Asymptotic properties of maximum likelihood estimator for two-step logit models

Changming Yin\textsuperscript{a,}\textsuperscript{*}, Zhanfeng Wang\textsuperscript{b}, Hong Zhang\textsuperscript{c}

\textsuperscript{a} School of Mathematics and Information Science, Guangxi University, Nanning 530004, PR China
\textsuperscript{b} Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, PR China
\textsuperscript{c} Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai 200433, PR China

ARTICLE INFO

Article history:
Received 30 June 2013
Received in revised form 15 October 2013
Accepted 22 November 2013
Available online 4 December 2013

Keywords:
Two-step models
Ordinal data
Rate of strong consistency
Asymptotic normality

ABSTRACT

Two-step logit models are extensions of the ordinary logistic regression model, which are designed for complex ordinal outcomes commonly seen in practice. In this paper, we establish some asymptotic properties of the maximum likelihood estimator (MLE) of the regression parameter vector under some mild conditions, which include existence of the MLE, convergence rate and asymptotic normality of the MLE. We relax the boundedness condition of the regressors required in most existing theoretical results, and all conditions are easy to verify.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Generalized linear models (GLMs) (Nelder and Wedderburn, 1972) have been widely used to handle non-normal responses such as categorical data and count data. Statistical inference of GLMs relies heavily on the asymptotic properties of the maximum likelihood estimator (MLE) or maximum quasi-likelihood estimator (MQLE). The asymptotic properties of the MLE and MQLE for GLMs have been extensively studied in the literature, see for example, Haberman (1977), Fahrmeir and Kaufmann (1985), Chen et al. (1999), Chang (1999), Yue and Chen (2004), Yin and Zhao (2005), Zhang and Liao (2008), and Gao et al. (2012). In all of these papers, the regressors are assumed to be either bounded or to satisfy some conditions that are hard to verify. Fahrmeir and Kaufmann (1986) established asymptotic normalities of the MLEs of regression parameters in the standard logistic regression model, multinomial logit model, cumulative logit model, and loglinear Poisson model by relaxing the boundedness assumption on the regressors.

Two-step logit models (Fahrmeir and Tutz, 1994; Morawitz and Tutz, 1990; Tutz, 1989) are used for complex ordinal responses. In such models, the response categories are divided into several subsets; in the first step, a model is used to model the categories between subsets, in the second step, another model is used to model the categories within each subset. The detailed description of the two-step models is given in Section 2.

Two-step logit models can be written as generalized linear models, and the theoretical properties are studied in Fahrmeir and Kaufmann (1985). However, the conditions for these properties are either too strong (for instance, the regressors are bounded) or too complicated to verify in practice. In this paper, we obtain the convergence rate of the MLE and the asymptotic distribution of the MLE under some mild conditions that are easy to verify. Particularly, relax the boundedness assumption on the regressors to a large extent.

* Corresponding author. Tel.: +86 15077161279.
E-mail address: yinchm@ustc.edu (C. Yin).

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2013.11.014
The rest of this paper is organized as follows. A description of the two-step model is given in Section 2. The asymptotic properties of the MLE of regression parameters are established in Section 3. Detailed proofs of the main results are provided in Section 4. To verify the main results, some numerical simulations are conducted in Section 5.

2. Model description

Consider a sequence \( \{ Y_n, n = 1, 2, \ldots \} \) of responses associated with non-random covariate vectors \( \{ X_n, n = 1, 2, \ldots \} \). Assume that \( \{ Y_n, n = 1, 2, \ldots \} \) are independent of each other given \( \{ X_n, n = 1, 2, \ldots \} \), and each of them takes \( k \) categories, i.e., \( Y_n \) takes value \( r \) if the subject locates in the \( r \)th category. Let \( \pi_{nr} := P(Y_n = r | X_n) \) for \( r = 1, \ldots, k \). Obviously, \( \sum_{r=1}^{k} \pi_{nr} = 1 \), so it suffices to model \( \pi_{nr} \) for \( r = 1, \ldots, q = (k - 1) \).

In this article, let \( ^\top \) denote the transpose of a matrix or vector. If the categories are unordered, then usually \( \pi_{nr} \) can be parameterized as

\[
\pi_{nr} = \exp(X'_n \beta_r) \left[ 1 + \sum_{r=1}^{q} \exp(X'_n \beta_r) \right], \quad r = 1, \ldots, q.
\]

If the categories are ordered, then a cumulative logit model can be used:

\[
\pi_{nr} = F(\delta_r + X'_n \lambda) - F(\delta_{r-1} + X'_n \lambda), \tag{1}
\]

where

\[
F(x) = \exp(x)/(1 + \exp(x)),
\]

\( \lambda \) is a regression parameter vector, and \(-\infty = \delta_0 < \delta_1 < \cdots < \delta_q < \delta_{q+1} = \infty \) are threshold values. Another strategy is to use the following sequential logit model for ordered categories:

\[
\pi_{nr} = F(\delta_r + X'_n \lambda) \prod_{l=1}^{r-1} \tilde{F}(\delta_l + \delta_{l-1} + X'_n \lambda) \quad \text{with} \quad \tilde{F}(x) = 1 - F(x), \quad r = 1, \ldots, k, \tag{2}
\]

where \( \delta_k = \infty \) and \( \prod_{l=1}^{0}(.) = 1 \). Refer to Fahrmeir and Tutz (1994) for more comprehensive explanation about these models.

All of the above models treat each response category in the same way. In some situations, the \( k \) response categories can be naturally divided into \( s \) subsets, and the categories between various subsets should be treated differentially. Mehta et al. (1984) analyzed a dataset, where each patient was categorized as ‘much improved’, ‘improved’, ‘no change’, ‘worse’, and ‘much worse’. There are three groups that are strictly different (group 1: ‘much improved’ and ‘improved’, group 2: ‘no change’, group 3: ‘worse’ and ‘much worse’), so it is natural to divide the five categories into three subsets and model them differentially.

Without loss of generality, denote the \( j \)th subset by

\[
S_j = \{ m_{j-1} + 1, \ldots, m_j \} \quad \text{with} \quad m_0 = 0 \quad \text{and} \quad m_t = k.
\]

Now one can adopt a two-step model as follows. In the first step, a cumulative logit model is used to model the categories between the subsets, i.e.,

\[
P(Y_n \in S_j | X_n) = F(\delta_{0j} + X'_n \lambda_{0j}) - F(\delta_{0,j-1} + X'_n \lambda_{0j}), \quad j = 1, \ldots, t, \tag{3}
\]

where \(-\infty = \delta_{00} < \delta_{01} < \cdots < \delta_{0,t-1} < \delta_{0,t} = \infty \). In step 2, the categories within the \( j \)th subset \(( j = 1, \ldots, t \) are determined by a sequential logit model, i.e.,

\[
P(Y_n = r | Y_n \in S_j, X_n) = F(\delta_{jr} + X'_n \lambda_{jr}) \prod_{l=m_{j-1}+1}^{r-1} \tilde{F}(\delta_{jl} + \delta_{l-1} + X'_n \lambda_{jl}), \quad r \in S_j, \tag{4}
\]

where \( \delta_{j,m_j} = \infty \) and \( \prod_{l=m_{j-1}+1}^{m_j}(.) = 1 \).

The resulting model determined by (3) and (4) jointly is called a cumulative–sequential two-step logit model. Similarly, one obtains a cumulative–cumulative two-step logit model if both steps are determined by cumulative logit models (Fahrmeir and Tutz, 1994; Morawitz and Tutz, 1990; Tutz, 1989).

3. Main results

In this section, we focus on the cumulative–sequential two-step logit model, since the cumulative–cumulative two-step logit model can be treated in a similar way. We show that the two-step logit model can be expressed as a generalized linear model (GLM), then obtain the strong convergence rate and asymptotic normality of the MLE under some mild conditions.
First we show that the cumulative–sequential two-step logit model defined by (3) and (4) belongs to the family of GLMs. Define 
\[
\beta := (\delta_{01}, \ldots, \delta_{0t-1}, \lambda'_{0}, \delta_{1, m_1-1}, \lambda'_1, \ldots, \delta_{t, m_t-1}, \lambda'_t)',
\]
\[
Z'_n := \text{diag}(Z_{n0}, Z_{n1}, \ldots, Z_{nt}),
\]
where \(Z_{n0} := (l_{t-1}, (X_n, \ldots, X_n)')\), \(Z_{nj} = (l_{m_j-m_{j-1}-1}, (X_n, \ldots, X_n)')\) for \(j = 1, \ldots, t\), and \(l_a\) is the \(a\)-identity matrix. Let the dimension of \(\beta\) be \(p\) and denote \(\gamma' := (\gamma_1, \ldots, \gamma_t)' := Z'_n \beta\).

Now we explicitly present the link functions defined by \(\pi_{ir} := P(Y_n = r | X_n), r = 1, \ldots, k\). For \(j = 2, \ldots, t - 1\), \(\pi_{ir}\) is equal to
\[
\begin{cases}
[F(\gamma_j) - F(\gamma_{j-1})] \left( \prod_{i=m_{j-1}+1}^{j-1} \bar{F}(\gamma_{i-j+i}) \right), & \text{if } m_j - m_{j-1} > 2 \text{ and } m_{j-1} + 1 < r < m_j; \\
[F(\gamma_j) - F(\gamma_{j-1})]F(\gamma_{j-j+i}), & \text{if } m_j - m_{j-1} \geq 2 \text{ and } r = m_{j-1} + 1; \\
[F(\gamma_j) - F(\gamma_{j-1})] \left( \prod_{i=m_{j-1}+1}^{j-1} \bar{F}(\gamma_{i-j+i}) \right), & \text{if } m_j - m_{j-1} \geq 2 \text{ and } r = m_j; \\
F(\gamma_j), & \text{if } m_j - m_{j-1} = 1 \text{ and } r = m_{j-1} + 1.
\end{cases}
\]

For \(j = 1,\)
\[
\pi_{ir} = \begin{cases}
F(\gamma_1) \left( \prod_{i=1}^{t-1} \bar{F}(\gamma_{i-1+i}) \right), & \text{if } m_1 > 2 \text{ and } 1 < r < m_1; \\
F(\gamma_1)F(\gamma_t), & \text{if } m_1 \geq 2 \text{ and } r = 1; \\
F(\gamma_1) \left( \prod_{i=1}^{t-1} \bar{F}(\gamma_{i-1+i}) \right), & \text{if } m_1 \geq 2 \text{ and } r = m_1; \\
F(\gamma_1), & \text{if } m_1 = 1 \text{ and } r = 1.
\end{cases}
\]

For \(j = t,\)
\[
\pi_{ir} = \begin{cases}
\bar{F}(\gamma_{t-1})F(\gamma_t) \left( \prod_{i=m_{t-1}+1}^{t-1} \bar{F}(\gamma_{i-1+i}) \right), & \text{if } m_t - m_{t-1} > 2 \text{ and } m_{t-1} + 1 < r < m_t; \\
\bar{F}(\gamma_{t-1})F(\gamma_t), & \text{if } m_t - m_{t-1} \geq 2 \text{ and } r = m_{t-1} + 1; \\
\bar{F}(\gamma_{t-1}) \left( \prod_{i=m_{t-1}+1}^{t-1} \bar{F}(\gamma_{i-1+i}) \right), & \text{if } m_t - m_{t-1} \geq 2 \text{ and } r = m_t; \\
\bar{F}(\gamma_{t-1}), & \text{if } m_t - m_{t-1} = 1 \text{ and } r = m_{t-1} + 1.
\end{cases}
\]

For \(r = 1, \ldots, k,\) let \(y_{ir} := 1_{\{Y_n = r\}},\) which takes value 1 if \(Y_n = r\) and 0 otherwise. Denote \(y_n := (y_{n1}, \ldots, y_{nk})'\) and \(\pi_n := (\pi_{n1}, \ldots, \pi_{nk})'\), then the log likelihood is
\[
l_n(\beta) := \sum_{i=1}^{n} \sum_{r=1}^{k} y_{ir} \log \pi_{ir}, \quad \pi_{ir} = \pi_{ir}(\beta).
\]
Denote
\[
\theta'_i := (\theta_{i1}, \ldots, \theta_{ik}) := \left( \log \frac{\pi_{i1}}{\pi_{ik}}, \ldots, \log \frac{\pi_{iq}}{\pi_{ik}} \right),
\]
then the score function and information matrix are
\[
s_n(\beta) := \partial l_n(\beta) / \partial \beta = \sum_{i=1}^{n} Z_i U_i(\beta) (y_i - \pi_i(\beta)),
\]
and
\[
F_n(\beta) := \text{cov}_\beta s_n(\beta) = \sum_{i=1}^{n} Z_i U_i(\beta) \Sigma_i(\beta) U_i'(\beta) Z_i',
\]
respectively, where \(\Sigma_i(\beta) := \text{cov}_\beta (y_{ir})\) and \(U_i(\beta) := \partial \theta'_i / \partial \gamma\) are evaluated at \(\gamma = Z'_n \beta\).

From now on, we assume that the unknown \(p \times 1\) parameter \(\beta\) lies in an open set and denote by \(\beta_0\) the true value of \(\beta\). Let \(\lambda_{\text{min}} A\) denote the smallest eigenvalue of any matrix \(A\). For the cumulative–sequential two-step logit model, we establish the following main result.
Theorem 1. Suppose that the following conditions hold:

\begin{align}
(C1) \max_{1 \leq i \leq m} \|X_i\| = o(\log(n)), \\
(C2) \lambda_{\min}(\sum_{i=1}^n z_i z_i') \geq c n^r \quad \text{for some } c > 0, \quad r > 0 \quad \text{and all large } n, \quad \text{where } z_n := (1, X_n')',
\end{align}

then there exists an MLE \( \hat{\beta}_n \) of \( \beta_0 \) satisfying

\begin{align}
P(s_n(\hat{\beta}_n) = 0, \text{ for } n \text{ large}) = 1, \\
\|\hat{\beta}_n - \beta_0\| = o(n^{\frac{1}{4} - \alpha/2}) \quad \text{a.s. for any small } \epsilon_0 > 0, \\
F_n^{-\frac{1}{2}}(\hat{\beta}_n - \beta_0) \to N(0, I_p) \quad \text{in distribution},
\end{align}

where \( F_n^{-\frac{1}{2}} \) is symmetric positive definite square root of \( F_n(\beta_0) \).

Remark 1. Obviously, the sequential logit model is a special case of the cumulative–sequential two-step logit model, so the results in Theorem 1 holds true for this model.

Remark 2. For the cumulative–sequential two-step logit model, \( \log(n) \) is the sharp upper bound of \( \{X_n\} \), since it is the sharp upper bound for MLE to be asymptotically normal in the standard logistic regression model (Fahrmeir and Kaufmann, 1986).

Remark 3. For the standard logistic regression model, cumulative logit model (special models of our considered models), and loglinear Poisson model, Fahrmeir and Kaufmann (1985, 1986) obtained the same asymptotic normality. For multinomial response model, Fahrmeir and Kaufmann (1986) obtained the asymptotic normality of MLE under four conditions, one of which is that \( \text{tr} \Sigma_n^{-1} \text{tr} \Sigma_n \) is bounded uniformly in \( n \). This condition is not satisfied in the cumulative–sequential two-step logit model.

Remark 4. We not only establish the asymptotic normality of \( \hat{\beta}_n \), but also obtain a rate of the strong convergence of \( \hat{\beta}_n \).

Remark 5. Theorem 1 holds true under the cumulative–cumulative two-step logit model. The proof is similar and is omitted.

4. Proof of main results

In this paper, let \( c \) be any positive constant independent of \( n \) whose value may change from one expression to another. In order to prove Theorem 1, we prove several lemmas first.

Lemma 1. For the cumulative–sequential two-step logit model, there exists a constant \( c_0 > 0 \) such that

\[ \inf_{i \geq 1} \lambda_{\min}(U_i(\beta_0)U_i'(\beta_0)) > c_0. \]

Proof. After tedious algebraic reasoning, from (5)–(7) and (9), we deduce that the determinant of \( U_i(\beta) \) is equal to

\[
|U_i(\beta)| = \pm \left\{ \prod_{t=1}^{\ell-1} \frac{\hat{F}(\gamma_t)}{F(\gamma_t) - F(\gamma_{t-1})} - 1 \right\} \frac{\hat{F}(\gamma_\ell)}{1 - F(\gamma_\ell)} = \pm \left\{ \prod_{t=2}^{\ell} \frac{1}{1 - e^{\gamma_{t-1} - \gamma_t}} \right\}
\]

where \( \hat{F} \) stands for the derivative of \( F \). Define \( \|A\| = (\sum_{j=1}^p \sum_{y=1}^q a_{ij})^{1/2} \) for a matrix \( A = (a_{ij}) \in \mathbb{R}^{p \times q} \). It is easy to prove that, for any neighborhood \( N \) of \( \beta_0 \), there exists a constant \( c_1 > 0 \) such that

\[
\sup_{i \geq 1, \beta \in N} \|U_i(\beta)\| < c_1.
\]

By (16) and (17), we get (15), and this completes the proof of Lemma 1.

Lemma 2. For the cumulative–sequential two-step logit model, the log likelihood is concavity, i.e.,

\[ \frac{\partial^2 l_n(\beta)}{\partial \beta \partial \beta'} \leq 0. \]

Proof. It is easy to prove that

\[
\frac{\partial^2 \log(F(\gamma_j) - F(\gamma_{j-1}))}{\partial \gamma_j \partial \gamma'} \leq 0, \quad j = 2, \ldots, t - 1,
\]

\[ d^2 \log F(x)/dx^2 \leq 0, \quad d^2 \log \hat{F}(x)/dx^2 \leq 0. \]
Therefore, by (5)–(7), we have that
\[ \partial^2 \log p_{\theta} / \partial \gamma \partial \gamma' \leq 0, \quad r = 1, \ldots, k. \] (19)

Now (19) and (8) imply (18).

Lemma 3. Under the conditions of Theorem 1, there exists \( \hat{\beta}_n \) such that (12) and (13) hold.

Proof. Write \( N_n := \{ \beta : \| \beta - \beta_0 \| \leq n^{\alpha_0/2} \} \), where \( \epsilon_0 > 0 \) is a small constant. Denote by \( \bar{N}_n \) and \( \partial \bar{N}_n \) the closure and boundary of \( N_n \), respectively. It is sufficient to prove that with probability one for large \( n \),
\[ \sup_{\beta \in \partial \bar{N}_n} l_n(\beta) < l_n(\beta_0). \] (20)

Write \( W_\theta(\beta) := \partial^2 \theta / \partial \gamma \partial \gamma', \) \( W_\theta := W_\theta(\beta_0) \). By (10), we have that
\[ H_n(\beta) := -\partial^2 l_n(\beta) / \partial \beta \partial \beta', \]
\[ = \sum_{i=1}^n Z_i U_i(\beta) \Sigma_i(\beta) U_i(\beta)' Z_i' - \sum_{i=1}^n \sum_{r=1}^q Z_i W_{\theta r}(\beta) Z_i' (y_{ir} - \pi_{\theta r}(\beta)). \] (21)

By Taylor’s expansion, (10) and (21), and using similar decomposition as in Fahrmeir and Kaufmann (1985), we have that
\[ l_n(\beta) - l_n(\beta_0) = G_n(\beta) - J_n(\beta) + B_n(\beta) + C_n(\beta) + D_n(\beta), \] (22)
with
\[ G_n(\beta) := \xi^\top \sum_{i=1}^n Z_i U_i(\beta), \quad J_n(\beta) := \frac{1}{2} \xi^\top \sum_{i=1}^n Z_i U_i(\beta) \Sigma_i(\beta) U_i(\beta)' \xi, \]
\[ B_n(\beta) := \frac{1}{2} \xi^\top \sum_{i=1}^n Z_i \sum_{r=1}^q W_{\theta r}(\beta) e_r Z_i' \xi, \quad C_n(\beta) := \frac{1}{2} \xi^\top \sum_{i=1}^n Z_i \sum_{r=1}^q [W_{\theta r}(\beta) - W_{\theta r}] e_r Z_i' \xi, \]
\[ D_n(\beta) := \frac{1}{2} \xi^\top \sum_{i=1}^n Z_i \sum_{r=1}^q [W_{\theta r}(\tilde{\beta}) - W_{\theta r}] \pi_{\theta r}(\beta_0) - \pi_{\theta r}(\tilde{\beta})] Z_i' \xi, \]
where \( \xi = \beta - \beta_0 \), \( e_i = y_i - \pi_{\theta i}(\beta_0) \), \( e_{ir} = y_{ir} - \pi_{\theta r}(\beta_0) \), and \( \tilde{\beta} \) lies between \( \beta \) and \( \beta_0 \).

To prove (20), it is enough to prove that with probability one for large \( n \),
\[ \sup_{\beta \in \partial \bar{N}_n} \left\{ G_n(\beta) - \frac{1}{4} J_n(\beta) \right\} < 0, \] (23)
\[ \sup_{\beta \in \partial \bar{N}_n} \{ B_n(\beta) - J_n(\beta)/4 \} < 0, \quad \sup_{\beta \in \partial \bar{N}_n} \{ C_n(\beta) - J_n(\beta)/4 \} < 0, \] (24)
and
\[ \sup_{\beta \in \partial \bar{N}_n} \{ D_n(\beta) - J_n(\beta)/4 \} < 0. \] (25)

Note that
\[ \Sigma_i(\beta) = \text{diag}(\pi_i(\beta)) - \pi_i(\beta) \pi_i'(\beta), \quad \pi_i(\beta) = (\pi_{i1}(\beta), \ldots, \pi_{i q}(\beta))'. \] (26)

In the following, we establish a lower bound of \( \lambda_{\min} \Sigma_i(\beta) \), the minimal eigenvalue of \( \Sigma_i(\beta) \). Hereafter, \( c \) stands for any positive constant independent of \( n \) whose value may change from one expression to another.

By (26) and Gershgorin’s Theorem (Wang and Jia, 1994), we get
\[ \lambda_{\min} \Sigma_i(\beta) \geq \min \{ \pi_{i1}(\beta) \pi_{i k}(\beta), \ldots, \pi_{i q}(\beta) \pi_{i k}(\beta) \}. \] (27)

By the definition of \( \gamma \) and \( Z_n \), we have that
\[ \gamma_j - \gamma_{j-1} = (\delta_{0j} + X_n^{\top} \lambda_0) - (\delta_{0j-1} + X_n^{\top} \lambda_0) = \delta_{0j} - \delta_{0j-1} > c > 0, \quad \text{for } j = 2, \ldots, t - 1. \]
Therefore,

\[
F(\gamma_j) - F(\gamma_{j-1}) = \frac{e^{\gamma_j}}{1 + e^{\gamma_j}} - \frac{e^{\gamma_{j-1}}}{1 + e^{\gamma_{j-1}}} = \frac{e^{\gamma_{j-1}}(e^{\gamma_j} - 1)}{1 + e^{\gamma_{j-1}}(1 + e^{\gamma_j})} = \frac{c}{(1 + e^{\gamma_{j-1}})(1 + e^{\gamma_j})} \geq \frac{e^{\gamma_{j-1}}}{1 + e^{\gamma_{j-1}}}
\]

\[
\geq \begin{cases} 
\frac{c}{1 + (1 + (1 + 1)(2e^{\gamma_j}))} = \frac{c}{4e^{-|\gamma_j|-|\gamma|}}, & \text{if } \gamma_{j-1} \leq 0 \text{ and } \gamma_j > 0; \\
\frac{c}{1 + (1 + 1 + 1 + 1)(2e^{\gamma_j})} = \frac{c}{4}, & \text{if } \gamma_{j-1} \leq 0 \text{ and } \gamma_j \leq 0; \\
\frac{c}{1 + (2e^{\gamma_j})(2e^{\gamma_j})} = \frac{c}{4}, & \text{if } \gamma_{j-1} > 0.
\end{cases}
\]

Similarly, we have

\[
F(\gamma_j) \geq \frac{1}{2} e^{-|\gamma_j|} \quad \text{and} \quad \tilde{F}(\gamma_j) = \frac{1}{1 + e^{\gamma_j}} \geq \frac{1}{2} e^{-|\gamma_j|} \quad \text{for } j = 1, \ldots, q.
\]

As a result, it follows from (5)–(7) and (27) that

\[
\lambda_{\min} \Sigma_n(\beta_0) \geq ce^{-|\gamma_1| + \cdots + |\gamma|}.
\]

From condition (C1), for any \( \epsilon > 0 \), if \( n \) is large enough, we have \( \|X_n\| < \epsilon \log n \). From the definition of \( \nu \) and \( Z'_n \), we have that \( \|\nu\| < c\epsilon (\log n) \), or equivalently \( e^{-\|\nu\|} > n^{-c\epsilon} \). Therefore, \( \lambda_{\min} \Sigma_n(\beta_0) > cn^{-c\epsilon} \), or equivalently, for any \( \epsilon > 0 \), as \( n \) is large enough,

\[
\lambda_{\min} \Sigma_n(\beta_0) > n^{-\epsilon}.
\]

Denote \( \Sigma_i := \Sigma_i(\beta_0), U_i := U_i(\beta_0), \) and \( F_n := F_n(\beta_0) \). From Lemma 1 and (28), we have that

\[
\lambda_{\min} U_i \Sigma_i U'_i \geq cn^{-\epsilon}, \quad i = 1, \ldots, n.
\]

It is easy to prove

\[
\lambda_{\min} \sum_{i=1}^n Z_i Z_i' \geq \lambda_{\min} \sum_{i=1}^n z_i z_i'.
\]

From (29), (30) and condition (C2), we have that

\[
F_n \geq cn^{n-\epsilon} l_n,
\]

and as \( n \to \infty \), for all \( \beta \in \partial \hat{N}_n \),

\[
J_n(\beta) \geq J_n^*(\beta) \geq n^{2\epsilon_0-\epsilon} \to \infty,
\]

where \( \epsilon > 0 \) is small enough and

\[
J_n^*(\beta) := -n^{-\epsilon} \xi^t \sum_{i=1}^n Z_i Z_i' \xi.
\]

Now we prove (23). We divide \( \partial \hat{N}_n \) into \( M_n \) parts, \( BS_1, BS_2, \ldots, BS_{M_n} \), such that the diameter of each part is less than \( n^{-2} \) and \( M_n \leq (2np^2 + 1)^p \). Take any fixed \( \beta_j \in BS_j \) for \( j = 1, \ldots, M_n \), and let \( \xi_j = \beta_j - \beta_0 \). From (17) and condition (C1), we have that

\[
|\xi^t Z_i U_i e_i| \leq cn^{\epsilon_0-\epsilon/2} \log n, \quad E(\xi^t Z_i U_i e_i)^2 \leq c \xi^t Z_i Z_i' \xi.
\]

By using the Bernstein inequality (Bennett, 1962), (34) and condition (C2), we have that

\[
P(G_n(\beta_j) \geq J_n^*(\beta_j)/5) \leq \exp\left(-cn^{2\epsilon_0-2\epsilon} \right) \leq \exp(-cn^{2\epsilon_0-2\epsilon}),
\]

and that

\[
\sum_{n=1}^\infty P \left\{ \bigcup_{1 \leq j \leq M_n} \left[ G_n(\beta_j) \geq J_n^*(\beta_j)/5 \right] \right\} \leq \sum_{n=1}^\infty (2np^2 + 1)^p \exp(-cn^{2\epsilon_0-2\epsilon}) < \infty.
\]

By the Borel–Cantelli lemma, with probability one for \( n \) large, we have that

\[
G_n(\beta_j) < J_n^*(\beta_j)/5, \quad j = 1, \ldots, M_n.
\]
For each \( \beta \in \partial \bar{N}_n \), there exists \( \beta_j \in BS_j \) such that \( \| \beta_j - \beta \| < n^{-2} \). By condition (C1) and (17), we have that

\[
|G_n(\beta) - G_n(\beta_j)| = \left| (\beta - \beta_j)' \sum_{i=1}^{n} Z_i U_i \right| \leq c, \tag{37}
\]

\[
|J_n^*(\beta) - J_n^*(\beta_j)| = n^{-c} \left| (\beta - \beta_j)' \sum_{i=1}^{n} Z_i Z_i'(\beta - \beta_0) + (\beta_j - \beta_0)' \sum_{i=1}^{n} Z_i Z_i'(\beta - \beta_j) \right| \leq c. \tag{38}
\]

Therefore, with probability one, for \( n \) large, (23) holds by (32) and (36)–(38).

Similarly, (24) holds true. Note, for \( r = 1, \ldots, q \),

\[
\sup_{i \geq 1, \beta \in \bar{N}_n} \|W_{tr}(\beta)\| < c. \tag{39}
\]

And the derivative of any element of \( W_{tr}(\beta) \) is bounded, that is, for any \( \lambda \) satisfying \( \| \lambda \| = 1 \),

\[
\sup_{i \geq 1, \beta \in \bar{N}_n} \| \partial \lambda' W_{tr}(\beta) \lambda / \partial \gamma \| < c. \tag{40}
\]

Now we prove (25). It is easy to prove that

\[
\sup_{i \geq 1, \beta \in \bar{N}_n} \| \partial \pi_{tr}(\beta) / \partial \gamma \| < c. \tag{41}
\]

By the mean-value theorem, (41), condition (C1) and \( \| \beta - \beta_0 \| = n^{\alpha_0 - \alpha/2} \) (noting \( \tilde{\beta} \) lies between \( \beta \) and \( \beta_0 \)), for \( n \) large,

\[
|\pi_{tr}(\tilde{\beta}) - \pi_{tr}(\beta_0)| \leq c \| Z_i(\tilde{\beta} - \beta_0) \| \leq n^{-\alpha/3}. \tag{42}
\]

From (42), (39), (28) and (15),

\[
\sum_{r=1}^{q} \left\{ W_{tr}(\beta) (\pi_{tr}(\beta_0) - \pi_{tr}(\tilde{\beta})) - U_{r}(\tilde{\beta}) \Sigma (\tilde{\beta}) U_{r}(\tilde{\beta})' / 4 \right\} < 0,
\]

and (25) holds. By (25), (24) and (23), we get (20), (12) and (13).

**Lemma 4.** Under the conditions of Theorem 1,

\[
\max_{\beta \in \bar{N}_n} \| F_n^{-1/2} H_n(\beta) F_n^{-1/2} - I \| \to 0 \quad \text{in probability.} \tag{43}
\]

**Proof.** Using the same decomposition as in Fahrmeir and Kaufmann (1985), we have that

\[
F_n^{-1/2} H_n(\beta) F_n^{-1/2} - I = A_n^*(\beta) - B_n^* - C_n^*(\beta) - D_n^*(\beta), \tag{44}
\]

where

\[
A_n^*(\beta) := F_n^{-1/2} F_n(\beta) F_n^{-1/2} - I, \quad B_n^* := \sum_{i=1}^{n} \sum_{r=1}^{q} F_n^{-1/2} Z_i W_{tr}(\beta) F_n^{-1/2} e_{ir}, \quad C_n^*(\beta) := \sum_{i=1}^{n} \sum_{r=1}^{q} F_n^{-1/2} Z_i (W_{tr}(\beta) - W_{ir}) Z_i F_n^{-1/2} e_{ir}, \]

\[
D_n^*(\beta) := \sum_{i=1}^{n} \sum_{r=1}^{q} F_n^{-1/2} Z_i W_{tr}(\beta) Z_i F_n^{-1/2} (\mu_{tr}(\beta_0) - \pi_{tr}(\beta)).
\]

To prove (43), we only need to prove

\[
\max_{\beta \in \bar{N}_n} \| A_n^*(\beta) \| \to 0, \tag{45}
\]

\[
\max_{\beta \in \bar{N}_n} \| D_n^*(\beta) \| \to 0, \quad \max_{\beta \in \bar{N}_n} \| C_n^*(\beta) \| \to 0 \quad \text{in probability,} \tag{46}
\]

and

\[
\| B_n^* \| \to 0 \quad \text{in probability.} \tag{47}
\]
By the mean-value theorem, (17), (26), (39), (41), and condition (C1), for any \( \lambda, \| \lambda \| = 1, \beta \in \bar{N}_n \), we have that
\[
|\lambda^t U(\beta) \sigma_i(\beta) U_i(\beta) \lambda - \lambda^t U(\beta) \sigma_i U_i \lambda| \leq c \| Z_i(\beta - \beta_0) \| \leq cn^{-a/3}.
\]
(48)
From (15), (28) and (48), for any \( \delta > 0 \), when \( n \) is large, we have for \( \beta \in \bar{N}_n \) and \( 1 \leq i \leq n \) that
\[
|\lambda^t U(\beta) \sigma_i(\beta) U_i(\beta) \lambda - \lambda^t U(\beta) \sigma_i U_i \lambda| < \delta \lambda^t U(\beta) \sigma_i U_i \lambda.
\]
(49)
From (49), we have (45). Similarly, we can prove (46) by noting (39)–(41).
By (31),
\[
F_n^{-1/2} \leq cn^{(-a+e)/2} I_p.
\]
(50)
Denote \( B_n^* := \sum_{i=1}^n F_n^{-1/2} Z_i W_i Z_i^t F_n^{-1} e_{it} \). By (29), (39), (50) and conditions (C1), noting \( \sum_{i=1}^n \lambda^t F_n^{-1/2} Z_i W_i Z_i^t F_n^{-1} \lambda = 1 \), we conclude that
\[
\text{Var}(\lambda^t B_n^* \lambda) \leq cn^{-a+e} \log(n) \leq c (\log(n))^{2n-a+2e} \to 0.
\]
(51)
Hence, (47) is proved using the Markov inequality. It follows that (43) holds by (47), (46), (45) and (44).

**Lemma 5.** Under the conditions of Theorem 1,
\[
F_n^{-1/2} s_n \to N(0, I_p) \quad \text{in distribution.}
\]
(52)

**Proof.** Let \( \alpha_i := \lambda^t F_n^{-1/2} Z_i U_i \) and \( S_m := \alpha_m^t (y_i - \pi_i(\beta_0)) \). By (50), (17), and condition (C1),
\[
\max_{1 \leq i \leq n} \| \alpha_i \| \leq n^{-(a-e)/3} \to 0.
\]
(53)
Note \( \{y_i, i \geq 1\} \) is bounded. From (53), for any \( \delta > 0 \), when \( n \) is large,
\[
\sum_{i=1}^{n} E S_i^2 I(S_i^2 > \delta^2) \leq \sum_{i=1}^{n} \| \alpha_i \|^2 \mathbb{E} \| y_i - \pi_i(\beta_0) \|^2 I(\| y_i - \pi_i(\beta_0) \| > \delta/\| \alpha_i \|) = 0.
\]
(54)
It follows from the Lindeberg–Feller theorem that (52) holds.

**Proof of Theorem 1.** By Lemma 3 and the mean value theorem for vector valued functions (Heuser, 1981), we have that
\[
s_n(\beta_0) = \left\{ \int_0^1 H_n(\beta_0 + t(\hat{\beta}_n - \beta_0)) dt \right\} (\hat{\beta}_n - \beta_0).
\]
(55)

Lemmas 4, 5, and (55) imply (14), and Theorem 1 is proved by Lemmas 2 and 3.

5. Simulation

Numerical studies are conducted in this section to demonstrate the main results on the MLE \( \hat{\beta}_n \).
Assume that \( S_1 = \{1\} \) and \( S_2 = \{2, 3\} \), i.e., \( k = 3 \) and \( t = 2 \). The simulation datasets are generated from the following model:
\[
\begin{align*}
P(y_{t1} = 1) &= 1 - P(y_{t1} = 0) = \pi_{t1} = F(y_{t1}) = e^{y_{t1}}/(1 + e^{y_{t1}}), \\
P(y_{t2} = 1) &= 1 - P(y_{t2} = 0) = \pi_{t2} = \bar{F}(y_{t1}) F(y_{t2}) = e^{y_{t2}}/(1 + e^{y_{t1}})^2 (1 + e^{y_{t2}}), \\
P(y_{t3} = 1) &= 1 - P(y_{t3} = 0) = \pi_{t3} = \bar{F}(y_{t1}) \bar{F}(y_{t2}) = 1/(1 + e^{y_{t1}}) (1 + e^{y_{t2}}),
\end{align*}
\]
where
\[
(y_{t1}, y_{t2})' = Z_i' \beta', \quad Z_i' = \begin{pmatrix} 1 & X_i & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad i = 1, \ldots, n, \quad \beta' = (\delta_{01}, \lambda_0, \delta_{22}, \lambda_2).
\]
Hence, the response \( (y_{t1}, y_{t2}, y_{t3}) \) follows a multinomial distribution. The covariate \( X_i \) is sampled from the normal distribution with mean 1 and standard deviation 1, which indicates that \( \alpha \) in the condition (C2) equals 1 and the convergence rate of the MLE is of order \( n^{-1/2} \). We considered two setups for the parameter \( \beta \), namely S1: \( \beta_0 = (-1.0, 1.0, -0.5, 1.0)' \) and S2: \( \beta_0 = (1.2, -1.5, -1.5, 1.5)' \). The sample size \( n \) is set to be \( n = 100, 200 \) and \( 400 \). All results are based on 1000 replicates of simulations.
approximation of confidence intervals are much closer to the nominal levels as the sample size gets larger. This demonstrates that the normal approximation is very accurate for moderately large samples sizes.

Table 1

<table>
<thead>
<tr>
<th>n</th>
<th>Mean (standard deviation)</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>δβ1 = −1.0</td>
<td>λ0 = 1.0</td>
</tr>
<tr>
<td>100</td>
<td>−1.038 (0.376)</td>
<td>1.048 (0.294)</td>
</tr>
<tr>
<td>200</td>
<td>−1.012 (0.243)</td>
<td>1.011 (0.187)</td>
</tr>
<tr>
<td>400</td>
<td>−1.005 (0.175)</td>
<td>1.006 (0.138)</td>
</tr>
</tbody>
</table>

|     | δβ1 = 1.2                 | λ0 = −1.5 | δβ2 = −1.5 | λ2 = 1.5 | 0.95^a | 0.99^a |
| 100 | 1.256 (0.420)             | −1.580 (0.371) | −1.680 (0.822) | 1.653 (0.596) | 0.953 | 0.996 |
| 200 | 1.231 (0.273)             | −1.541 (0.235) | −1.589 (0.530) | 1.572 (0.382) | 0.965 | 0.998 |
| 400 | 1.214 (0.196)             | −1.513 (0.170) | −1.541 (0.348) | 1.534 (0.242) | 0.951 | 0.989 |

S1: β0 = (−1.0, 1.0, −0.5, 1.0)’
S2: β0 = (1.2, −1.5, −1.5, 1.5)’

Mean (standard deviation) Coverage
1.513 (0.170) 1.006 (0.200) 0.949 0.990
1.541 (0.235) 1.006 (0.138) 0.949 0.990
1.589 (0.530) 1.006 (0.138) 0.949 0.990
1.653 (0.596) 1.006 (0.138) 0.949 0.990

^a Nominal level.

Theorem 1 shows that the asymptotic distribution of $F_n^{1/2} (\hat{\beta}_n - \beta_0)$ is the standard multivariate normal distribution with mean 0 and identity covariance matrix, hence the asymptotic distribution of $T_n = (\hat{\beta}_n - \beta_0)' F_n (\hat{\beta}_n - \beta_0)$ is the centralized chi-square distribution with df 4. In order to illustrate the accuracy of normal approximation, we calculate the empirical coverage probabilities of the confidence interval with confidence level 1 − ρ defined by $(\hat{\beta}_n - \beta_0)' F_n (\hat{\beta}_n - \beta) \leq \chi^2_k (\rho)$. The simulation results are displayed in Table 1.

As expected, the biases of the MLEs (average $\hat{\beta}_n$ minus the true value of $\beta_0$) are virtually small. As the sample size gets larger, both the biases and the standard errors get smaller. The empirical coverage probabilities of the 95% and 99% confidence intervals are much closer to the nominal levels as the sample size gets larger. This demonstrates that the normal approximation of $\hat{\beta}_n$ is very accurate for moderately large sample sizes.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (11061002, 11101396, 11371101) and the Natural Science Foundation of Guangxi (2011gxnsfa018126).

References