS
β_0	4.53	1.66	1.0	929	4.40	1.75	1.0	258
β_1	5.99	4.54	1.0	103	6.05	4.61	1.1	35
β_2	−0.04	0.04	1.0	1088	−0.04	0.04	1.0	1400
β_3	−3.08	0.89	1.0	988	−3.01	0.90	1.0	234

symptotic and bootstrap-based bias reduction methods and Bayesian methods using weakly informative priors. However, the performance of these methods have not been fully investigated yet with respect to bias, precision, and accuracy measures in the context of logistic regression.

In the simulation, three methods were used to obtain the estimates of model parameters and their standard errors: Firth's penalized maximum likelihood estimation, Firth's method with nonparametric bootstrapping, and Bayesian approach with seven different priors. In a concrete real life example, parameter estimation was performed using these three methods for the endometrial cancer data. Supplementary material contains the relevant R code for obtaining the estimates of model parameters and their standard errors for each estimation method presented in this paper. Results of the simulation study and the analysis of the endometrial cancer data have showed that although most of the methods perform well in coping with the consequences of separation problem in logistic regression, Bayesian estimation with Log-F(2, 2) prior performs better than other methods.

The choice of prior distribution in Bayesian approach plays an essential role to overcome separation in logistic regression. It was shown both by the simulation and real life example that Bayesian approach with Cauchy(0, 2.5) or Ridge prior does not provide a reliable solution to separation in logistic regression, since these priors incorporate too much detrimental information into the analysis. A more coherent weakly informative prior such as Normal(0, 2.5), Student-t(0, 2.5, df = 7), Log-F(1, 1), Log-F(2, 2), or Lasso prior should be utilized in place of Cauchy(0, 2.5) prior when dealing with separation in the data.

REFERENCES

[1] Albert, A., Anderson, J. A., On the existence of maximum likelihood estimates in logistic regression models, *Biometrica*, 71 (1984), 1-10.

[2] Betancourt, M. J., Byrne, S., Livingstone, S., Girolami, M., The Geometric Foundations of Hamiltonian Monte Carlo. ArXiv e-prints 1410.5110, 2014.

[3] Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., Weidman, L., Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression, *Journal of the American Statistical Association*, 86 (1991), 68-78.
[4] Denwood, M. J., runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS, *Journal of Statistical Software*, 71 (2016), 1-25.
[5] Duane, S., Kennedy, A. D., Pendleton, B. J., Roweth, D., Hybrid Monte Carlo, *Physics Letters B*, 195 (1987), 216-222.
[6] Efron, B., Tibshirani, R. J. An introduction to the bootstrap, New York: Chapman & Hall, 1993.
[7] Firth, D., Bias reduction of maximum likelihood estimates, *Biometrika*, 80 (1993), 27-38.
[8] Fisher, R. A., On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society*, 222 (1922), 309-368.
[9] Gelman, A., Jakulin, A., Pittau, M. G., Su, Y., A weekly informative prior distribution for logistic and other regression models, *Annals of Applied Statistics*, 2 (2008), 1360-83.
[10] Geman, S. and Geman, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 (1984), 721-741.
[11] Ghosh, J., Li., Y., Mitra, R., On the use of Cauchy prior distributions for Bayesian logistic regression, *International Society for Bayesian Analysis*, 13 (2018), 359-383.
[12] Goodrich, B., Gabry, J., Ali, I., Brilleman, S., rstanarm: Bayesian applied regression modeling via Stan, *R package version 2.17.4,* 2018.
[13] Hastings, W., Monte Carlo sampling methods using Markov chains and their applications, *Biometrika,* (1970), 57, 97-109.
[14] Heinze, G., Schemper, M., A solution to the problem of separation in logistic regression, *Statistics in Medicine*, 21 (2002), 2409-2419.
[15] Hox, J. J., Multilevel analysis: Techniques and applications (2nd ed.), New York, NY: Routledge, 2010.
[16] Jeffreys, H., An invariant form of the prior probability in estimation problems, *Proceedings of the Royal Society of London A*, 186 (1946), 453-61.
[17] Lokhorst, J., The lasso and generalised linear models, Honors Project, University of Adelaide, Adelaide, 1999.
[18] Maciejewski, R., Data representations, transformations, and statistics for visual reasoning, *Synthesis Lectures on Visualization*, 2 (2011), 1-85.
[19] Makalic, E., Schmidt, D. High-dimensional Bayesian regularised regression with the BayesReg package, 2016, arXiv:1611.06649.
[20] Mansournia, M. A., Geroldinger, A., Greenland, S., Heinze, G., Separation in logistic regression: Causes, consequences, and control, *American Journal of Epidemiology*, 187 (2018), 864-870.
[21] McCullagh, P., Nelder, J., Generalized linear models (2nd ed.), Boca Raton, FL: Chapman & Hall / CRC, 1989.
[22] Mehta, C. R., Patel, N. R., Exact logistic regression: theory and examples, *Statistics in Medicine*, 14 (1995), 2143-2160.
[23] Metropolis, N., Rosenbuth, A., Rosenbuth, M., Teller, A., Teller, E., Equations of state calculations by fast computing machines, *The Journal of Chemical Physics*, 21 (1953), 1087-1092.
[24] Mood, C., Logistic regression: Why we cannot do what we think we can do, and what we can do about it, *European Sociological Review*, 26 (2010), 67-82.
[25] Muth, C., Oravecz, Z., Gabry, J., User-friendly Bayesian regression modeling: A tutorial with rstanarm an shinystan, *The Quantitative Methods for Psychology,* 14 (2018), 99-119.

[26] Neal, R., MCMC using Hamiltonian Dynamics, *In Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds. CRC Press, New York.), 2013.
[27] Newton, I., Philosophiae naturalis principia mathematica, Colonia Allobrogum: sumptibus CI. et Ant. Philibert, 1760.
[28] Ohkura, M., Kamakura, T. Test for a regression parameter in a logistic regression model under the small sample size and the high event occurrence probability, *Japanese Applied Statistics (in Japanese)*, 40 (2011), 41-51.
[29] Rainey, C., Dealing with separation in logistic regression models. *Political Analysis*, 24 (2016), 339-355.
[30] Roth, V., The generalized lasso,*IEEE Transactions on Neural Networks*, 15 (2004), 16-28.
[31] Schaefer, R. L., Roi, L. D., Wolfe, R. A., A ridge logistic estimator, *Communications in Statistics - Theory and Methods*, 13 (1984), 99-113.
[32] Walther, B. A., Moore, J. L., The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance, *Ecography*, 28 (2005), 815-829.
[33] Webb, M. C., Wilson, J. R., Chong, J., An analysis of quasi-complete binary data with logistic models: Applications to alcohol abuse data, *Journal of Data Science*, 2 (2004), 273-285.
[34] Yuan, K. H., Hayashi, K., Standard errors in covariance structure models: Asymptotic versus bootstrap. *British Journal of Mathematical and Statistical Psychology*, 59 (2006), 397-417.
[35] Zorn, C., A solution to separation in binary response models, *Political Analysis*, 13 (2005), 157-170.