Dealing with separation or near-to-separation in the model for multinomial response with application to childhood health seeking behavior data from a complex survey

Nowrin Nusrat and M. S. Rahman

Institute of Statistical Research and Training, University of Dhaka, Dhaka, Bangladesh

ABSTRACT
Separation or monotone-likelihood can be observed in fitting process of a multinomial logistic model using maximum likelihood estimation (MLE) when sample size is small and/or one of the outcome categories is rare and/or there is one or more influential covariates, resulting in infinite or biased estimate of at least one regression coefficient of the model. This study investigated empirically to identify the optimal data condition to define both ‘separation’ and ‘near-to-separation’ (partial separation) and explored their consequences in MLE and provided a solution by applying a penalized likelihood approach, which has been proposed in the literature, by adding a Jeffreys prior-based penalty term to the original likelihood function to remove the first-order bias in the MLEs of the multinomial logit model via equivalent Poisson regression. Furthermore, the penalized estimating equation (PMLE) is extended to a weighted estimating equation allowing for survey-weight for analyzing data from a complex survey. The simulation study suggests that the PMLE outperforms the MLE by providing smaller amount of bias and mean squared of error and better coverage. The methods are applied to analyze data on choice of health facility for treatment of childhood diseases.

1. Introduction
Separation or monotone-likelihood may occur in the fitting process of the multinomial logistic regression model using maximum likelihood estimation (MLE) procedure, resulting in an infinite estimate of at least one regression coefficients of the model. As described in many studies [5,15,16], separation occurs when the responses and non-responses are separated by a predictor, particularly binary predictor, or a non-trivial linear combination of several predictors. The phenomenon of separation is more likely to occur when the sample size is small, or number of response category is high or one of the categories is rare, or there is one or more strong predictors or combination of them. Let us consider an example (Table 1) describing different forms of separation caused by a dichotomous covariate (X) in
Table 1. Example of separation (left) or near-to-separation (right) due to a dichotomous predictor $X$ (covariate) against outcome $Y$. Number in each cell indicates number of observations.

|         | $Y$ 1 | $Y$ 2 | $Y$ 3 |
|---------|-------|-------|-------|
| $X = 0$ | 15    | 0     | 15    |
| $X = 1$ | 15    | 15    | 0     |

multinomial response $Y$. In the first case, the response is completely separated (more than one empty cell if we compare response categories 2 and 3) or quasi-completely separated (one empty cell if we consider response categories 1 and 3) by the predictor. In the second case, all are non-zero cells but there is at least one cell with a few observations, which we termed as ‘near-to-separation’ (partial separation). For multinomial response, we empirically determined a cutoff (threshold) of such observations, which is maximum 15% of the total sample in the data, to define the case as ‘near-to-separation’. Details of the empirical evaluation are discussed later in Section 3. Although complete or quasi-complete separation is rare in practice, the situation of ‘near-to-separation’ is more common. As reported by Mondol and Rahman [20] in correlated binary data, this form of separation often makes trouble to achieve convergence and yields bias in the estimate to some extent, and hence it cannot be ignored.

Such forms of separation can also be found in the multinomial data from a complex survey design, for example, in ‘childhood health seeking behavior’ data extracted from the database of the Bangladesh Demographic and Health Survey (BDHS) 2014. The BDHS employed two-stage stratified cluster sampling design to make a stratum-specific estimate of the outcome for some stratum such as urban and rural in addition to seven administrative divisions of the country, which may help the policy-makers for designing evidence-based policies. Therefore, when looking for a stratum-specific estimate of health seeking behavior (choice of health facilities: private/public/local-pharmacy, for treatment of acute respiratory infection (ARI)), one can observe the existence of both complete or quasi-complete separation (for both division and urban stratum) and near-to-separation (for rural stratum) created by the covariate ‘household economic condition’ (poorest/poorer/middle/richer/richest) calculated from household assets owned by the parents (Table 2). The MLE of the regression coefficient associated with covariate ‘economic condition’ in the multinomial logit model of the ‘choice of facilities’ fitted for the urban stratum (where separation exists) is reported with an extreme high value of the odds ratio (OR) and very wider Wald-based confidence interval, which lead to invalid inference [24]. For the rural stratum (where near-to-separation exists), the corresponding odds ratio value is also reported to be large in some cases. Similar results are observed for a divisional stratum (results not shown). These results suggest that maximum likelihood-based multinomial logit models for health seeking behavior data may be failed to provide consistent estimates and valid inference. The results are similar to those found in many other studies [4,12,15,16], which described consequences of separation in the MLE including frequent convergence failure, and even if it converges, it provides biased or often infinite estimates of at least one regression coefficient and standard error (SE) leading to misleading inference.

A number of studies discussed small sample bias adjustment in the ML estimates of the regression coefficient of the generalized linear model [2,3,6–8,10,21,23] while other
Table 2. Observed data and odds ratio of the three different population: urban, rural and both for the covariate economic status obtained from the fitted MLE (adjusted for other covariates) to the data on childhood health seeking behavior for treatment of ARI.

| Economic status | Urban | Rural | Both |
|-----------------|-------|-------|------|
|                 | Local | Public | Private | Local | Public | Private | Local | Public | Private |
| Observed data   |       |        |        |       |        |        |       |        |        |
| Poorest         | 6     | 3      | 0      | 57    | 13     | 10     | 63    | 16     | 10     |
| Poorer          | 3     | 5      | 4      | 42    | 10     | 18     | 45    | 15     | 22     |
| Middle          | 6     | 4      | 5      | 41    | 5      | 16     | 47    | 9      | 21     |
| Richer          | 14    | 7      | 9      | 18    | 4      | 11     | 32    | 11     | 20     |
| Richest         | 5     | 4      | 3      | 6     | 1      | 9      | 11    | 6      | 27     |
| Total           | 34    | 24     | 36     | 164   | 33     | 64     | 198   | 57     | 100    |
| Estimated parameter |       |        |        |       |        |        |       |        |        |
| Poorest         | (RC)  |        |        | (RC)  |        |        | (RC)  |        |        |
| Poorer          | 24.98 | $6.42 \times 10^5$ | 0.41 | 2.57 | 0.42 | 2.80 |
| Middle          | 16.91 | $9.54 \times 10^5$ | 0.17 | 1.51 | 0.21 | 1.96 |
| Richer          | 4.16  | $2.28 \times 10^5$ | 0.34 | 2.83 | 0.31 | 2.25 |
| Richest         | 4.24  | $1.12 \times 10^6$ | 0.39 | 5.69 | 0.31 | 6.82 |

Note: OR, odds ratio; RC, reference category.

studies discussed solution to the problem of separation [15,16]. Of them, Firth’s [10] proposal was based on reducing the first-order ($O(n^{-1})$) bias in the MLEs by adding a penalty term in the score function, which was equivalent to Jeffrey’s invariant prior. Heinze and Schemper [16] extended Firth’s penalized approach for solving the problem of separation in binary logistic regression. Firth’s approach was also extended to conditional logit [22], proportional odds [19], multinomial logistic regression [4] and non-linear GLM [17] for a small sample bias reduction. Furthermore, Kosmidis and Firth [18] explored the connection between multinomial logistic regression models and Poisson log-linear models for cell counts and proposed a penalized maximum likelihood estimators procedure to remove the first-order bias in the MLEs of the multinomial logit model using Poisson trick. Although small sample bias reduction and solution to separation has been proposed for multinomial outcome by Bull et al. [4] and later by Kosmidis and Firth [18] via an equivalent Poisson regression, none of them elaborately investigate the performance methods in the presence of separation, particularly ‘near-to-separation’, which is more common in practice. The present study investigates the consequences of all forms of separation in the maximum likelihood and penalized maximum likelihood approaches using an extensive simulation study and compared the results to provide some practical recommendations. In addition, the paper extended the penalized methods incorporating sampling weight for analyzing the above-mentioned health seeking behavior data extracted from a survey with complex design. The resulting weighted penalized method achieves convergence and obtains finite estimate of the regression in the presence of both separation and near-to-separation.

The paper is organized as follows. Section 2 reviews the existing maximum likelihood-based multinomial model and describes the penalized methods incorporating the survey-weight. The simulation study is described in Section 3 and an application of the methods for analyzing health seeking behavior data in Section 4. Finally, in Section 5, the main findings of the paper are discussed.
2. Methodology

2.1. Multinomial logit model

Let us consider a multinomial response \( Y \) with \( k \) categories and their corresponding probabilities \( \pi_1, \pi_2, \ldots, \pi_k \). Let \( q = k - 1 \) and \( \beta^T = (\beta_1^T, \beta_2^T, \ldots, \beta_q^T) \) be the vector of the \( pq \) model parameters. Assume that we have observed \( n \) pairs \( (y_i, x_i) \) with \( y_i = (y_{i1}, \ldots, y_{iq})^T \) the vector of observed frequencies which are realizations of a multinomially distributed random vector \( Y_i \) with index \( m_i \), and \( x_i \), a \( p \times 1 \) vector of known covariate values. The observed frequency from the \( k \)th category is \( y_{ik} = m_i - \sum_{s=1}^{q} y_{is} \) and \( \pi_i = (\pi_{i1}, \ldots, \pi_{iq})^T \) is the \( q \times 1 \) vector of the corresponding category probabilities. By definition, the probability of the \( k \)th category is \( \pi_{ik} = 1 - \sum_{s=1}^{q} \pi_{is} \). The matrix \( X \) with rows \( x_i^T \) is assumed to be of full rank and for the presence of an intercept parameter in the model, the first element of \( x_i \) is set to one for every \( i = 1, 2, \ldots, n \). Let \( Z_i = 1_q \otimes x_i^T \) be the \( q \times pq \) model matrix. The log-odds can be expressed as

\[
\log \frac{\pi_{is}}{\pi_{ik}} = \eta_{is} = \sum_{t=1}^{pq} \beta_t z_{ist} \quad (i = 1, 2, \ldots, n; s = 1, 2, \ldots, q),
\]

where \( z_{ist} \) is the \((s, t)\)th element of \( Z_i \) and \( 1_q \) is the \( q \times q \) identity matrix.

2.2. MLE procedure

According to the notations in Ref. [9], the multinomial log-likelihood can be written as

\[
l(\beta) = \sum_i \sum_{s=1}^{q} y_{is} \log \frac{\pi_{is}}{\pi_{ik}} + \sum_i m_i \log \pi_{ik}.
\]

The corresponding score function for the coefficient vector \( \beta \) has the following form:

\[
U(\beta) = \sum_i U_i(\beta) = q \sum_i Z_i^T (y_i - m_i \pi_i).
\]

The maximum likelihood estimator (MLE) for the regression coefficient is the solution to the equation \( U(\beta) = 0 \), for which Newton–Raphson approximation is commonly applied [9].

2.3. Review of Firth’s penalized method for GLM

To remove the first-order \( (O(n^{-1})) \) bias in the MLE of the \( t \)th regression parameter of GLM, Firth [10] proposed the following score function:

\[
U^*_t(\beta) \equiv U_t(\beta) + a_t(\beta) \quad (t = 1, 2, \ldots, pq)
\]

with the penalty term

\[
a_t(\beta) = \frac{1}{2} \text{trace}[I(\beta)^{-1} \{ \partial I(\beta)/\partial \beta_t \}] = \frac{\partial}{\partial \beta_t} \left[ \frac{1}{2} \log |I(\beta)| \right],
\]

where \( U_t(\beta) = \partial l(\beta)/\partial \beta_t = 0 \) is the standard score function based on the log-likelihood function \( l(\beta) = \log L(\beta) \), and \( I(\beta)^{-1} \) is the inverse of the corresponding information.
matrix \( I(\beta) = -\partial^2 l(\beta)/\partial \beta^2 \) evaluated at \( \beta \). The corresponding penalized log-likelihood and likelihood functions against the above-modified score equation are \( l^*(\beta) = l(\beta) + 1/2 \log |I(\beta)| \) and \( L^*(\beta) = L(\beta)|I(\beta)|^{1/2} \), respectively. The penalty function is known as Jeffreys invariant prior for this problem and its influence is asymptotically negligible. Bull et al. [4] applied Firth’s general idea to the multinomial logit models for finite sample bias reduction in the regression estimate. For multinomial logit model, the observed information matrix \( F(\beta) = Z^TWZ \), with \( W \) being a \( nq \times nq \) block-diagonal matrix with \( q \times q \) blocks \( W_i = \{ w_{isk} \}, w_{iss} = \pi_{is}(1 - \pi_{is}) \) for \( s = k \) and \( w_{isk} = -\pi_{is}\pi_{ik} \) otherwise. Then, the penalized score function given in Equation (3) can be replaced by the following form for the multinomial logit model:

\[
U^*_t(\beta) = U_t(\beta) + \frac{1}{2} \sum_i \sum_{s=1}^q \text{trace}\{H_iW^{-1}_iK_{is}\}z_{ist}
\]

\[
= \sum_i \sum_{s=1}^q \left( y_{is} - m_{i}\pi_{is} + \frac{1}{2} \text{trace}\{H_iW^{-1}_iK_{is}\} \right) z_{ist} \quad (t = 1, \ldots, pq),
\]

where \( K_{is} \) is a \( q \times q \) symmetric matrix with \( (u,v) \)th element, the third-order cumulants of \( Y_i \) and \( H_i \) denotes the \( i \)th diagonal block of the \( nq \times nq \) ‘hat’ matrix \( H = Z(Z^TWZ)^{-1}Z^TW \) consisting of \( n^2 \) blocks, each of dimension \( q \times q \). After some algebra, the above penalized score function takes the following form:

\[
U^*_t(\beta) = \sum_i \sum_{s=1}^q \left[ y_{is} + \frac{1}{2} h_{iss} - \left( m_{i} + \frac{1}{2} \text{trace}H_i \right) \pi_{is} - \frac{1}{2} \sum_{i=1}^q \pi_{is}h_{iis} \right] z_{ist}, \quad (4)
\]

where \( h_{iis} \) is the \( (s,u) \)th element of \( H_i \). The solution of the equation \( U^*_t(\beta) = 0 \) (from Equation (4)) is computationally tedious due to the complex structure of the hat matrix. Kosmidis and Firth [18] proposed an alternative penalized score function for the multinomial logit model using the equivalent Poisson trick, which is more flexible (simple and computationally efficient) for illustrative purpose. The details of this approach are discussed below.

### 2.4. Solving the problem of separation in multinomial response using equivalent Poisson regression

The multinomial model in Equation (1) can be embedded conveniently into a Poisson log-linear model. Using the indicator function \( \delta_{sk} \) (1, if \( s = k \) and 0 otherwise), the log-linear model for multinomial response can be written as (using the similar notations used in Ref. [18]):

\[
\log \mu_{is} = \{ \phi_i + (1 - \delta_{sk})x_i^T \beta_s \},
\]

\[
\mu_{is} = \exp(\phi_i + (1 - \delta_{sk})x_i^T \beta_s) \quad (i = 1, 2, \ldots, n; s = 1, 2, \ldots, k), \quad (5)
\]

where \( \mu_{is} \) are the expectations of the independent Poisson random variables \( Y_{is} \), \( \tau_i = \sum_{s=1}^k \mu_{is} \) and \( \phi_i \) nuisance parameters. As described by Kosmidis and Firth [18], both the multinomial logit model given in Equation (1) and the Poisson log-linear model in
Equation (5) are full exponential families and hence Firth’s [10] bias-reduced penalty term can be directly applied to the likelihood of each model. However, they argued that under the given parameterization \( \theta^\dagger = (\beta^T, \tau^T)^T \), the Firth-type penalized Poisson likelihood cannot be factorized as the product of the required multinomial likelihood (for \( \beta \)) and a factor free of \( \beta \). Therefore, standard computation of bias-reduced estimates for the full parameter vector \( \theta^\dagger \) in the Poisson log-linear model does not deliver bias-reduced estimates of \( \beta \) of the multinomial model. Kosmidis and Firth [18] provided a solution to work with the restricted version of the Poisson model derived by imposing the constraint \( m_i = \tau_i \). For more details about this constraint, see Ref. [18]. Under this constraint, the Poisson model in Equation (5) becomes into a canonically linked generalized non-linear model of the form:

\[
\log \mu_{is} = \log \left( \tau_i \left\{ 1 + \sum_{u=1}^{q} \exp(x_i^T \beta_u) \right\}^{-1} + (1 - \delta_{sk})x_i^T \beta_s. \right.
\]

To simplify the computational algorithm of the likelihood score equation of the above non-linear model, Kosmidis and Firth [18] applied the general results of the penalized likelihood of the generalized non-linear model (Equation (13)) derived in Ref. [17] and proposed the following bias-reducing adjusted score functions for the above model:

\[
U^\dagger_t = \sum_{i=1}^{n} \sum_{s=1}^{k} \left[ y_{is} + \frac{1}{2} h_{iss} + \frac{1}{2} \mu_{is} \text{tr}( (F^\dagger)^{-1} D^2(\xi_{is}; \theta^\dagger) ) - \mu_{is} \right] z^\dagger_{ist} \quad (t = 1, \ldots, n + pq),
\]

where \( F^\dagger \) is the expected information on \( \theta^\dagger \), \( D^2(\xi_{is}; \theta^\dagger) \) denotes the \((n + pq) \times (n + pq)\) Hessian matrix of \( \xi_{is} \) with respect to \( \theta^\dagger \), and \( z^\dagger_{ist} \) is the \((s, t)\)th component of the \( k \times (n + pq)\) matrix

\[
Z^\dagger_i = \begin{bmatrix}
G_i - 1_q \otimes (\pi^T_i G_i) \\
- \pi^T_i G_i G_i \\
\end{bmatrix} \begin{bmatrix}
1_q \otimes (\tau_i^{-1} e_i^T) \\
- \tau_i^{-1} e_i^T \\
\end{bmatrix} 
\quad (i = 1, \ldots, n),
\]

with \( \pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iq})^T \) and \( \pi_{is} = \mu_{is}/\tau_i \). After noting that \( D^2(\xi_{is}; \theta^\dagger) \) does not depend on \( s \) and substituting for \( z^\dagger_{ist} \), Kosmidis and Firth [18] proposed the following simple form of the adjusted score functions for \( \beta \):

\[
U^\dagger_t = \sum_{i=1}^{n} \sum_{s=1}^{k} \left[ y_{is} + \frac{1}{2} h_{iss} - \left\{ m_i + \frac{1}{2} \text{tr}(H_i) \right\} \pi_{is} \right] g^\dagger_{ist} \quad (t = 1, \ldots, pq),
\]

where \( g^\dagger_{ist} \) is the \((s, t)\)th component of \( G_i \). Only the quantity \( h_{iss} \) in the above bias-reducing score equation will be affected due to the constraint \( \tau_i = m_i \). Kosmidis and Firth [18] also proved using Theorem 1 that these adjusted score functions are identical to those (Equation (4)) obtained directly from the penalization of the likelihood for the model in Equation (1). The penalized estimating equation (PMLE) estimates of \( \beta \) can be estimated using the following standard iterative process proposed by Kosmidis and Firth [18]. New values \( \beta^{(j+1)} \) obtained from the candidate \( \beta^{(j)} \) by solving the following iterative equation:
with $\tilde{h}_{iss}$ calculated for the restricted parameterization. As suggested by Kosmidis and Firth [18], Equation (7) can be implemented by the following steps:

(i) set $\tilde{\phi}_i^{(j)} = \log m_i - \log \{1 + \sum_{s=1}^{q} \exp(x_i^T \beta_s^{(j)})\}$.
(ii) use $\tilde{\theta}^{(j)} = (\beta_1^{(j)}, \ldots, \beta_n^{(j)})$ to calculate new values $\tilde{H}_i^{(j)}$.
(iii) fit model in Equation (5) by maximum likelihood but using the adjusted responses $y_{is} + \tilde{h}_{iss}/2$ in place of $y_{is}$ to get new estimates $\phi^{(j+1)}$ and $\beta^{(j+1)}$.

The $\beta$-block of the inverse of the expected information matrix evaluated at the reduced-bias estimates (termed as PMLE) can be used to produce valid SEs for the estimators.

The penalized method using the Poisson trick presented above was originally proposed for finite sample (first-order) bias reduction in the MLE of the multinomial model. However, these bias-reducing score functions in Equation (6) can be used for solving the problems due to separation and near-to-separation discussed earlier in the multinomial response. Because the functions given in Equation (6) guarantee finite estimates of $\beta$ even while there is empty cell in one or more response categories, which was not possible by the standard score functions (Equation (2)) derived without penalization. This idea is similar to those proposed by Heinze and Schemper [16] for solving the problem of separation in the logistic regression. The corresponding estimate of SEs of the PMLE can be obtained from the roots of the diagonal elements of the hat matrix $H_i$.

### 2.5. Weighted PMLE for complex survey data

The main motivation of the paper is to apply the above method to analyze data (mentioned earlier) extracted from a cross-sectional survey with stratified cluster design (a multi-stage design), which is a complex design. Therefore, incorporating individual sampling weight calculated from the inverse probability of selection in estimation of the model is extremely required to have accurate estimate of the regression parameters and associated SE. Let $Y_{hijk}$ be a multinomial response with $k$ categories for the $i$th subject ($i = 1, 2, \ldots, n_{hi}$) from the $j$th cluster ($j = 1, 2, \ldots, m_{hj}$) and from the $h$th stratum ($h = 1, 2, \ldots, H$). Again, let $\delta_{hji}$ be an indicator variable with value 1 if the subject $hji$ is selected in the sample and 0 otherwise and $P(\delta_{hji} = 1)$ be the probability of being selected into the survey, which is fixed by study design and may depend on covariates or additional variables such as screening variables not in the Poisson model for multinomial response of interest [11]. This probability can be calculated by multiplying the probability of selection in each stage in the multi-stage sampling scheme. Therefore, each subject in the sample has known weight $w_{hji} = \delta_{hji}/p_{hji}$. Incorporating the sampling weight, the penalized estimating function given in Equation (6) becomes to the following weighted estimating function for survey data:

$$U_t^{+, w} = \sum_{h} \sum_{j=1}^{m_h} \sum_{i=1}^{n_{hi}} \sum_{s=1}^{k} \left[ y_{hjis} + \frac{1}{2} h_{hjiss} - \left\{ m_{hji} + \frac{1}{2} \text{tr}(H_{hji}) \right\} \pi_{hjis} \right] g_{hjist} w_{hji}$$

\[ (t = 1, \ldots, pq). \]
Solving $U_i^{*,w} = 0$ provides weighted PMLE for the survey data. The equation can be solved using the same procedure as described earlier and the SE of the weighted PMLE can be obtained from the squared root of the robust variance estimator.

3. Simulation study

An extensive simulation study was performed to investigate the consequence of separation and near-to-separation in MLE for multinomial model and evaluate the performance of PMLE over MLE. We performed simulation considering data from a study with simple random sampling.

3.1. Simulation design

3.1.1. Data generation

Let us consider multinomial data $y_{is}$ for the $i$th observation and $s$th response category ($i = 1, 2, \ldots, n; s = 1, 2, \ldots, k$) with $p = 2$ covariates, one of which was considered as binary and other one as continuous. The realizations $x_{ib}$ of the binary covariate $X_b$ were generated from Bernoulli distribution with probability of event $\gamma$ and $x_{isc}$ of the continuous covariate $X_c$ from standard normal distribution. We considered a multinomial response with three response categories and the first response category as the reference category, then the multinomial responses $y_i = (y_{i1}, y_{i2}, y_{i3})$ were generated using the multinomial distribution with the probabilities derived from the following model:

$$
\pi_{i1} = \frac{1}{1 + \exp(\beta_{02} + \beta_{12b}x_{ib} + \beta_{22c}x_{isc}) + \exp(\beta_{03} + \beta_{13b}x_{ib} + \beta_{23c}x_{isc})},
$$

$$
\pi_{i2} = \frac{\exp(\beta_{02} + \beta_{12b}x_{ib} + \beta_{22c}x_{isc})}{1 + \exp(\beta_{02} + \beta_{12b}x_{ib} + \beta_{22c}x_{isc}) + \exp(\beta_{03} + \beta_{13b}x_{ib} + \beta_{23c}x_{isc})},
$$

$$
\pi_{i3} = \frac{\exp(\beta_{03} + \beta_{13b}x_{ib} + \beta_{23c}x_{isc})}{1 + \exp(\beta_{02} + \beta_{12b}x_{ib} + \beta_{22c}x_{isc}) + \exp(\beta_{03} + \beta_{13b}x_{ib} + \beta_{23c}x_{isc})}.
$$

The probabilities corresponding to the multinomial response categories satisfied the condition $\pi_{i1} + \pi_{i2} + \pi_{i3} = 1$. The values of $\beta_{02}$ and $\beta_{03}$ determine the overall prevalence of the response category in the data, and the values of $\beta_{12b}$, $\beta_{22c}$, $\beta_{13b}$ and $\beta_{23c}$ represent the effect size indicating the strength of association between the covariate and the response.

3.1.2. Simulation scenarios

Using the above settings, data were generated for several simulation scenarios with the evidence of separation or near-to-separation. To create separation, we treated binary covariate $X_b$ as an influential covariate by considering a relatively large value of $\beta_{12b}$ and $\beta_{13b}$ (compared to $\beta_{22c}$ and $\beta_{23c}$ associated with $X_c$) so that it creates any form of separation defined earlier (Table 1). This setting confirmed complete or quasi-complete separation (at least one zero cell in a contingency table between the binary covariate $X_b$ and response $Y$) for some simulated datasets and created near-to-separation (there is at least one cell with a few observations in the contingency table) for most of the other simulation sets. In order to determine the optimal cutoff point of the cell with few observations, we initially performed an extra simulation study before conducting the final simulation by considering
Figure 1. Bias and MSE of the MLE and PMLE estimates of $\beta_{12b}$ and $\beta_{13b}$ (coefficients of binary covariate that made near-to-separation) over percentage of cell frequency for both $N = 50$ and $N = 100$ number of observations.

different cutoff points as the percentage of observations of the total sample, which are 5%, 8%, 10%, 12%, 15%, 18% and 20%. The idea is to determine the optimal cutoff point (as percentage) for which there is negligible bias in the estimate of the regression coefficient. We then consider this cutoff point as the rule of thumb below which one can consider the dataset with evidence of near-to-separation. This simulation was performed for scenarios with the true values of the coefficients associated with the binary covariate $\beta_{12b} = 1.3$, $\beta_{13b} = 1.2$ and sample size $n = 50$ and 100. For each scenarios, we plot the bias of the estimated regression coefficient over 1000 simulations. The empirical result showed that the bias in both the MLE and PMLE is near to zero for cell frequency greater than 15% of the total sample (Figure 1). We then defined near-to-separation in the final simulation if there is at least one cell with observations less than 15% of the total sample.

For the final simulation, we considered several scenarios created by varying the sample size $n$ and degree of imbalance in the prevalence of the response category, proportion of event in binary covariate ($\gamma$), magnitude of the log-odds ratio relating the binary covariate and the response ($\beta_{12b}, \beta_{13b}$). We considered the sample size $n$ as 15, 20, 30, 50, 100 and 200. Under each sample size scenario, first for fixed $\gamma = 0.5$ and $(\beta_{02}, \beta_{22c}, \beta_{03}, \beta_{23c}) = (-0.6, 0.65, -0.5, 0.5)$, we varied the value of $\beta_{12b}$ as 0.3, 0.9, 1.1, 1.6
and $\beta_{13b}$ as 0.4, 0.8, 1.2, 1.5, to see if the magnitude of the effect size affects the likelihood of separation and the estimates. Second, for fixed $\beta = (\beta_{02}, \beta_{12b}, \beta_{22c}, \beta_{03}, \beta_{13b}, \beta_{23c}) = (-0.6, 1.3, 0.65, -0.5, 1.2, 0.5)$, we varied the proportion of event in the binary covariate $X_b$ with the values of $\gamma$ as 0.2, 0.3, ..., 0.8. Third, for fixed $\gamma = 0.5$ and $(\beta_{12b}, \beta_{22c}, \beta_{13b}, \beta_{23c}) = (1.3, 0.65, 1.2, 0.5)$, we varied the value of $\beta_{02}$ as $-1.0, -0.1, 0.1, 0.3, 0.5, 0.6, 1.0, 1.5$ and $\beta_{03}$ as $-0.5, -0.2, 0.2, 0.1, 0.4, 0.8, 1.25, 2.0$, to see how likely the separation or near-to-separation occur in the data and if there is any difference in the performance of both MLE and PMLE.

The above simulation scenarios were considered for two cases based on the distribution of the response category. The case 1 is termed as ‘homogeneous distribution of the response category’, where the percentage distribution of the response categories are approximately equal (approx. 33% in each category), and case 2 is termed as ‘heterogeneous distribution of the response category’, where such distribution of the response category is unequal. These two cases were considered by fixing such a value of $\beta_{02}$ and $\beta_{03}$ which control the overall prevalence of the response category in the data.

### 3.2. Fitting the model and evaluating the performance

For each simulation scenario, we created 1000 replications of dataset and fitted the model in Equation (1) using MLE and the model in Equation (5) using PMLE to each of the simulated datasets. For both the MLE and PMLE estimates, we calculated the bias as $\text{bias}(\hat{\beta}_r) = \hat{\beta}_r - \beta_r$, where $\hat{\beta}_r = \frac{1}{1000} \sum^{1000}_{s=1} \hat{\beta}_{rs}$, the mean squared of error (MSE) as $\text{MSE}(\hat{\beta}_r) = \frac{1}{1000} \sum^{1000}_{r=1} (\hat{\beta}_{rs} - \beta_r)^2$. We also reported analytical SE as the average of the analytical estimates over the number of simulations, and the simulated standard error (SimSE) as the standard deviation of the estimates over the number of simulations. The coverage of the confidence interval for both MLE and PMLE estimators was calculated as the percentage of confidence intervals which include the true value over 1000 simulations. In all aspects of performance, we reported the number of convergence failure and summarized the results over the simulation for which convergence achieved.

All computations were performed using R version 3.5.2. For the standard MLE, the function ‘multinom’ belonging to R library ‘nnet’ was used, and for PMLE the function ‘brmultinom’ belonging to R library ‘brglm2’ was used.

### 3.3. Results

We first summarized likelihood of separation (complete or quasi-complete) or near-to-separation (in terms of percentage) over 1000 simulations under different scenarios considered (supplementary Table A.1). The results revealed that the likelihood of separation was positively correlated with degree of uneven proportions of event and non-event in the binary covariate ($\gamma$) and the magnitude of log-odds ratio $(\beta_{12b}, \beta_{13b})$. However, the chances of separation (complete or quasi-complete) decreased with the increasing sample size. The opposite scenarios can be observed for the near-to-separation. The results suggest that the separation is likely to occur for small sample or even for large sample with rare outcome along with extremely uneven distribution of event and non-event in binary covariate.
Table 3. The estimates of MLE and PMLE in the presence of complete or quasi-complete separation. The estimates are the mean over the number of simulations in which complete or quasi-complete separation occurred out of 1000 simulations under sample size $N = 20, 30$ and $\gamma = 0.5$.

| Sample size | True coefficient | Methods | N.Sim | N.Conv.failure | Estimate | SE | SimSE |
|-------------|-----------------|---------|-------|----------------|----------|----|-------|
| 20          | $\beta_{12b} = 1.3$ | MLE     | 215   | 19             | 5.66     | 62.91 | 26.83 |
|             |                 | PMLE    | 0     | 2.10           | 1.72     | 1.38 |
|             | $\beta_{13b} = 1.2$ | MLE     | 10    | 7.55           | 59.30    | 5.91 |
|             |                 | PMLE    | 0     | 2.20           | 1.73     | 1.48 |
|             | $\beta_{22c} = 0.65$ | MLE    | 10    | 1.13           | 1.34     | 2.02 |
|             |                 | PMLE    | 0     | 0.61           | 0.82     | 0.81 |
|             | $\beta_{23c} = 0.5$ | MLE     | 5     | 0.93           | 1.32     | 1.99 |
|             |                 | PMLE    | 0     | 0.50           | 0.82     | 0.86 |
| 30          | $\beta_{12b} = 1.3$ | MLE     | 62    | 8              | 8.88     | 74.85 | 5.80 |
|             |                 | PMLE    | 0     | 2.77           | 1.60     | 1.33 |
|             | $\beta_{13b} = 1.2$ | MLE     | 7     | 7.06           | 61.83    | 6.75 |
|             |                 | PMLE    | 0     | 2.29           | 1.51     | 1.57 |
|             | $\beta_{22c} = 0.65$ | MLE    | 0     | 0.93           | 0.81     | 0.95 |
|             |                 | PMLE    | 0     | 0.63           | 0.68     | 0.68 |
|             | $\beta_{23c} = 0.5$ | MLE     | 0     | 0.74           | 0.78     | 0.97 |
|             |                 | PMLE    | 0     | 0.54           | 0.66     | 0.70 |

Note: $\gamma$, proportion of events in binary covariate; N.Sim, number of simulation set in which complete or quasi-complete separation occurs; N.Conv.failure, number of simulation set which fail to converge; SimSE, simulated standard error.

### 3.3.1. Case 1: homogeneous distribution of the response categories

For the parameter set $\beta = (\beta_{02}, \beta_{12b}, \beta_{22c}, \beta_{03}, \beta_{13b}, \beta_{23c}) = (-0.6, 1.3, .65, -0.5, 1.2, 0.5)$ and for $\gamma = 0.5$, the percentage distribution of the response category is approximately equal even if the sample size vary. Under this homogeneous condition, the results revealed that in the presence of separation (complete or quasi-complete), the MLE either failed to achieve convergence or, if achieved, provided large estimates of the regression coefficients ($\beta_{12b}, \beta_{22c}, \beta_{13b}, \beta_{23c}$) (Table 3). The number of convergence failure decreased with the increasing sample size. The MLE also reported large SE, particularly for the regression coefficients ($\beta_{12b}, \beta_{13b}$) associated with the binary covariate ($X_b$), which created separation. In contrast, the PMLE achieved convergence for all these cases and provided finite estimate of the regression coefficients. In particular, the PMLE slightly overestimated the true value of the regression coefficients ($\beta_{12b}, \beta_{13b}$) associated with $X_b$ and underestimated the true value of those associated with $X_c$. Similar results can be observed for all other simulation scenarios with evidence of separation (results not shown). As the MLE showed convergence failure, large amount of bias in the estimate and high SE, we therefore restricted to evaluate the other properties such as MSE and coverage of the confidence interval for these scenarios.

For the simulations with evidence of near-to-separation, the amount of bias and MSE for the coefficients associated with the binary covariate decreased with increasing sample size, which is true for both the MLE and PMLE (Figure 2). However, the amount of both bias and MSE were observed to be low for the PMLE compared to those associated with the MLE. The amount of bias and MSE for both the MLE and PMLE slightly increased with increasing value of the true regression coefficients ($\beta_{12b}$ and $\beta_{13b}$) associated with the binary covariate (Figure 3). The PMLE showed substantially low amount of bias and MSE compared to the MLE. The bias and MSE for the both estimates were observed to be slightly lower when the proportion of event was around 0.5 (Figure 4). As usual, the PMLE
Figure 2. Bias and MSE of the MLE and PMLE estimates of $\beta_{12b}$ and $\beta_{13b}$ (coefficients of binary covariate that made near-to-separation) over sample size, where Max.MCE is the maximum Monte Carlo error.

showed smaller amount of bias and MSE than the MLE. Furthermore, we evaluated the performance of MLE and PMLE by increasing the number of binary covariate ($K$). The results suggested that the amount of bias and MSE in both the MLE and PMLE increased with increasing number of binary covariate but were greater for MLE than PMLE (Figure 5).

For the regression coefficients ($\beta_{22c}, \beta_{23c}$) of the continuous covariate which is not responsible for creating near-to-separation, both the MLE and PMLE, in general, provided smaller amount of bias and MSE compared to those associated with the binary covariate that created near-to-separation (Figure 6). Here, the MLE reported bias and MSE to some extent when sample size is small and the amount of bias and MSE decreased with increasing sample size. In contrast, the PMLE showed negligible amount of bias and MSE, even for small sample.

Some additional results (analytical SE, SimSE, width of confidence interval and coverage) of the MLE and PMLE from the simulation sets with evidence of near-to-separation were summarized in Table 4. The results revealed that in the presence of near-to-separation, both MLE and PMLE achieve convergence and provide finite estimates for the regression coefficients. However, the MLE provided large estimate of the SE (but smaller than the simulation SE) and wider confidence interval compared to those associated with PMLE.
Figure 3. Bias and MSE of the MLE and PMLE estimates of $\beta_{12b}$ and $\beta_{13b}$ (coefficients of binary covariate that made near-to-separation) over the magnitude of true coefficient value of binary covariate for $N = 50$, where Max.MCE is the maximum Monte Carlo error.

3.3.2. Case 2: heterogeneous distribution of the response categories

For the parameter set $\beta = (\beta_{02}, \beta_{12b}, \beta_{22c}, \beta_{03}, \beta_{13b}, \beta_{23c}) = (0.3, 1.3, .65, 0.1, 1.2, 0.5)$ and for $\gamma = 0.5$, the percentage distribution of the response category was heterogeneous (unequal) even if the sample size vary. Under this heterogeneous condition, similar pattern of the results can be observed here in the presence of complete or quasi-complete separation with convergence failure and infinitely large estimates for the MLE of the regression coefficients particularly that associated with the binary covariate (supplementary Table A.2). In contrast, the PMLE showed greater improvement by achieving convergence and providing finite estimate. The number of simulations with complete or quasi-complete separation was greater under heterogeneous case than the homogeneous case. The number of simulations with such conditions decreased with the increasing sample size. Similar results were also observed for all other simulation scenarios with evidence of complete or quasi-complete separation (results not shown).

In the presence of near-to-separation, the results suggested that the amount of MSE in both the MLE and PMLE decreased with increasing sample size and MSE were observed to be low for the PMLE estimate compared to those associated with MLE estimate (supplementary Figure A.1). Again, the amount of MSE is observed to be substantially high for
Figure 4. Bias and MSE of the MLE and PMLE estimates of $\beta_{12b}$ and $\beta_{13b}$ (coefficients of binary covariate that made near-to-separation) over the magnitude of proportion of binary covariate for $N = 50$, where Max.MCE is the maximum Monte Carlo error.

the MLE compared to the PMLE for any value of the true effect size and the proportion of the binary covariate. For the regression coefficient of the continuous covariate ($\beta_{22c}, \beta_{23c}$) that is not responsible for creating near-to-separation, both the MLE and PMLE, in general, provided smaller amount of MSE compared to those associated with the coefficient of the binary covariate ($\beta_{12b}, \beta_{13b}$) that created near-to-separation. Here, the PMLE also provided comparatively smaller amount of MSE than the MLE (supplementary Figure A.2). The MSE also decreased with increasing sample size and for large sample, the MLE and PMLE provide comparable results. Like the results for homogeneous distribution of the response category, similar pattern of the results can be observed here with the large value of SE, wider confidence interval for the MLE in comparison with the PMLE (supplementary Table A.3).

4. Application for analyzing health seeking behavior data

4.1. Data and variables

The methods are applied to analyze the ‘childhood health seeking behavior’ data extracted from the database of the BDHS 2014 conducted between June 2014 and November 2014.
Figure 5. Bias and MSE of the MLE and PMLE estimates of $\beta_{12b}$ and $\beta_{13b}$ (coefficients of binary covariate that made near-to-separation) over no. of binary covariates, where Max.MCE is the maximum Monte Carlo error. *Higher MSE value which is near 100.

through the collaborative efforts of the National Institute of Population Research and Training (NIPORT), ICF International (USA), and Mitra and Associates as part of the world-wide demographic and health survey (DHS) program. The BDHS is a nationally representative cross-sectional survey that employed two-stage stratified cluster sampling scheme, where in the first stage, enumeration areas (EA) were randomly selected (75% from rural area and 25% from urban area) from each of the seven administrative divisions, and in the second stage, households were selected following systematic sampling method from the each selected EA [25]. The main objective of such national level survey is to provide updated facts and figures on health and demographic indicators for the policy-makers.

Childhood health seeking behavior information (choice of service providers/facilities) has been collected for all children of age under 5 years for their illness (due to diarrhea, ARI, fever) in the two weeks preceding the survey. Of the total children ($n = 7886$), 406 were reported to suffer from ARI, 2768 from fever and 371 from diarrhea. The information about the choice of service providers for the sick children is also collected from the respondent (their parents and caregiver). Of the children with any illness, 355 with ARI, 1999 with fever and 280 with diarrhea received treatment from medical facilities. The health seeking
Figure 6. Bias and MSE of the MLE and PMLE estimates of $\beta_{22c}$ and $\beta_{23c}$ (coefficients of continuous covariate that made near-to-separation) over sample size, where Max.MCE is the maximum Monte Carlo error.

Table 4. The estimates of MLE and PMLE in the presence of near-to-separation. The estimates are the mean over the number of simulations in which near-to-separation occurred out of 1000 simulations under sample size and $\gamma = 0.5$.

| Sample size | True coefficient | Estimate | SE    | SimSE | SE/SimSE | Conf. width | Cov (%) |
|-------------|------------------|----------|-------|-------|-----------|-------------|--------|
| 30          | $\beta_{12b} = 1.3$ | MLE      | 1.47  | 1.21  | 1.98      | 0.61        | 5.02   | 98     |
|             |                   | PMLE     | 1.18  | 1.08  | 0.94      | 1.15        | 4.46   | 98     |
|             | $\beta_{13b} = 1.2$ | MLE      | 1.38  | 1.19  | 1.98      | 0.61        | 4.90   | 98     |
|             |                   | PMLE     | 1.11  | 1.06  | 0.95      | 1.12        | 4.37   | 99     |
|             | $\beta_{22c} = 0.65$ | MLE      | 0.84  | 0.68  | 1.47      | 0.46        | 2.81   | 99     |
|             |                   | PMLE     | 0.64  | 0.58  | 0.60      | 0.97        | 2.41   | 99     |
|             | $\beta_{23c} = 0.5$ | MLE      | 0.66  | 0.66  | 1.45      | 0.45        | 2.74   | 97     |
|             |                   | PMLE     | 0.50  | 0.57  | 0.60      | 0.95        | 2.35   | 99     |
| 50          | $\beta_{12b} = 1.3$ | MLE      | 1.43  | 0.84  | 0.91      | 0.92        | 3.39   | 96     |
|             |                   | PMLE     | 1.28  | 0.81  | 0.80      | 1.01        | 3.28   | 97     |
|             | $\beta_{13b} = 1.2$ | MLE      | 1.29  | 0.82  | 0.85      | 0.96        | 3.31   | 96     |
|             |                   | PMLE     | 1.16  | 0.80  | 0.75      | 1.07        | 3.21   | 98     |
|             | $\beta_{22c} = 0.65$ | MLE      | 0.75  | 0.45  | 0.51      | 0.88        | 1.81   | 95     |
|             |                   | PMLE     | 0.66  | 0.43  | 0.44      | 0.98        | 1.74   | 97     |
|             | $\beta_{23c} = 0.5$ | MLE      | 0.58  | 0.43  | 0.48      | 0.89        | 1.75   | 97     |
|             |                   | PMLE     | 0.51  | 0.42  | 0.42      | 1.00        | 1.69   | 98     |

Note: $\gamma$, proportion of events in binary covariate; SimSE, simulated standard error; Conf. width, confidence width; Cov, coverage.
behavior is defined as their choice of service providers or place from where they first sought treatment. The outcome variable ‘choice of health facilities’ is categorized as

(i) local facilities – drug dealer of pharmacy or unqualified doctor such as so-called untrained village doctor,
(ii) public facilities – public hospital, district hospital, maternal and child welfare center, upazila health complex or other public sector,
(iii) private facilities – private hospital, NGO static clinic, private chamber or other private sector.

Of these facilities, the most expensive one is the private facility, which is followed by the public and the local facilities. The local facilities are equipped with all unqualified medical providers who basically sell the medicine and hence most of the time they charge for medicine but not for consultation. Generally, people from poor household seek for treatment from such local facilities [1]. In addition, due to lack of knowledge, some people often use these facilities, even in most cases government facility provides health service with very low cost. However, government facilities often suffer from shortage of health-care providers, and hence it takes time to get service.

Several covariates are considered here including wealth index (poorest, poorer, middle, richer, richest), place of residence (urban, rural), current working status of the parent’s (yes, no), mother’s education (none, primary, secondary, higher) and respondent’s age. Of them wealth index (the socio-economic status of the parents) is considered to be an influential covariate that may greatly influence the choice of health facilities, because medical expenses significantly differ among the above three facilities.

4.2. Analysis and results

As this survey used a two-stage stratified sampling scheme where urban and rural are two important strata, obtaining separate estimates for urban and rural strata, in addition to national estimate, is a major concern for the decision makers to make appropriate policies for urban and rural part of the country. Therefore, we performed separate analysis for urban and rural strata in addition to combined analysis with full dataset. Summary statistics from the contingency table of the outcome variable (choice of facilities) for each type of disease (ARI/fever/diarrhea) suggest that the outcome variables related to ARI and diarrhea for urban sample is separated by the wealth index (Table 2) and those for fever is nearly separated (near-to-separation) by the same covariate (results not shown). For rural sample, the outcome variables for both ARI and diarrhea are also nearly separated by the wealth index. The education level also created a sort of near-to-separation for ARI and diarrhea for the urban sample (results not shown).

For the choice of health facilities for each childhood illness (ARI, fever, diarrhea), multinomial models with both MLE and PMLE version were fitted separately for the urban and rural strata and combined data. Again, as the sampling design of this survey is based on a multi-stage design, we therefore fitted the model incorporating households sampling weight and reported the weighted estimate of the regression coefficients and its SE. In each model, we explored the relationship between health seeking behavior of parents for illness (ARI, fever, diarrhea) and the covariates wealth index, place of residence, current
Table 5. Goodness-of-fit statistics of the models fitted using MLE and PMLE for the data from different population (urban, rural and both) on childhood health seeking behavior for treatment of ARI, fever and diarrhea.

| Criterion                  | ARI     | Fever    | Diarrhea |
|----------------------------|---------|----------|----------|
|                            | MLE     | PMLE     | MLE      | PMLE     | MLE      | PMLE     |
| Urban AIC                  | 157.85  | 159.60   | 928.80   | 929.03   | 191.29   | 193.34   |
| Residual deviance          | −60.93  | −61.80   | −444.40  | −444.51  | −75.64   | −76.67   |
| Log-likelihood             | 31.41*  | 29.67*   | 94.32*   | 94.09*   | 26.58    | 24.54    |
| LR test                    | 22.50   | 17.31    | 39.84*   | 40.85*   | 17.64    | 17.43    |
| Hosmer–Lemeshow test       |         |          |          |          |          |          |
| Rural AIC                  | 554.78  | 555.18   | 2378.89  | 2379.02  | 430.02   | 430.56   |
| Residual deviance          | 518.78  | 519.18   | 2338.89  | 2339.02  | 390.02   | 390.56   |
| Log-likelihood             | −259.39 | −259.59  | −1169.50 | −1169.51 | −195.01  | −195.28  |
| LR test                    | 41.86*  | 41.47*   | 149.66*  | 149.53*  | 72.94*   | 72.40*   |
| Hosmer–Lemeshow test       | 16.75   | 20.12    | 23.58    | 23.02    | 31.22*   | 32.52*   |
| Both AIC                   | 705.39  | 705.63   | 3312.13  | 3312.21  | 624.74   | 625.09   |
| Residual deviance          | 665.39  | 665.63   | 3268.13  | 3268.21  | 580.74   | 581.09   |
| Log-likelihood             | −332.69 | −332.82  | −1634.10 | −1634.12 | −290.37  | −290.54  |
| LR test                    | 58.24*  | 57.99*   | 254.85*  | 254.76*  | 85.32*   | 84.98*   |
| Hosmer–Lemeshow test       | 17.75   | 21.66    | 21.12    | 19.60    | 21.84    | 20.25    |

Note: AIC, Akaike information criterion; LR, likelihood ratio; *, significant test statistics.

working status, mother’s education and respondent’s age. Furthermore, in each model, we compared the public and private facilities with the local facilities as reference category. The goodness-of-fit statistics (AIC, deviance residuals, likelihood ratio test, and Hosmer–Lemeshow test) of each model fitted using both MLE and PMLE showed comparable results (Table 5). The goodness-of-fit for all models (models for different types of illness and sub-groups in Tables 6–8) are found to be satisfactory for at least one of the goodness-of-fit criterion applied here.

The results for urban stratum revealed that the MLE provided infinitely large estimates of the odds ratio associated with the wealth index that created quasi-complete separation in the outcome (choice of health facilities) for the illness due to ARI, which is not interpretable (Table 6). In contrast, the PMLE provided comparatively very smaller odds ratio associated with the respective covariate. In addition, the PMLE provided comparatively smaller value for each of the other estimates of the OR associated with education than the MLE. When the model fitted for rural stratum (for which sample size is larger than the urban stratum), a remarkable difference in the estimates between the MLE and PMLE is also observed, particularly for the OR associated with the wealth index that created near-to-separation, with relatively smaller value for the PMLE. For the combined data, a negligible difference in the estimates between two approaches is observed.

The fever data are comparatively large and there was no evidence of complete separation except for near-to-separation in the urban stratum. However, there was evidence of separation due to wealth index in diarrhea data for the urban stratum and a sort of near-to separation in the rural stratum. Similar results were observed in the analysis of fever (Table 7) and diarrhea (Table 8) data with noticeable difference in the estimates of regression coefficients while separation exists and small difference when there is evidence of near-to-separation and negligible difference when there is no evidence of any kind of separation.
Table 6. The summary of the fitted MLE and PMLE to the child health seeking behavior data for treatment of ARI.

| Covariate   | MLE | Private | PMLE | Private |
|-------------|-----|---------|------|---------|
|             | OR  | SE      | P-value | OR  | SE      | P-value | OR  | SE      | P-value |
| Urban       |     |         |        |       |         |        |       |         |        |
| Wealth      |     |         |        |       |         |        |       |         |        |
| Poorest     | 24.98 (50.71) | – | 6.42×10⁵ (0.73×10⁶) | – | 11.24 (20.68) | – | 14.72 (32.24) | – |
| Middle      | 16.91 (0.52×10⁴) | – | 9.54×10⁶ (0.73×10⁶) | – | 8.80 (14.52) | – | 21.06 (41.27) | – |
| Richer      | 4.16 (5.91) | – | 2.28×10⁵ (0.15×10⁶) | – | 2.66 (3.46) | – | 6.06 (10.97) | – |
| Richest     | 4.24 (7.04) | – | 1.12×10⁶ (0.78×10⁶) | – | 2.76 (4.14) | – | 25.18 (46.83) | – |
| Education   |     |         |        |       |         |        |       |         |        |
| None        | 0.03 (0.04) | – | 0.53 (0.94) | – | 0.07 (0.09) | – | 0.54 (0.78) | – |
| Primary     | 0.12 (0.17) | – | 1.04 (1.85) | – | 0.19 (0.24) | – | 1.01 (1.44) | – |
| Higher      | 0.39 (0.74) | – | 2.08 (4.49) | – | 0.51 (0.89) | – | 1.50 (2.76) | – |
| Age         | 1.09 (0.07) | – | 1.02 (0.05) | – | 1.07 (0.06) | – | 1.02 (0.05) | – |
| Rural       |     |         |        |       |         |        |       |         |        |
| Wealth      |     |         |        |       |         |        |       |         |        |
| Poorest     | 0.41 (0.18) | 0.039 | 2.57 (1.08) | 0.024 | 0.42 (0.18) | 0.046 | 2.49 (1.02) | 0.028 |
| Middle      | 0.17 (0.10) | 0.004 | 1.51 (0.69) | 0.373 | 0.19 (0.11) | 0.005 | 1.49 (0.69) | 0.381 |
| Richer      | 0.34 (0.22) | 0.093 | 2.83 (1.42) | 0.037 | 0.38 (0.24) | 0.117 | 2.75 (1.38) | 0.041 |
| Richest     | 0.39 (0.39) | 0.341 | 5.69 (3.70) | 0.007 | 0.48 (0.44) | 0.422 | 5.30 (3.44) | 0.009 |
| Education   |     |         |        |       |         |        |       |         |        |
| None        | 5.78 (4.10) | 0.013 | 1.26 (0.60) | 0.635 | 4.90 (3.28) | 0.017 | 1.22 (0.59) | 0.678 |
| Primary     | 4.34 (3.30) | 0.054 | 1.41 (0.69) | 0.487 | 3.70 (2.66) | 0.069 | 1.36 (0.67) | 0.530 |
| Higher      | 9.26 (11.48) | 0.073 | 0.70 (0.68) | 0.709 | 8.15 (9.62) | 0.075 | 0.77 (0.74) | 0.783 |
| Age         | 1.06 (0.03) | 0.073 | 1.05 (0.03) | 0.052 | 1.06 (0.03) | 0.075 | 1.05 (0.03) | 0.061 |
| Both        |     |         |        |       |         |        |       |         |        |
| Wealth      |     |         |        |       |         |        |       |         |        |
| Poorest     | 0.50 (0.21) | 0.087 | 2.80 (1.15) | 0.012 | 0.51 (0.21) | 0.101 | 2.71 (1.11) | 0.014 |
| Middle      | 0.21 (0.11) | 0.003 | 1.96 (0.85) | 0.122 | 0.23 (0.12) | 0.004 | 1.91 (0.82) | 0.130 |
| Richer      | 0.31 (0.16) | 0.023 | 2.25 (1.03) | 0.076 | 0.33 (0.17) | 0.030 | 2.20 (0.99) | 0.081 |
| Richest     | 0.31 (0.23) | 0.106 | 6.82 (3.63) | 0.000 | 0.35 (0.25) | 0.131 | 6.35 (3.37) | 0.000 |
| Residence   |     |         |        |       |         |        |       |         |        |
| Urban       | 0.35 (0.15) | 0.016 | 0.59 (0.21) | 0.131 | 0.36 (0.15) | 0.018 | 0.59 (0.21) | 0.139 |
| Rural       |     |         |        |       |         |        |       |         |        |
| Education   |     |         |        |       |         |        |       |         |        |
| None        | 2.06 (0.98) | 0.130 | 1.38 (0.62) | 0.475 | 1.96 (0.92) | 0.153 | 1.34 (0.60) | 0.517 |
| Primary     | 1.79 (0.95) | 0.273 | 1.66 (0.76) | 0.268 | 1.70 (0.88) | 0.308 | 1.60 (0.72) | 0.302 |
| Higher      | 4.68 (4.23) | 0.088 | 1.28 (0.99) | 0.747 | 4.42 (3.93) | 0.094 | 1.27 (0.98) | 0.754 |
| Age         | 1.06 (0.03) | 0.050 | 1.04 (0.02) | 0.081 | 1.05 (0.03) | 0.054 | 1.04 (0.02) | 0.092 |

Note: OR, odds ratio; SE, standard error; P, P-value; RC, reference category.

5. Discussion

Complete separation or near-to-separation (partial separation) is a serious problem in the fitting process of a model. Although the data with complete or quasi-complete separation
Table 7. The summary of the fitted MLE and PMLE to the data on childhood health seeking behavior for treatment of fever.

| Covariates | MLE       | Public | OR (SE) | P-value | Private | OR (SE) | P-value |
|------------|-----------|--------|---------|---------|---------|---------|---------|
|            |           |        |         |         |         |         |         |
| Urban      | Wealth    |         |         |         |         |         |         |
| Poorer     | 1.13 (0.72) | 0.845  | 2.95 (2.02) | 0.114 | 1.14 (0.71) | 0.827  | 2.77 (1.86) | 0.129 |
| Middle     | 0.95 (0.50) | 0.916  | 1.15 (0.75) | 0.833 | 0.94 (0.49) | 0.906  | 1.11 (0.70) | 0.874 |
| Richer     | 1.14 (0.53) | 0.784  | 3.54 (1.94) | 0.021 | 1.11 (0.51) | 0.826  | 3.24 (1.72) | 0.028 |
| Richest    | 0.43 (0.21) | 0.089  | 4.86 (2.63) | 0.003 | 0.43 (0.21) | 0.086  | 4.42 (2.33) | 0.004 |
| Work       | No        |         |         |         |         |         |         |
| Yes        | 1.20 (0.39) | 0.581  | 1.41 (0.38) | 0.200 | 1.20 (0.39) | 0.566  | 1.40 (0.37) | 0.211 |
| Education  | None      |         |         |         |         |         |         |
| Primary    | 1.67 (0.83) | 0.304  | 2.07 (0.93) | 0.106 | 1.60 (0.78) | 0.332  | 1.98 (0.88) | 0.123 |
| Secondary  | 1.76 (0.85) | 0.242  | 3.76 (1.57) | 0.001 | 1.68 (0.80) | 0.273  | 3.55 (1.46) | 0.002 |
| Higher     | 9.28 (5.24) | 0.000  | 11.44 (5.57) | 0.000 | 8.47 (4.70) | 0.000  | 10.50 (5.06) | 0.000 |
| Age        | 1.01 (0.03) | 0.621  | 1.00 (0.02) | 0.838 | 1.01 (0.03) | 0.605  | 1.00 (0.02) | 0.844 |
| Rural      | Wealth    |         |         |         |         |         |         |
| Poorer     | 0.87 (0.22) | 0.586  | 1.82 (0.43) | 0.010 | 0.88 (0.22) | 0.602  | 1.81 (0.42) | 0.011 |
| Middle     | 1.27 (0.30) | 0.312  | 3.71 (0.80) | 0.000 | 1.26 (0.30) | 0.309  | 3.61 (0.78) | 0.000 |
| Richer     | 0.88 (0.25) | 0.635  | 2.28 (0.57) | 0.001 | 0.88 (0.24) | 0.654  | 2.26 (0.56) | 0.001 |
| Richest    | 0.35 (0.17) | 0.027  | 6.24 (1.72) | 0.000 | 0.37 (0.17) | 0.032  | 6.13 (1.68) | 0.000 |
| Work       | No        |         |         |         |         |         |         |
| Yes        | 0.68 (0.14) | 0.050  | 0.55 (0.10) | 0.000 | 0.68 (0.13) | 0.053  | 0.56 (0.09) | 0.001 |
| Education  | None      |         |         |         |         |         |         |
| Primary    | 0.77 (0.20) | 0.322  | 0.75 (0.17) | 0.213 | 0.77 (0.20) | 0.316  | 0.75 (0.17) | 0.211 |
| Secondary  | 1.12 (0.30) | 0.667  | 0.87 (0.19) | 0.521 | 1.11 (0.29) | 0.686  | 0.87 (0.19) | 0.511 |
| Higher     | 3.51 (1.33) | 0.001  | 1.53 (0.48) | 0.173 | 3.47 (1.31) | 0.001  | 1.52 (0.47) | 0.178 |
| Age        | 1.02 (0.02) | 0.163  | 0.98 (0.01) | 0.073 | 1.02 (0.02) | 0.160  | 0.98 (0.01) | 0.077 |
| Both       | Wealth    |         |         |         |         |         |         |
| Poorer     | 0.89 (0.21) | 0.615  | 1.82 (0.40) | 0.007 | 0.89 (0.21) | 0.629  | 1.81 (0.40) | 0.007 |
| Middle     | 1.22 (0.26) | 0.349  | 3.02 (0.61) | 0.000 | 1.21 (0.25) | 0.347  | 2.99 (0.60) | 0.000 |
| Richer     | 0.92 (0.21) | 0.732  | 2.46 (0.53) | 0.000 | 0.93 (0.21) | 0.740  | 2.44 (0.52) | 0.000 |
| Richest    | 0.37 (0.11) | 0.001  | 4.99 (1.12) | 0.000 | 0.38 (0.12) | 0.002  | 4.93 (1.11) | 0.000 |
| Residence  | Urban     |         |         |         |         |         |         |
| Rural      | 0.49 (0.09) | 0.000  | 0.72 (0.10) | 0.019 | 0.49 (0.09) | 0.000  | 0.72 (0.10) | 0.019 |
| Work       | No        |         |         |         |         |         |         |
| Yes        | 0.77 (0.13) | 0.124  | 0.72 (0.10) | 0.015 | 0.78 (0.13) | 0.129  | 0.72 (0.10) | 0.016 |
| Education  | None      |         |         |         |         |         |         |
| Primary    | 0.91 (0.21) | 0.674  | 0.90 (0.18) | 0.603 | 0.91 (0.21) | 0.665  | 0.89 (0.20) | 0.595 |
| Secondary  | 1.19 (0.27) | 0.455  | 1.22 (0.23) | 0.297 | 1.18 (0.27) | 0.471  | 1.21 (0.23) | 0.308 |
| Higher     | 4.40 (1.33) | 0.000  | 2.78 (0.69) | 0.000 | 4.33 (1.31) | 0.000  | 2.75 (0.68) | 0.000 |
| Age        | 1.02 (0.01) | 0.158  | 0.99 (0.01) | 0.240 | 1.01 (0.01) | 0.155  | 0.99 (0.01) | 0.245 |

Note: OR, odds ratio; SE, standard error; P, P-value; RC, reference category.
Table 8. The summary of the fitted MLE and PMLE to the data on childhood health seeking behavior for treatment of diarrhea.

| Covariate | Public MLE | Private MLE | Public PMLE | Private PMLE |
|-----------|------------|-------------|-------------|--------------|
|           | OR (SE) P-value | OR (SE) P-value | OR (SE) P-value | OR (SE) P-value |
| Urban     |             |             |             |              |
| Wealth    |             |             |             |              |
| Poorest   | 3.85 (25.3) – | 1.27 × 10³ (15.7 × 10³) – | 4.38 (0.76) 0.238 | 2.64 (0.22) 0.098 |
| Poorer    | 3.83 (8.57) – | 28.3 × 10³ (249 × 10³) – | 2.99 (0.77) 0.456 | 4.75 (0.52) 0.740 |
| Middle    | 1.14 × 10⁻¹ (0.23) – | 1.70 × 10⁴ (11.6 × 10⁴) – | 0.19 (0.41) 0.445 | 3.14 (0.31) 0.331 |
| Richer    | 6.15 × 10⁻¹ (1.24) – | 4.50 × 10⁴ (32.7 × 10⁴) – | 0.76 (0.78) 0.812 | 7.71 (0.80) 0.360 |
| Richest   |             |             |             |              |
| Work      |             |             |             |              |
| No (RC)   | 3.72 (4.12) 0.370 | 1.70 (1.43) 0.234 | 2.98 (0.16) 0.021 | 1.53 (0.16) 0.022 |
| Education |             |             |             |              |
| None (RC) |             |             |             |              |
| Primary   | 0.12 (0.35) 0.448 | 0.57 (0.77) 0.680 | 0.34 (0.16) 0.023 | 0.64 (0.46) 0.805 |
| Secondary | 8.26 (19.17) 0.363 | 0.43 (0.57) 0.521 | 3.88 (0.25) 0.190 | 0.50 (0.84) 0.272 |
| Higher    | 5.04 (12.07) 0.499 | 0.70 (0.98) 0.794 | 2.78 (0.61) 0.830 | 0.73 (2.00) 0.079 |
| Age       | 1.03 (0.09) 0.753 | 0.91 (0.06) 0.148 | 1.02 (0.03) 0.059 | 0.92 (0.03) 0.331 |
| Rural     |             |             |             |              |
| Wealth    |             |             |             |              |
| Poorest   | 1.80 (0.83) 0.205 | 0.40 (0.20) 0.075 | 1.73 (0.79) 0.226 | 0.43 (0.21) 0.087 |
| Poorer    | 1.52 (0.90) 0.480 | 0.94 (0.47) 0.911 | 1.51 (0.87) 0.477 | 0.96 (0.48) 0.933 |
| Middle    | 1.50 (1.11) 0.584 | 0.27 (0.19) 0.062 | 1.53 (1.09) 0.551 | 0.30 (0.21) 0.083 |
| Richer    | 1.33 (1.28) 0.765 | 1.22 (0.78) 0.754 | 1.45 (1.33) 0.685 | 1.21 (0.77) 0.764 |
| Richest   |             |             |             |              |
| Work      |             |             |             |              |
| No (RC)   | 0.18 (0.11) 0.005 | 0.21 (0.11) 0.004 | 0.21 (0.12) 0.006 | 0.23 (0.12) 0.005 |
| Education |             |             |             |              |
| None (RC) |             |             |             |              |
| Primary   | 0.41 (0.20) 0.063 | 0.70 (0.46) 0.586 | 0.42 (0.20) 0.070 | 0.69 (0.44) 0.568 |
| Secondary | 0.39 (0.20) 0.068 | 3.42 (1.96) 0.032 | 0.41 (0.21) 0.078 | 3.11 (1.74) 0.042 |
| Higher    | 0.91 (0.81) 0.917 | 7.56 (6.09) 0.012 | 0.95 (0.82) 0.952 | 6.45 (5.12) 0.019 |
| Age       | 1.09 (0.04) 0.033 | 0.97 (0.03) 0.355 | 1.08 (0.04) 0.042 | 0.97 (0.03) 0.386 |
| Both      |             |             |             |              |
| Wealth    |             |             |             |              |
| Poorest   | 1.73 (0.77) 0.217 | 0.45 (0.21) 0.086 | 1.68 (0.76) 0.238 | 0.47 (0.22) 0.098 |
| Poorer    | 1.49 (0.79) 0.453 | 1.16 (0.52) 0.742 | 1.48 (0.77) 0.456 | 1.16 (0.52) 0.740 |
| Middle    | 0.56 (0.39) 0.404 | 0.60 (0.30) 0.304 | 0.60 (0.41) 0.445 | 0.62 (0.31) 0.331 |
| Richer    | 1.16 (0.79) 0.831 | 1.60 (0.81) 0.350 | 1.17 (0.78) 0.812 | 1.59 (0.80) 0.360 |
| Richest   |             |             |             |              |
| Residence |             |             |             |              |
| Urban     |             |             |             |              |
| Rural     | 0.43 (0.22) 0.106 | 0.38 (0.14) 0.011 | 0.45 (0.23) 0.115 | 0.40 (0.15) 0.014 |
| Work      |             |             |             |              |
| No (RC)   | 0.32 (0.15) 0.016 | 0.39 (0.15) 0.016 | 0.34 (0.16) 0.021 | 0.41 (0.16) 0.022 |
| Education |             |             |             |              |
| None (RC) |             |             |             |              |
| Primary   | 0.33 (0.16) 0.019 | 0.90 (0.48) 0.841 | 0.35 (0.16) 0.023 | 0.88 (0.46) 0.805 |
| Secondary | 0.54 (0.25) 0.179 | 1.80 (0.90) 0.239 | 0.55 (0.25) 0.190 | 1.71 (0.84) 0.272 |
| Higher    | 0.83 (0.60) 0.798 | 3.38 (2.20) 0.061 | 0.86 (0.61) 0.830 | 3.10 (2.00) 0.079 |
| Age       | 1.07 (0.04) 0.048 | 0.97 (0.03) 0.313 | 1.06 (0.03) 0.059 | 0.97 (0.03) 0.331 |

Note: OR, odds ratio; SE, standard error; P, P-value; RC, reference category.
are rare but nearly separated data (near-to-separation) are very common in practice. However, it is essential to explore the data condition (rule of thumb) to define near-to-separation. This study determined a threshold value (<15%) of the observed count in at least one cell of a contingency table of a binary covariate and the outcome. Furthermore, this study investigated the consequences of both separation and near-to-separation in the maximum likelihood-based standard multinomial model and addressed the problems by applying the penalized likelihood-based Poisson regression for multinomial response. The simulation results showed that in the presence of complete or quasi-complete separation, the MLE often failed to achieve convergence and/or provided large or infinite estimates of the regression coefficient and Wald-confidence interval, which is not interpretable. In contrast, the PMLE showed great improvement by achieving convergence and providing finite estimate with very small amount of bias and MSE for the corresponding regression coefficient. In the presence of near-to-separation, the standard MLE provided large amount of bias and MSE in the estimate of the regression coefficient associated with the covariate that creates near-to-separation. The amount of bias and MSE were reported to be comparatively high for small sample, large true regression coefficient, the uneven distribution of the event and non-event of the binary covariate. Because for each of these cases, the chance of such near-to-separation was high. For all these scenarios, the PMLE showed improvement over MLE by reducing bias and MSE to some extent and providing narrower confidence interval and accurate coverage. Even for large sample with rare outcome category, the MLE still provided large amount of bias and MSE while the PMLE offered improvement to some extent.

In contrast, the amount of bias and MSE for both the MLE and PMLE for the regression coefficient associated with the continuous covariate that is not responsible for creating any kind of separation were comparatively lower than those associated with the binary covariate. Likewise, the PMLE reported smaller amount of bias and MSE than those in the MLE when sample size was small.

This study also extended the penalized estimation method to analyze data from a complex survey by incorporating individual survey-weight (inverse probability weighting) and provided an application of the methods for analyzing health seeking behavior (choice of health-care facilities for childhood illness) data with evidence of separation or near-to-separation. The results supported the simulation findings. All the findings of this paper are similar to those reported in Ref. [4], which investigated finite sample bias in both Firth’s type penalized estimate and MLE for the multinomial logistic regression. However, they ignored the issues related to near-to-separation in the multinomial outcome, which may occur frequently even for moderate-to large sample with rare outcome category.

Finally, the PMLE for the multinomial logit model via Poisson regression framework showed significant improvement over the standard MLE-based multinomial logit model in the presence of separation or near-to-separation created by a covariate. The chance of separation or near-to-separation is reported to be high when sample size is small or even large but the prevalence of one or more outcome category is low or there is a number of influential covariates or combination of them, and the problems due to such separation is non-negligible for multinomial response. Based on the findings of the study, we recommend to use PMLE if sample size is small ($n \leq 50$). Again, if the sample size is large, we suggest to explore the data to identify any evidence of separation or near-to-separation, and if exists, recommend to use PMLE.
This study explored the use of penalized methods to analyze the sparse multinomial data from a complex survey, however, further study is required to develop appropriate goodness-of-fit methods aligned with penalized maximum likelihood approach. Residual-based goodness-of-fit assessments discussed in some recent studies [13,14] may be an alternative choice in addition to traditional approaches such as Akaike information criteria, Hosmer–Lameshow test and likelihood ratio test.

Acknowledgments
The authors acknowledge the authority of the DHS program for providing data for this study. The authors also acknowledge the contribution of the associate editor, two anonymous reviewers for their constructive comments and feedback throughout the review process.

Disclosure statement
No potential conflict of interest was reported by the author(s).

ORCID
M. S. Rahman http://orcid.org/0000-0001-5256-7453

References
[1] S.S. Andaleeb, Public and private hospitals in Bangladesh: service quality and predictors of hospital choice, J. Health Policy Plann. 15 (2000), pp. 95–102.
[2] J.A. Anderson and S.C. Richardson, Logistic discrimination and bias correction in maximum likelihood estimation, Technometrics 21 (1979), pp. 71–78.
[3] S.B. Bull, W.W. Hauck, and C.M.T. Greenwood, Two-step jackknife bias reduction for logistic regression MLEs, Commun. Stat. Simul. Comput. 23 (1994), pp. 59–88.
[4] S.B. Bull, C. Mak, and C.M.T. Greenwood, A modified score function estimator for multinomial logistic regression in small samples, Comput. Stat. Data Anal. 39 (2002), pp. 57–74.
[5] S.J. Cook, J. Niehaus, and S. Zuhlke, A warning on separation in multinomial logistic models, Res. Polit. 5 (2018), pp. 1–5.
[6] J.B. Copas, Binary regression models for contaminated data, J. R. Stat. Soc.: Ser. B (Methodol.) 50 (1988), pp. 225–253.
[7] G.M. Cordeiro and P. McCullagh, Bias correction in generalized linear models, J. R. Stat. Soc.: Ser. B (Methodol.) 53 (1991), pp. 629–643.
[8] D.R. Cox and E.J. Snell, A general definition of residuals, J. R. Stat. Soc. Ser. B (Methodol.) 43 (1968), pp. 248–275.
[9] A.J. Dobson and A.G. Barnett, An Introduction to Generalized Linear Models, CRC Press (A Chapman & Hall Book), London, New York, 2018.
[10] D. Firth, Bias reduction of maximum likelihood estimates, Biometrika 80 (1993), pp. 27–38.
[11] G.M. Fitzmaurice, D.P. Harrington, M. Parzen, D. Sinha, K.A. Rader, and S.R. Lipsitz, Bias-corrected estimates for logistic regression models for complex surveys with application to the united states' nationwide inpatient sample, Stat. Methods Med. Res. 26 (2017), pp. 2257–2269.
[12] B.D. Frischknecht, C. Eckert, J. Geweke, and J.J. Louviere, Incorporating prior information to overcome complete separation problems in discrete choice model estimation. Technical report, University of Technology, Sydney, 2013.
[13] J.J. Goeman and S. le Cessie, A goodness-of-fit test for multinomial logistic regression, Biometrics 62 (2006), pp. 980–985.
[14] J.J. Goeman and S. le Cessie, Dealing with the phenomenon of quasi-complete separation and a goodness of fit test in logistic regression models in the case of long data sets, Stat. Biosci. 11 (2019), pp. 567–596.
[15] G. Heinze, *A comparative investigation of methods for logistic regression with separated or nearly separated data*, Stat. Med. 25 (2006), pp. 4216–4226.

[16] G. Heinze and M. Schemper, *A solution to the problem of separation in logistic regression*, Stat. Med. 21 (2002), pp. 2409–2419.

[17] I. Kosmidis and D. Firth, *Bias reduction in exponential family nonlinear models*, Biometrika 96 (2009), pp. 793–804.

[18] I. Kosmidis and D. Firth, *Multinomial logit bias reduction via the Poisson log-linear model*, Biometrika 98 (2011), pp. 755–759.

[19] S.R. Lipsitz, G.M. Fitzmaurice, S.E. Regenbogen, D. Sinha, J.G. Ibrahim, and A.A. Gawande, *Bias correction for the proportional odds logistic regression model with application to a study of surgical complications*, J. R. Stat. Soc.: Ser. C (Appl. Stat.) 62 (2013), pp. 233–250.

[20] M.H. Mondol and M.S. Rahman, *Bias-reduced and separation-proof GEE with small or sparse longitudinal binary data*, Stat. Med. 38 (2019), pp. 2544–2560.

[21] M. Noh and Y. Lee, *REML estimation for binary data in GLMMs*, J. Multivar. Anal. 98 (2007), pp. 896–915.

[22] R. Puhr and G. Heinze, *Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets*, Stat. Med. 29 (2010), pp. 770–777.

[23] R.L. Schafer, *Bias correction in maximum likelihood logistic regression*, Stat. Med. 2 (1983), pp. 71–78.

[24] M. Siino, S. Fasola, and V.M.R. Muggeo, *Inferential tools in penalized logistic regression for small and sparse data: a comparative study*, Stat. Methods Med. Res. 27 (2018), pp. 1365–1375.

[25] NIPORT: IFC-International, and Mitra-Associate, *Bangladesh Demographic and Health Survey*, National Institute of Population Research and Training (NIPORT), Dhaka, 2014.