Statistica Sinica **16**(2006), 1335-1365

STRONG LIMIT THEOREMS ON MODEL SELECTION IN GENERALIZED LINEAR REGRESSION WITH BINOMIAL RESPONSES

Guoqi Qian and Yuehua Wu

University of Melbourne and York University

Abstract: We prove a law of iterated logarithm for the maximum likelihood estimator of the parameters in a generalized linear regression model with binomial response. This result is then used to derive an asymptotic bound for the difference between the maximum log-likelihood function and the true log-likelihood. It is further used to establish the strong consistency of some penalized likelihood based model selection criteria. We have shown that, under some general conditions, a model selection criterion will select the simplest correct model almost surely if the penalty term is an increasing function of the model dimension and has an order between $O(\log\log n)$ and $O(n)$. Cases involving the commonly used link functions are discussed for illustration of the results.

Key words and phrases: Generalized linear models, law of the iterated logarithm, maximum likelihood estimator, model selection, strong consistency.

1. Introduction

An important task in linear regression is to identify an optimal subset of available explanatory variables to form a model for best predicting the response variable. We refer to George (2002) and Rao and Wu (2001) for a detailed survey in this area of research. Among the many model selection methods, the classical ones like AIC and BIC are still widely used in practice. It is therefore of interest to investigate the asymptotic properties of model selection criteria which have not yet been established for many problems.

In this paper, we focus on variable selection in generalized linear models with binomial responses. We consider a set of model selection criteria, such as AIC, BIC, C_p and the stochastic complexity criterion, that follow the form of a penalized log-likelihood. We assume that all the explanatory variables affecting the response variable are available in observations, so that selecting the simplest correct model is possible. We establish a strong representation for the maximum log-likelihood function relative to the true log-likelihood under some general conditions. Based on this representation we show that, when the sample size n is sufficiently large, the simplest correct model is selected almost surely if

the penalty term in the selection criterion is an increasing function of the model dimension and is of an order higher than $O(\log\log n)$ but lower than $O(n)$. During this asymptotic study we also obtain the law of iterated logarithm for the maximum likelihood estimator $\hat{\beta}_n$ of the regression coefficient vector β, with its unknown true value denoted as β_0, in the binomial regression model. Namely, $\limsup_{n\to\infty}(n^{-1}\log\log n)^{-1/2}||\hat{\beta}_n - \beta_0|| = c$ almost surely for some constant c, where $||\cdot||$ is the Euclidean norm.

An earlier study related to this paper is Qian and Field (2002a), where the focus is limited to the logistic regression models. Here we consider the more general and applicable binomial regression models where any meaningful link function is allowed. The method developed in Qian and Field (2002a) cannot be carried forward automatically to this seemingly simple generalization, and it entails a substantial new proof technique. This can be seen from the following. First, the log-likelihood function under this generalization loses simplicity and some good properties, for example global convexity, that it possesses under the logistic link. Consequently, it becomes substantially more difficult to derive various almost sure uniform bounds for establishing a strong representation for the log-likelihood function. That the response variable follows a binomial distribution does not help much in easing this difficulty. Second, the lack of specificity about the link function presents another complication. Some general conditions need be sorted out to properly regulate the link function for desired performance of model selection. In this paper we carry out a detailed study of the asymptotic properties of the log-likelihood function with regard to its use in model selection, and how they depend on the link function. Our asymptotic results also provide a justification to the empirical findings that some link functions, such as the logistic and the probit, behave quite well in practice, while some others, such as the complementary log-log and the log-log, do so only in some specific situations (McCullagh and Nelder (1989, Section 4.3.1)).

The paper is organized as follows. Section 2 provides an overview of binomial regression models and a model selection framework. Section 3 presents the main results. Proofs are given in Section 4. In Section 5 we discuss binomial regression model selection. The Appendix contains proofs of the lemmas in Section 4, and verification of the conditions imposed for the link function.

2. Binomial Regression Models and Model Selection

Suppose the response variable Y measures the proportion of "successes" in m independent and identical trials. Thus we can write $Y = Z/m$ where Z follows a binomial(m, π) distribution. Suppose the "success" probability for Y is dependent on a set of explanatory variables $\mathbf{x} = (x_1, \ldots, x_p)^t$. The dependence may be formulated by a binomial regression model $g(\pi) = \mathbf{x}^t\beta$, where $\beta =$

$(\beta_1, \ldots, \beta_p)^t$ is the unknown coefficient parameter vector to be estimated and $g(\cdot)$ is the link function relating the linear predictor $\eta = \mathbf{x}^t\beta$ to the probability π. A wide range of link functions are available, four commonly used in practice are the following.

1. The *logit* or *logistic* link $g_1(\pi) = \log(\pi/(1-\pi))$.
2. The *probit* or *inverse normal* link $g_2(\pi) = \Phi^{-1}(\pi)$, where $\Phi(\cdot)$ is the standard normal distribution function.
3. The *complementary log-log* link $g_3(\pi) = \log\{-\log(1-\pi)\}$.
4. The *log-log* link $g_4(\pi) = -\log\{-\log(\pi)\}$.

We denote $\pi = h(\eta)$ as the inverse link function corresponding to $g(\pi)$. As π is a probability, it is reasonable to regard $h(\cdot)$ as a cumulative distribution function defined on $(-\infty, +\infty)$.

Now let $\mathbf{Y}_n = (y_1, \ldots, y_n)^t$ be n independent observations of Y, with the corresponding binomial "success" probabilities being $\pi_1, \ldots, \pi_n$. The corresponding observations of the explanatory variables are $X_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^t$. Under the binomial regression model we have $\pi_i = h(\mathbf{x}_i^t\beta)$ $(i = 1, \ldots, n)$. The log-likelihood function for the parameter β is then

$$\ell(\beta|\mathbf{Y}_n, X_n) = \sum_{i=1}^{n} \log \binom{m_i}{m_i y_i} - \sum_{i=1}^{n} \rho(\pi_i; y_i, m_i), \tag{1}$$

where

$$\rho(\pi; y, m) = -my \log \pi - m(1-y)\log(1-\pi). \tag{2}$$

Note that $\rho(0; 0, m) = \rho(1; 1, m) = 0$ by convention. When the likelihood function is smooth enough, the maximum likelihood estimator (MLE) of β may be obtained by solving the likelihood equation

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{m_i h'(\mathbf{x}_i^t\beta)}{\pi_i(1-\pi_i)}(y_i - \pi_i)\mathbf{x}_i = 0. \tag{3}$$

Actually, if (3) has a finite solution and the log-likelihood function $\ell(\beta|\mathbf{Y}_n, X_n)$ is strictly concave, the solution is the unique estimator maximizing the likelihood function. This is the case when the link function is the logistic, probit, complementary log-log or log-log; see Wedderburn (1976). But (3) may have multiple solutions in general and not all of them maximize the likelihood function. An example involving multiple stationary points and local maximizers can be constructed if the inverse of the link function is taken to be $h_0(\eta) = \varpi^{-1}\arctan\eta + 0.5 + 0.1\eta^{-2}\sin^2\eta$ where $\varpi = 3.14159\cdots$. We provide details in the Appendix. Now if we can find a solution $\hat{\beta}_n$ of (3) which is a local maximizer of $\ell(\beta|\mathbf{Y}_n, X_n)$ and satisfies $\lim_{n\to\infty} ||\hat{\beta}_n - \beta_0|| = 0$ a.s. with $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^t$

being the finite true value of β, we know $\hat{\beta}_n$, if different from the global maximizer of $\ell(\beta|\mathbf{Y}_n, X_n)$, is asymptotically at least as good as the global maximizer in terms of consistency. In this paper, such a $\hat{\beta}_n$ is taken to be the MLE of β. (In Lehmann and Casella (1998, p.449), the solution of the likelihood equation is called an efficient likelihood estimator.) We show in Theorem 1 that $\hat{\beta}_n$ exists and $\limsup_{n\to\infty}(n^{-1}\log\log n)^{-1/2}||\hat{\beta}_n - \beta_0|| = c$ almost surely under some general conditions. The asymptotic efficiency of $\hat{\beta}_n$ has been established by Fahrmeir and Kaufmann (1985), who proved that $\hat{\beta}_n$ in generalised linear models has an asymptotic normal distribution with the inverse Fisher information matrix as the asymptotic variance, subject to some mild general conditions.

Based on the binomial regression model, the effect of any x variable on Y can be measured by the corresponding β component. There is no need to include in the model those x variables whose β components equal 0. Since the true value β_0 of β has to be estimated, this induces the problem of model selection or variable selection: find those x variables that have significant effects on Y. But the best subset of x variables is better chosen as a whole in terms of a submodel, because an x variable may have significant effect on Y in the presence of some x variables, but not in the presence of other x variables.

Many approaches have been proposed for selecting an optimum model in general parametric settings, see e.g., Rao and Wu (2001) or George (2002) for a detailed survey. In the context of binomial regression models, some of these approaches, such as AIC (Akaike (1973)), BIC (Schwarz (1978)), C_p (Mallows (1973)) and stochastic complexity criterion (SCC, Rissanen (1989, 1996) and Qian and Künsch (1998)), lead to a model selection criterion function that has the following general form for each candidate model $g(\pi_\alpha) = \eta_\alpha = \mathbf{x}_\alpha^t\beta(\alpha)$:

$$S(\eta_\alpha) = \sum_{i=1}^{n} \rho(h(\mathbf{x}_{i\alpha}^t\hat{\beta}_n(\alpha)); y_i, m_i) + C(n, \hat{\beta}_n(\alpha)). \tag{4}$$

Here α is a p_α-component sub-vector of $(1, 2, \ldots, p)$ for indexing; $\mathbf{x}_\alpha$ and $\mathbf{x}_{i\alpha}$ are the corresponding sub-vectors of $\mathbf{x}$ and $\mathbf{x}_i$ indexed by α; $\hat{\beta}_n(\alpha)$ is the MLE of $\beta(\alpha)$ — the sub-vector of β indexed by α. The first term in (4) is basically the negative maximum log-likelihood, while the second term $C(n, \hat{\beta}_n(\alpha))$ is a penalty term measuring the complexity of the underlying candidate model indexed by α. For AIC and C_p, $C(n, \hat{\beta}_n(\alpha)) = p_\alpha$; for BIC, $C(n, \hat{\beta}_n(\alpha)) = (p_\alpha \log n)/2$; and for SCC, $C(n, \hat{\beta}_n(\alpha)) = \log|I_n(\hat{\beta}_n(\alpha)|/2 + \sum_{i=2}^{p_\alpha}\log(|\hat{\beta}_n(\alpha)_i| + \varepsilon n^{-1/4})$ where $I_n(\beta(\alpha))$ is the Fisher information for $\beta(\alpha)$, $\hat{\beta}_n(\alpha)_i$ is the ith component of $\hat{\beta}_n(\alpha)$, and ε is a specified quantity to ensure the invariance of the SCC (see Qian and Künsch (1998) for details).

The candidate model that minimizes (4) is regarded as the optimum model. To see how this optimum model is related to the true model $\eta_0 = \mathbf{x}^t\beta_0$ is one

of the major objectives of this paper. Suppose all the candidate models under consideration have an intercept term that corresponds to the first component of $\mathbf{x}$, and the model $\eta = \mathbf{x}^t\beta$ that includes all the p explanatory variables is the full model. Of the $2^p - 1$ candidate models for selection, we see that each candidate model can be uniquely represented by α. Thus all the candidate models can be classified into $\mathcal{A}_c = \{\alpha : \beta_{0i} = 0 \quad \text{for any} \quad i \notin \alpha\}$ or $\mathcal{A}_w = \{\alpha : \beta_{0i} \neq 0 \quad \text{for some} \quad i \notin \alpha\}$. Each model in $\mathcal{A}_w$ is wrong because it misses at least one x variable that has non-zero effect on Y; each model in $\mathcal{A}_c$ is correct. Still, many models in $\mathcal{A}_c$ may contain some redundant x variables that have no effects on Y. The model in $\mathcal{A}_c$ that contains no redundant x variables is the most desirable. Here we assume the simplest correct model is unique for simplicity of the presentation, which would be the case if all components of $\mathbf{x}$ are linearly independent of each other. In this paper we show that the simplest correct model is almost surely selected by the criterion (4) under some general conditions. If there are multiple simplest correct models among the candidates, the results in the next section suggest that the criterion (4) selects one of the simplest correct models almost surely. In practice, if we find multiple models have criterion values close to the smallest and fit the data well, we take it that there are multiple simplest correct models.

3. Conditions and Main Results

In this paper, c is a constant independent of n and may represent different values in each appearance.

The properties of the MLE $\hat{\beta}_n$ and the model selection criterion $S(\eta_\alpha)$ depend on the link function, the design matrix and the Fisher information in binomial regression models. The Fisher information for β is

$$\begin{aligned} I_n(\beta) &= -E\frac{\partial^2 \ell}{\partial\beta\partial\beta^t} = \sum_{i=1}^{n} \frac{m_i h'(\mathbf{x}_i^t\beta)^2}{\pi_i(1-\pi_i)}\mathbf{x}_i\mathbf{x}_i^t \\ &= X_n^t M_n \,\mathrm{diag}\{\frac{h'(\mathbf{x}_1^t\beta)^2}{\pi_1(1-\pi_1)}, \ldots, \frac{h'(\mathbf{x}_n^t\beta)^2}{\pi_n(1-\pi_n)}\}X_n, \end{aligned}$$

where $M_n = \mathrm{diag}(m_1, \ldots, m_n)$. In the following we describe conditions be needed on various occasions for proving our main results.

(C.1) The function h is a strictly increasing cumulative distribution function, and is second order differentiable with h' and h'' uniformly continuous.

(C.2) There exists a constant $t_0 > 0$ such that $h''(t) \leq 0$ if $t > t_0$, and $h''(t) \geq 0$ if $t < -t_0$.

(C.3) $\sup_{t>t_0} |\frac{d^2}{dt^2}\log(1-h(t))| = \sup_{t>t_0} \left|\frac{h'(t)^2}{(1-h(t))^2} + \frac{h''(t)}{1-h(t)}\right| < \infty,$
$\sup_{t<-t_0} |\frac{d^2}{dt^2}\log h(t)| = \sup_{t<-t_0} \left|\frac{h'(t)^2}{h(t)^2} - \frac{h''(t)}{h(t)}\right| < \infty.$

(C.4) $\inf_{t>t_0}\left\{\frac{h'(t)^2}{(1-h(t))^2}+\frac{h''(t)}{1-h(t)}\right\}>0$, $\inf_{t<-t_0}\left\{\frac{h'(t)^2}{h(t)^2}-\frac{h''(t)}{h(t)}\right\}>0$.

(C.5) If

$$u(t,s)=\left[\frac{(1-h(s))^2h(s)}{(1-h(t))^2}+\frac{(1-h(s))h(s)^2}{h(t)^2}\right]\frac{h'(t)^2}{h'(s)^2},$$
$$v(t,s)=\left[\frac{(1-h(s))^2h(s)}{1-h(t)}-\frac{(1-h(s))h(s)^2}{h(t)}\right]\frac{h''(t)}{h'(s)^2},$$

there exist constants $\Delta_0>0$ and $s_0>0$ such that $\inf_{|t-s|\le\Delta_0,|s|>s_0}\{u(t,s)+v(t,s)\}>0$.

(C.6) $\sup_s\left|\frac{h'(s)h''(s)}{h(s)(1-h(s))}\right|<\infty$.

(C.7) Let $\lambda_1\{G\}\le\cdots\le\lambda_p\{G\}$ be the eigenvalues of a $p\times p$ symmetric matrix G. Then $\lim_{n\to\infty}\lambda_k\{I_n(\beta_0)\}=\infty$, $k=1,\ldots,p$. Also, there exists a constant $d_0>0$ such that $0<\lambda_p\{I_n(\beta_0)\}\le d_0\lambda_1\{I_n(\beta_0)\}$.

(C.8) If $\delta_n=\{\max_{1\le i\le n}m_i^2[h'(\mathbf{x}_i^t\beta_0)/(\pi_{0i}(1-\pi_{0i}))]^2\mathbf{x}_i^tI_n(\beta_0)^{-1}\mathbf{x}_i\}^{1/2}$ where $\pi_{0i}=h(\mathbf{x}_i^t\beta_0)$, then $\delta_n(\log\log\lambda_p\{I_n(\beta_0)\})^{1/2}=o(1)$.

(C.9) $d_1n\le\lambda_p\{I_n(\beta_0)\}\le d_2n$ holds for some positive constants d_1 and d_2.

(C.10) If $\xi_n=\{\max_{1\le i\le n}m_i\mathbf{x}_i^t(X_n^tM_nX_n)^{-1}\mathbf{x}_i\}^{1/2}$, then $\xi_n(\log\log\lambda_p\{X_n^tM_nX_n\})^{1/2}=o(1)$.

(C.11) $d_3n\le\lambda_p\{X_n^tM_nX_n\}\le d_4n$ for some positive constants d_3 and d_4.

(C.12) $\sum_{k=1}^n m_k(x_{ki}x_{kj})^2=O(n)$ for all $i,j=1,\ldots,p$, where x_{ki} is the ith component of $\mathbf{x}_k$.

(C.13) If $\Lambda_n=\mathrm{diag}\{\min\{\frac{h'(\mathbf{x}_1\beta_0)^2}{\pi_{01}(1-\pi_{01})},\pi_{01}(1-\pi_{01})\},\ldots,\min\{\frac{h'(\mathbf{x}_n\beta_0)^2}{\pi_{0n}(1-\pi_{0n})},\pi_{0n}(1-\pi_{0n})\}\}$, there exists a positive constant d_5 such that $\lambda_1\{X_n^tM_n\Lambda_nX_n\}\ge d_5n$.

(C.14) Let $b=(1/2)\min_{1\le i\le p_0}|\beta_0(\alpha_0)_i|$ where α_0 is the correct model in $\mathcal{A}_c$ that has the minimum dimension, and suppose $\beta_0(\alpha_0)_i$ is the ith component of $\beta_0(\alpha)$. Define $A_0=\{\beta:||\beta-\beta_0||\le b\}$. Then there exist a constant $d_6>0$ and a positive integer n_0 such that

$$\sup_{\beta\in A_0}\ell(\beta|\mathbf{Y}_n,X_n)-\sup_{\beta\notin A_0}\ell(\beta|\mathbf{Y}_n,X_n)\ge d_6n\quad\text{a.s. when } n\ge n_0.$$

Note that Conditions (C.1) to (C.6) are about the behaviour of the link function and its various derivatives. It may be difficult to understand (C.5), but it is not required for the proof of our results if (C.2) and (C.4) are satisfied. On the other hand, it can be shown that

$$u(t,s)+v(t,s)=\frac{\frac{d^2}{dt^2}\log h(t)}{\{\frac{d}{ds}\log h(s)\}^2}(1-h(s))-\frac{\frac{d^2}{dt^2}\log(1-h(t))}{\{\frac{d}{ds}\log(1-h(s))\}^2}h(s).\qquad(5)$$

Thus (C.5) implies that, when t is in a neighbourhood of s, the coefficient of h in (5) should not go to zero in the limit as $s\to+\infty$, and the coefficient of $1-h$

in (5) should not go to zero in the limit either as $s \to -\infty$. In the Appendix we show the following.

1. For the logistic link, Conditions (C.1), (C.2), (C.3), (C.5) and (C.6) hold while (C.4) does not.
2. For the probit link, condtions (C.1) to (C.4) and (C.6) hold while (C.5) does not.
3. For the complementary log-log link, Conditions (C.1), (C.2) and (C.6) hold while (C.3) to (C.5) do not, but (C.3) and (C.5) hold if one only considers $h(t) < 1 - \delta'$ for some constant δ'.
4. For the log-log link, Conditions (C.1), (C.2) and (C.6) hold while (C.3) to (C.5) do not, but (C.3) and (C.5) hold if one focuses on $h(t) > \delta''$ for some constant δ''.

These four link functions and their first three derivatives are plotted in Figure 1.

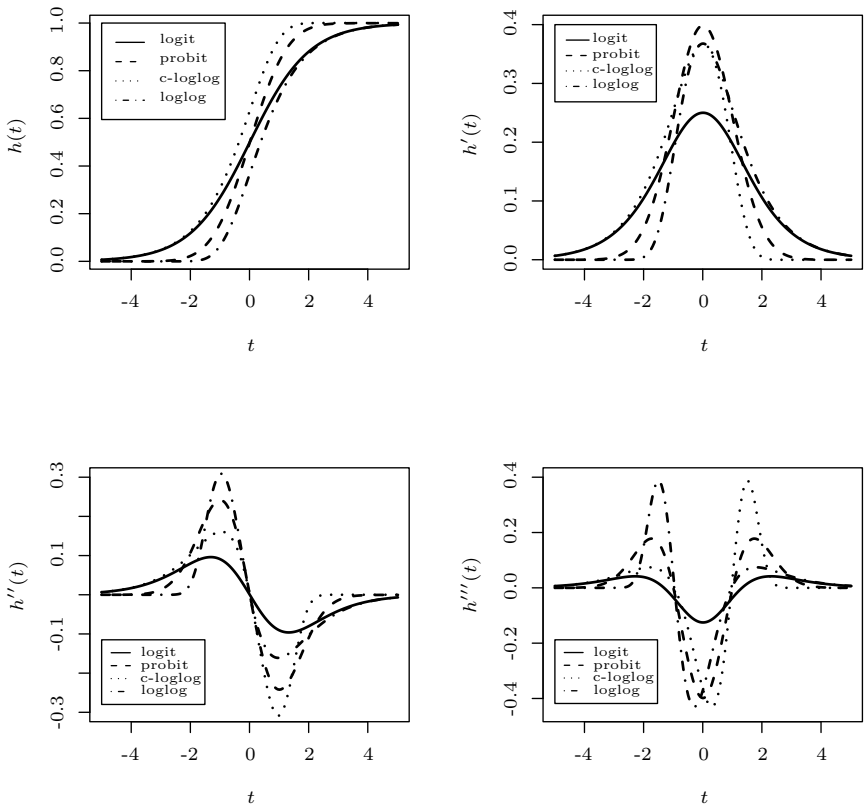

Figure 1. Plots of the four inverse link functions (logit, probit, complementary log-log and log-log) and their first three derivatives.

The conditions (C.7) to (C.13) are essentially about the behaviour of the explanatory variables $\mathbf{x}$. They suggest that most of the $\mathbf{x}$ observations should be finite and stay away from 0. One can follow Qian and Field (2002a) to provide some sufficient conditions for (C.7) to (C.13) by assuming the $\mathbf{x}$ variables are random, but we do not pursue that here. Condition (C.14) is about the behaviour of the log-likelihood function relative to the full model: its maximum value is attained in a neighbourhood of the true value β_0 and is distinctly greater than any log-likelihood outside this neighbourhood. Condition (C.14) becomes quite natural by the results of Theorems 1 and 2 plus the assumption that the MLE can be uniquely solved at (3).

The main results of this paper are listed below.

Theorem 1. *Suppose Conditions* (C.1) *to* (C.4) (*or alternatively* (C.1), (C.3) *and* (C.5)) *hold. Further suppose Conditions* (C.6) *to* (C.13) *hold. Then for any correct model* $\alpha \in \mathcal{A}_c$, *there exists an estimator* $\hat{\beta}_n(\alpha)$ *such that* $\hat{\beta}_n(\alpha)$ *is a local maximizer of* $\ell(\beta|\mathbf{Y}_n, X_{n\alpha})$ *and*

$$||\hat{\beta}_n(\alpha) - \beta_0(\alpha)|| = O((n^{-1}\log\log n)^{\frac{1}{2}}) \quad a.s.. \tag{6}$$

Further, there exists a constant $c > 0$ *such that for* $\alpha \in \mathcal{A}_c$

$$\limsup_{n\to\infty} \frac{||\hat{\beta}_n(\alpha) - \beta_0(\alpha)||}{(n^{-1}\log\log n)^{\frac{1}{2}}} = c \quad \text{a.s..} \tag{7}$$

Theorem 2. *Under the same conditions as given in Theorem* 1 *we have, for any correct model* $\alpha \in \mathcal{A}_c$,

$$0 \le \ell(\hat{\beta}_n(\alpha)|\mathbf{Y}_n, X_{n\alpha}) - \ell(\beta_0(\alpha)|\mathbf{Y}_n, X_{n\alpha}) = O(\log\log n) \quad a.s., \tag{8}$$

where $X_{n\alpha}$ *is the matrix comprising those columns of* X_n *indexed by* α; *equivalently,*

$$\begin{aligned} 0 &\le \sum_{k=1}^{n} \{\rho(h(\mathbf{x}_{k\alpha}^t \beta_0(\alpha)); y_k, m_k) - \rho(h(\mathbf{x}_{k\alpha}^t \hat{\beta}_n(\alpha)); y_k, m_k)\} \\ &= O(\log\log n) \quad a.s., \end{aligned} \tag{9}$$

where $\mathbf{x}_{k\alpha}$ $(k = 1, \ldots, n)$ *is the subvector of* $\mathbf{x}_k$ *indexed by* α.

Theorem 3. *In addition to the conditions of Theorem* 1, *suppose Condition* (C.14) *holds and* $\ell(\hat{\beta}_n|\mathbf{Y}_n, X_n) = \sup_{\beta\in A_0} \ell(\beta|\mathbf{Y}_n, X_n)$. *Then for any incorrect model* $\alpha \in \mathcal{A}_w$,

$$\limsup_{n\to\infty} n^{-1}\{\ell(\hat{\beta}_n(\alpha)|\mathbf{Y}_n, X_{n\alpha}) - \ell(\beta_0|\mathbf{Y}_n, X_n)\} < 0 \quad a.s.; \tag{10}$$

equivalently,

$$\liminf_{n\to\infty} \sum_{k=1}^{n} n^{-1}\{\rho(h(\mathbf{x}_{k\alpha}^t\hat{\beta}_n(\alpha)); y_k, m_k) - \rho(h(\mathbf{x}_k^t\beta_0); y_k, m_k)\} > 0 \quad \text{a.s..} \tag{11}$$

From Theorems 2 and 3 we know that the maximum log-likelihood for any correct model is almost surely greater than the log-likelihood of the true model, with the difference bounded by $O(\log\log n)$ almost surely. On the other hand, the maximum log-likelihood for any incorrect model is almost surely smaller than the log-likelihood of the true model by a term of order $|O(n)|$. Therefore, if we carry out a model selection by minimizing (4), we almost surely select the simplest correct model in $\mathcal{A}_c$ provided the penalty term $C(n, \hat{\beta}_n(\alpha))$ is an increasing function of the model dimension p_α and is of an order higher than $O(\log\log n)$ but smaller than $O(n)$. We call a model selection criterion *strongly consistent* if it selects the simplest correct model almost surely. From the above, we have the following.

Theorem 4. *Consider a binomial regression model. Under the conditions of Theorem* 1 *and* (C.14), *the model selection criterion based on stochastic complexity and the BIC criterion are both strongly consistent, while the AIC criterion is not strongly consistent.*

Proof. It is easy to see that the criterion BIC is strongly consistent because it has a penalty term $C(n, \hat{\beta}_n(\alpha)) = (p_\alpha \log n)/2$, while AIC is not strongly consistent because $C(n, \hat{\beta}_n(\alpha)) = p_\alpha$. Because the Fisher information $|I(\beta(\alpha))|$ is typically of order $O(n^{p_\alpha})$, it follows that the stochastic complexity criterion is also strongly consistent.

4. Proof of the Results

In the proofs we make use of the local convexity properties of the negative log-likelihood function, and locally approximate it with bounded quadratic errors. The main difficulties lie on how to properly regulate the link function, and on establishing uniform bounds for the error term in the almost sure expansion of the log-likelihood function. The idea of using convexity is widely seen in the context of M-estimators for linear models, see e.g., Rao and Zhao (1992) and others.

First of all we define a sequence of real numbers $\{\nu_n\}$ such that, for both $a = 1$ and 2,

$$\nu_n^a \uparrow \infty, \quad \nu_n^a \xi_n (\log\log n)^{\frac{1}{2}} \to 0 \ \text{ and } \ \nu_n^a (n^{-1}\log\log n)^{\frac{1}{2}} \downarrow 0. \tag{12}$$

Using $\{\nu_n\}$ we introduce the sequences $A_n = \{\beta\colon ||\beta - \beta_0|| \le \nu_n (n^{-1}\log\log n)^{1/2}\}$, $\partial A_n = \{\beta : ||\beta - \beta_0|| = \nu_n (n^{-1}\log\log n)^{1/2}\}$, $B_n = \{\beta : ||\beta - \beta_0|| \le \nu_n^2$

$(n^{-1}\log\log n)^{1/2}\}$, and $\partial B_n = \{\beta : ||\beta - \beta_0|| = \nu_n^2(n^{-1}\log\log n)^{1/2}\}$, so $A_1 \supset A_2 \supset A_3 \supset \cdots$, $B_1 \supset B_2 \supset B_3 \supset \cdots$, and $B_n \supset A_n$. We also define

$$H(\beta) = \sum_{k=1}^{n}\{\rho(h(\mathbf{x}_k^t\beta); y_k, m_k) - \rho(h(\mathbf{x}_k^t\beta_0); y_k, m_k)\}. \tag{13}$$

To prove the theorems we need some lemmas.

Lemma 1. *Let*

$$\begin{aligned} D(t;s,y) &= \rho(h(t);y,m) - \rho(h(s);y,m) - \frac{d}{dt}\rho(h(t);y,m)|_{t=s}(t-s) \\ &= -my\log\frac{h(t)}{h(s)} - m(1-y)\log\frac{1-h(t)}{1-h(s)} - m\left[\frac{1-y}{1-h(s)} - \frac{y}{h(s)}\right]h'(s)(t-s) \end{aligned}$$

and suppose (C.1) *and* (C.3) *hold.*

(R.1) *There exists a constant c such that $|D(t;s,y)| \le cm(t-s)^2$ for any real numbers s, t, and $y \in [0,1]$.*

(R.2) $D(t;s,h(s)) \ge 0$.

Lemma 2. *Let $K(t,s) = \rho(h(t);h(s),m) - \rho(h(s);h(s),m)$ and suppose* (C.1), (C.2) *and* (C.4) (*or alternatively* (C.1) *and* (C.5)) *hold. Then there exist positive constants c and Δ such that*

$$K(t,s) \ge c\min\left\{\frac{mh'(s)^2}{h(s)(1-h(s))}, mh(s)(1-h(s))\right\}(t-s)^2$$

for any s and t satisfying $|t-s| \le \Delta$.

Lemma 3. *Let $R(t,s) = \log(h(t)/(1-h(t))) - \log(h(s)/(1-h(s))) - (h'(s)/(h(s)(1-h(s))))(t-s)$ and suppose* (C.1) *and* (C.3) *hold. Then there exists a positive constant c such that $|R_t''(t,s)| \le c$ and $|R(t,s)| \le c(t-s)^2$ for any t and s. Further, for any t_1 and t_2, $|R(t_1,s) - R(t_2,s)| \le c(|t_1-s| + |t_2-s|)|t_1-t_2|$.*

Lemma 4. *Under* (C.7) *and* (C.8) *we have*

$$\begin{aligned} &\limsup_{n\to\infty} \frac{\pm\sum_{i=1}^n m_i h'(\mathbf{x}_i^t\beta_0)\pi_{0i}^{-1}(1-\pi_{0i})^{-1}(y_i-\pi_{0i})x_{ij}}{\{2I_n(\beta_0)(j,j)\log\log I_n(\beta_0)(j,j)\}^{\frac{1}{2}}} \\ &\quad = 1 \text{ a.s. for } j = 1,\ldots,p. \end{aligned} \tag{14}$$

Here x_{ij} is the jth component of $\mathbf{x}_i$ and $I_n(\beta_0)(j,j)$ is the component of $I_n(\beta_0)$ at the jth row and jth column. If, in addition, (C.9) *is satisfied, then we have*

$$\begin{aligned} \frac{\partial\ell}{\partial\beta}\Big|_{\beta=\beta_0} &= \sum_{i=1}^n \frac{m_i h'(\mathbf{x}_i^t\beta_0)}{\pi_{0i}(1-\pi_{0i})}(y_i-\pi_{0i})\mathbf{x}_i \\ &= X_n^t M_n \operatorname{diag}\{\frac{h'(\mathbf{x}_1^t\beta_0)}{\pi_{01}(1-\pi_{01})},\ldots,\frac{h'(\mathbf{x}_n^t\beta_0)}{\pi_{0n}(1-\pi_{0n})}\}(\mathbf{Y}_n - \Pi_{n0}) \\ &= O(\sqrt{n\log\log n}) \quad a.s., \end{aligned} \tag{15}$$

where $\Pi_{n0} = (\pi_{01}, \ldots, \pi_{0n})^t$ *is the true value of* $\Pi_n = (\pi_1, \ldots, \pi_n)^t$.

Lemma 5.(Law of the Iterated Logarithm). *Let* $\{Z_n, n \geq 1\}$ *be a sequence of independent random variables with* $EZ_n = 0$, $EZ_n^2 = \sigma_n^2$ *and* $B_n = \sum_{k=1}^n \sigma_k^2 \to \infty$. *If* $|Z_n| \leq o\{(B_n/\log\log B_n)^{1/2}\}$ *a.s., then*

$$\limsup_{n\to\infty} \frac{\pm\sum_{k=1}^n Z_k}{\{2B_n \log\log B_n\}^{\frac{1}{2}}} = \limsup_{n\to\infty} \frac{|\sum_{k=1}^n Z_k|}{\{2B_n \log\log B_n\}^{\frac{1}{2}}} = 1 \quad \text{a.s..}$$

The proof can be found in Chow and Teicher (1997, pp.373-374) or Petrov (1995, pp.239-246).

Lemma 6. *Under* (C.1), (C.3), (C.6), (C.7) *and* (C.9)−(C.12), *the function* $H(\beta) = \sum_{k=1}^n \{\rho(h(\mathbf{x}_k^t\beta); y_k, m_k) - \rho(h(\mathbf{x}_k^t\beta_0); y_k, m_k)\}$ *is strictly convex on* $\beta \in B_n$ *when* n *is sufficiently large, for all sample sequences* $\{y_1, \ldots, y_n, \ldots\}$ *except a subset with probability* 0. *Further, the eigenvalues of* $\partial^2 H(\beta)/(\partial\beta\partial\beta^t)$ *at* $\beta = \beta_0$ *satisfy*

$$cn \leq \lambda_1\Big\{\frac{\partial^2 H(\beta)}{\partial\beta\partial\beta^t}\Big|_{\beta=\beta_0}\Big\} \leq \cdots \leq \lambda_p\Big\{\frac{\partial^2 H(\beta)}{\partial\beta\partial\beta^t}|_{\beta=\beta_0}\Big\} \leq Cn \quad a.s.$$

for some positive constants c *and* C *when* n *is sufficiently large.*

The proofs of Lemmas 1, 2, 3, 4 and 6 are given in the Appendix.

Proof of Theorem 1. It suffices to prove (6) for the full model:

$$||\hat{\beta}_n - \beta_0|| = O((n^{-1}\log\log n)^{\frac{1}{2}}) \quad \text{a.s..} \tag{16}$$

By the definition of $H(\beta)$ we have

$$\begin{aligned} H(\beta) &= \sum_{k=1}^n \{\rho(h(\mathbf{x}_k^t\beta); \pi_{0k}, m_k) - \rho(h(\mathbf{x}_k^t\beta_0); \pi_{0k}, m_k)\} \\ &\quad + \sum_{k=1}^n \{\rho(h(\mathbf{x}_k^t\beta); y_k, m_k) - \rho(h(\mathbf{x}_k^t\beta); \pi_{0k}, m_k)\} \\ &\quad - \sum_{k=1}^n \{\rho(h(\mathbf{x}_k^t\beta_0); y_k, m_k) - \rho(h(\mathbf{x}_k^t\beta_0); \pi_{0k}, m_k)\} \\ &\stackrel{\text{denote}}{=} T_1 + T_2 - T_3. \end{aligned}$$

Here one can show that $T_1 = \sum_{k=1}^n K(\mathbf{x}_k^t\beta, \mathbf{x}_k^t\beta_0)$ by the definition of $K(t,s)$ in Lemma 2, and

$$\begin{aligned} T_2 - T_3 &= -\sum_{k=1}^n \frac{m_k h'(\mathbf{x}_k^t\beta_0)}{\pi_{0k}(1-\pi_{0k})}(y_k - \pi_{0k})\mathbf{x}_k^t(\beta - \beta_0) \\ &\quad - \sum_{k=1}^n m_k R(\mathbf{x}_k^t\beta, \mathbf{x}_k^t\beta_0)(y_k - \pi_{0k}) \stackrel{\text{denote}}{=} T_4 - T_5 \end{aligned}$$

by the definition of $R(t,s)$ in Lemma 3.

We first show that T_5 satisfies

$$\sup_{\beta\in A_n}|T_5| = o(1)\nu_n^2\log\log n \quad \text{a.s..} \tag{17}$$

For each integer n there exists a unique integer m such that $2^m < n \le 2^{m+1}$. Define a subset $C_m = \{\bar\beta_{mj} \in A_{2^m} : j = 1,\dots,h_m\}$ such that for any $\beta \in A_{2^m}$, there exists j to satisfy $||\beta - \bar\beta_{mj}|| \le \nu_{2^m}2^{-m/2}\sqrt{\log\log 2^m}/m$. It is easy to see that at least one such C_m exists with $h_m \le 4^p m^p$. Also define another subset $D_m = \{\tilde\beta_{ml} \in A_{2^m} : l = 1,\dots,r_m\}$ such that for any $\beta \in A_{2^m}$, there exists l to satisfy $||\beta - \tilde\beta_{ml}|| \le \nu_{2^m}2^{-m/2}\sqrt{\log\log 2^m}/2^m$. It is easy to see that at least one such D_m exists with $r_m \le 4^p 2^{mp}$. For each $l = 1,\dots,r_m$, let

$$A_{2^m,l} = \left\{\beta : ||\beta - \tilde\beta_{ml}|| \le \nu_{2^m}2^{-\frac{m}{2}}\frac{\sqrt{\log\log 2^m}}{2^m}\right\}$$

be a hyper-ball centered at $\tilde\beta_{ml}$. By the definition of D_m we have $\bigcup_{l=1}^{r_m} A_{2^m,l} \supseteq A_{2^m}$. If $Z_k(\beta) = m_k R(\mathbf{x}_k^t\beta, \mathbf{x}_k^t\beta_0)(y_k - \pi_{0k})$, it has a mean of 0, and we see that for each $\beta \in A_{2^m}$, there exist l with $||\beta - \tilde\beta_{ml}|| \le \nu_{2^m}2^{-m/2}\sqrt{\log\log 2^m}/2^m$ and j with $||\tilde\beta_{ml} - \bar\beta_{mj}|| \le \nu_{2^m}2^{-m/2}\sqrt{\log\log 2^m}/m$, such that $\beta \in A_{2^m,l}$ and

$$|Z_k(\beta)| \le |Z_k(\beta) - Z_k(\tilde\beta_{ml})| + |Z_k(\tilde\beta_{ml}) - Z_k(\bar\beta_{mj})| + |Z_k(\bar\beta_{mj})|. \tag{18}$$

From (18) it follows that for any $\varepsilon > 0$

$$\begin{aligned}
&\left\{\max_{2^m<n\le 2^{m+1}}\sup_{\beta\in A_{2^m}}\Big|\sum_{k=1}^n Z_k(\beta)\Big| \ge \varepsilon\nu_{2^m}^2\log\log 2^m\right\}\\
&= \bigcup_{n=2^m+1}^{2^{m+1}}\left\{\sup_{\beta\in A_{2^m}}\Big|\sum_{k=1}^n Z_k(\beta)\Big| \ge \varepsilon\nu_{2^m}^2\log\log 2^m\right\}\\
&\subseteq \left[\bigcup_{n=2^m+1}^{2^{m+1}}\bigcup_{l=1}^{r_m}\left\{\sup_{\beta\in A_{2^m,l}}\Big|\sum_{k=1}^n[Z_k(\beta) - Z_k(\tilde\beta_{ml})]\Big| \ge \frac{1}{3}\varepsilon\nu_{2^m}^2\log\log 2^m\right\}\right]\\
&\quad\bigcup\left[\bigcup_{n=2^m+1}^{2^{m+1}}\bigcup_{l=1}^{r_m}\bigcup_{j\in S_l}\left\{\Big|\sum_{k=1}^n[Z_k(\tilde\beta_{ml}) - Z_k(\bar\beta_{mj})]\Big| \ge \frac{1}{3}\varepsilon\nu_{2^m}^2\log\log 2^m\right\}\right]\\
&\quad\bigcup\left[\bigcup_{n=2^m+1}^{2^{m+1}}\bigcup_{j=1}^{h_m}\left\{\Big|\sum_{k=1}^n Z_k(\bar\beta_{mj})\Big| \ge \frac{1}{3}\varepsilon\nu_{2^m}^2\log\log 2^m\right\}\right]
\end{aligned}$$

$$= \left[\bigcup_{l=1}^{r_m}\bigcup_{n=2^m+1}^{2^{m+1}}\left\{\sup_{\beta\in A_{2^m,l}}\Big|\sum_{k=1}^{n}[Z_k(\beta)-Z_k(\tilde\beta_{ml})]\Big|\ge\frac{1}{3}\varepsilon\nu_{2^m}^2\log\log 2^m\right\}\right]$$
$$\bigcup\left[\bigcup_{l=1}^{r_m}\bigcup_{j\in S_l}\left\{\max_{2^m<n\le 2^{m+1}}\Big|\sum_{k=1}^{n}[Z_k(\tilde\beta_{ml})-Z_k(\bar\beta_{mj})]\Big|\ge\frac{1}{3}\varepsilon\nu_{2^m}^2\log\log 2^m\right\}\right]$$
$$\bigcup\left[\bigcup_{j=1}^{h_m}\left\{\max_{2^m<n\le 2^{m+1}}\Big|\sum_{k=1}^{n}Z_k(\bar\beta_{mj})\Big|\ge\frac{1}{3}\varepsilon\nu_{2^m}^2\log\log 2^m\right\}\right], \tag{19}$$

where $S_l=\{j:||\tilde\beta_{ml}-\bar\beta_{mj}||\le\nu_{2^m}2^{-m/2}\sqrt{\log\log 2^m}/m.\}$ for each l given.

By Lemma 3, (21), the Cauchy-Schwarz inequality and (C.11), we have

$$\sum_{k=1}^{n}|Z_k(\beta)-Z_k(\tilde\beta_{ml})|\le 2\sum_{k=1}^{n}m_k|R(\mathbf{x}_k^t\beta,\mathbf{x}_k^t\beta_0)-R(\mathbf{x}_k^t\tilde\beta_{ml},\mathbf{x}_k^t\beta_0)|$$
$$\le 2c_3\sum_{k=1}^{n}m_k||\mathbf{x}_k||^2(||\beta-\beta_0||+||\tilde\beta_{ml}-\beta_0||)||\beta-\tilde\beta_{ml}||$$
$$\le c2^{-m}\nu_{2^m}^2\log\log 2^m=o(1)\nu_{2^m}^2\log\log 2^m\quad\text{for every }\beta\in A_{2^m,l},$$

which implies that for large m,

$$P\Big\{\bigcup_{l=1}^{r_m}\bigcup_{n=2^m+1}^{2^{m+1}}\Big[\sup_{\beta\in A_{2^m,l}}\Big|\sum_{k=1}^{n}(Z_k(\beta)-Z_k(\tilde\beta_{ml})\Big|\ge\frac{1}{3}\varepsilon\nu_{2^m}^2\log\log 2^m\Big]\Big\}=0. \tag{20}$$

By the definition of ξ_n, and (C.11),

$$\xi_{2^{m+1}}^2\ge m_k\mathbf{x}_k^t(X_{2^{m+1}}^tM_{2^{m+1}}X_{2^{m+1}})^{-1}\mathbf{x}_k$$
$$\ge m_k\mathbf{x}_k^t\mathbf{x}_k(\lambda_p\{X_{2^{m+1}}^tM_{2^{m+1}}X_{2^{m+1}}\})^{-1}\ge d_4^{-1}2^{-(m+1)}m_k||\mathbf{x}_k||^2 \tag{21}$$

for any $k=1,\dots,2^{m+1}$. For any $\tilde\beta_{ml}$ and $\bar\beta_{mj}$ with $j\in S_l$, it follows from Lemma 3, (21), the Cauchy-Schwarz inequality and (C.11), that

$$|Z_k(\tilde\beta_{ml})-Z_k(\bar\beta_{mj})|\le 4c_3d_4m^{-1}\xi_{2^{m+1}}^2\nu_{2^m}^2\log\log 2^m,\quad k=1,\dots,2^{m+1}, \tag{22}$$

$$\sum_{k=1}^{2^{m+1}}E[Z_k(\tilde\beta_{ml})-Z_k(\bar\beta_{mj})]^2$$
$$\le\sum_{k=1}^{2^{m+1}}m_k^2c_3^2[|\mathbf{x}_k^t(\tilde\beta_{ml}-\beta_0)|+|\mathbf{x}_k^t(\bar\beta_{mj}-\beta_0)|]^2|\mathbf{x}_k^t(\tilde\beta_{ml}-\bar\beta_{mj})|^2$$

$$\le c_3^2 \max_{1\le k\le 2^{m+1}} m_k ||\mathbf{x}_k||^2 [||\tilde{\beta}_{ml} - \beta_0|| + ||\bar{\beta}_{mj} - \beta_0||]^2 (\tilde{\beta}_{ml} - \bar{\beta}_{mj})^t \Big(\sum_{k=1}^{2^{m+1}} m_k \mathbf{x}_k \mathbf{x}_k^t\Big)(\tilde{\beta}_{ml} - \bar{\beta}_{mj})$$
$$\le 16 c_3^2 d_4^2 \xi_{2^{m+1}}^2 \nu_{2^m}^4 m^{-2} (\log\log 2^m)^2. \tag{23}$$

By (23) and the well-known relationship among the median, mean and variance, $|\mathrm{med}(X) - E(X)| \le \sqrt{\mathrm{Var}(X)}$ (Chow and Teicher (1997, p.109)), it follows that

$$\Big|\mathrm{med}\Big(\sum_{k=n+1}^{2^{m+1}} [Z_k(\tilde{\beta}_{ml}) - Z_k(\bar{\beta}_{mj})]\Big)\Big| \le 4 c_3 d_4 \xi_{2^{m+1}} \nu_{2^m}^2 m^{-1} \log\log 2^m$$
$$= o(1) \nu_{2^m}^2 \log\log 2^m. \tag{24}$$

By (24) and Lévy's inequality,

$$P\Big\{ \max_{2^m < n \le 2^{m+1}} \Big| \sum_{k=1}^{n} [Z_k(\tilde{\beta}_{ml}) - Z_k(\bar{\beta}_{mj})] \Big| \ge \frac{1}{3} \varepsilon \nu_{2^m}^2 \log\log 2^m \Big\}$$
$$\le P\Big\{ \max_{2^m < n \le 2^{m+1}} \Big| \sum_{k=1}^{n} [Z_k(\tilde{\beta}_{ml}) - Z_k(\bar{\beta}_{mj})] - \mathrm{med}\Big(\sum_{k=n+1}^{2^{m+1}} [Z_k(\tilde{\beta}_{ml}) - Z_k(\bar{\beta}_{mj})]\Big)\Big| \ge c \nu_{2^m}^2 \log\log 2^m \Big\}$$
$$\le 2P\Big\{ \Big| \sum_{k=1}^{2^{m+1}} [Z_k(\tilde{\beta}_{ml}) - Z_k(\bar{\beta}_{mj})] \Big| \ge c \nu_{2^m}^2 \log\log 2^m \Big\}. \tag{25}$$

Before we proceed, we give Bernstein's inequality, it can be found in Chow and Teicher (1997, p.111).

Bernstein's Inequality. If $S_n = \sum_{j=1}^n Z_j$ where Z_j's are independent random variables with $EZ_j = 0$ and $|Z_j| \le a$ for each j, then for any $x > 0$,

$$P\{|S_n| > x\} \le 2\exp\{\frac{-x^2}{2(ax + \sum_{j=1}^n EZ_j^2)}\}.$$

By (25), (22), (23), Bernstein's inequality, (C.10) and (C.11), it follows that

$$P\Big\{ \bigcup_{l=1}^{r_m} \bigcup_{j\in S_l} \Big\{ \max_{2^m < n \le 2^{m+1}} \Big| \sum_{k=1}^{n} [Z_k(\tilde{\beta}_{ml}) - Z_k(\bar{\beta}_{mj})] \Big| \ge \frac{1}{3} \varepsilon \nu_{2^m}^2 \log\log 2^m \Big\} \Big\}$$
$$\le 2 \sum_{l=1}^{r_m} \sum_{j\in S_l} P\Big\{ \Big| \sum_{k=1}^{2^{m+1}} [Z_k(\tilde{\beta}_{ml}) - Z_k(\bar{\beta}_{mj})] \Big| \ge c \nu_{2^m}^2 \log\log 2^m \Big\}$$

$$\le 4^{2p+1}m^p 2^{mp}\exp\Big\{\frac{-c\nu_{2^m}^4(\log\log 2^m)^2}{m^{-1}\xi_{2^{m+1}}^2\nu_{2^m}^4(\log\log 2^m)^2+m^{-2}\xi_{2^{m+1}}^2\nu_{2^m}^4(\log\log 2^m)^2}\Big\}$$
$$\le 4^{2p+1}m^p 2^{mp}\exp\{-cm\xi_{2^{m+1}}^{-2}\} \le \exp\{-cm\log\log 2^{m+1}\}, \tag{26}$$

when m is sufficiently large.

For any $\bar\beta_{mj}$, it follows from Lemma 3, (21), the Cauchy-Schwarz inequality and (C.11) that

$$|Z_k(\bar\beta_{mj})| \le 2c_3d_4\xi_{2^{m+1}}^2\nu_{2^m}^2\log\log 2^m, \quad k=1,\dots,2^{m+1}, \tag{27}$$

$$\sum_{k=1}^{2^{m+1}} EZ_k(\bar\beta_{mj})^2 \le \sum_{k=1}^{2^{m+1}} m_k^2c_3^2[\mathbf{x}_k^t(\bar\beta_{mj}-\beta_0)]^4$$
$$\le c_3^2\max_{1\le k\le 2^{m+1}} m_k||\mathbf{x}_k||^2||\bar\beta_{mj}-\beta_0||^2(\bar\beta_{mj}-\beta_0)^t\Big(\sum_{k=1}^{2^{m+1}} m_k\mathbf{x}_k\mathbf{x}_k^t\Big)(\bar\beta_{mj}-\beta_0)$$
$$\le c_3^2d_4 2^{m+1}\xi_{2^{m+1}}^2||\bar\beta_{mj}-\beta_0||^4\lambda_p\{X_{2^{m+1}}^tM_{2^{m+1}}X_{2^{m+1}}\}$$
$$\le 4c_3^2d_4^2\xi_{2^{m+1}}^2\nu_{2^m}^4(\log\log 2^m)^2. \tag{28}$$

Similar to (24), following (C.10) and (28), we have

$$\Big|\mathrm{med}\Big(\sum_{k=n+1}^{2^{m+1}} Z_k(\bar\beta_{mj})\Big)\Big| \le 2c_3d_4\xi_{2^{m+1}}\nu_{2^m}^2\log\log 2^m = o(1)\nu_{2^m}^2\log\log 2^m. \tag{29}$$

Therefore, using the Lévy and Beinstein's inequalities, (12), (27), (28), (29), (C.10) and (C.11), we find that, when m is sufficiently large,

$$P\Big\{\bigcup_{j=1}^{h_m}\Big\{\max_{2^m<n\le 2^{m+1}}\Big|\sum_{k=1}^n Z_k(\bar\beta_{mj})\Big| \ge \frac{1}{3}\varepsilon\nu_{2^m}^2\log\log 2^m\Big\}\Big\}$$
$$\le \sum_{j=1}^{h_m} P\Big\{\max_{2^m<n\le 2^{m+1}}\Big|\sum_{k=1}^n Z_k(\bar\beta_{mj}) - \mathrm{med}\Big(\sum_{k=n+1}^{2^{m+1}} Z_k(\bar\beta_{mj})\Big)\Big| \ge c\nu_{2^m}^2\log\log 2^m\Big\}$$
$$\le 2\sum_{j=1}^{h_m} P\Big\{\Big|\sum_{k=1}^{2^{m+1}} Z_k(\bar\beta_{mj})\Big| \ge c\nu_{2^m}^2\log\log 2^m\Big\}$$
$$\le \exp\Big\{-c\nu_{2^m}^2\log\log 2^{m+1}\Big\}. \tag{30}$$

Combining (19), (20), (26) and (30), we see for any $\varepsilon>0$ and for large m,

$$\sum_{m=1}^{\infty} P\Big\{\max_{2^m<n\le 2^{m+1}}\sup_{\beta\in A_{2^m}}|\sum_{k=1}^n Z_k(\beta)| \ge \varepsilon\nu_{2^m}^2\log\log 2^m\Big\}$$
$$\le c+\sum_{m=1}^{\infty}\{e^{-cm\log\log 2^{m+1}}+e^{-c\nu_{2^m}^2\log\log 2^{m+1}}\} < \infty. \tag{31}$$

It follows from the Borel-Cantelli Lemma that

$$\max_{2^m<n\le 2^{m+1}} \sup_{\beta\in A_{2^m}} |\sum_{k=1}^{n} Z_k(\beta)| = o(1)\nu_{2^m}^2 \log\log 2^m \quad \text{a.s..} \tag{32}$$

Hence, $\sup_{\beta\in A_{2^m}} |\sum_{k=1}^{n} Z_k(\beta)| = o(1)\nu_{2^m}^2 \log\log 2^m = o(1)\nu_n^2 \log\log n$ a.s.. Accordingly $\sup_{\beta\in A_n} |\sum_{k=1}^{n} Z_k(\beta)| = o(1)\nu_n^2 \log\log n$ a.s. and $\sup_{\beta\in \partial A_n} |\sum_{k=1}^{n} Z_k(\beta)| = o(1)\nu_n^2 \log\log n$ a.s., which proves that (17) is true.

Concerning the uniform bound of T_4, by Lemma 4 and the Cauchy-Schwarz inequality, we see that

$$\sup_{\beta\in\partial A_n} |T_4| \le O(\sqrt{n\log\log n}) \sup_{\beta\in\partial A_n} ||\beta-\beta_0|| = O(1)\nu_n \log\log n \quad \text{a.s..} \tag{33}$$

Concerning T_1, by Lemma 2 and (C.13),

$$\begin{aligned}\inf_{\beta\in\partial A_n} T_1 &\ge \inf_{\beta\in\partial A_n} c_2 \sum_{k=1}^{n} \min\left\{ \frac{m_k h'(\mathbf{x}_k^t\beta_0)^2}{\pi_{0k}(1-\pi_{0k})}, m_k\pi_{0k}(1-\pi_{0k}) \right\} [\mathbf{x}_k^t(\beta-\beta_0)]^2 \\ &= \inf_{\beta\in\partial A_n} c_2 (\beta-\beta_0)^t X_n^t \Lambda_n X_n (\beta-\beta_0) \\ &\ge \inf_{\beta\in\partial A_n} c_2 ||\beta-\beta_0||^2 \lambda_1\{X_n^t \Lambda X_n\} \ge c_2 d_5 \nu_n^2 \log\log n. \end{aligned} \tag{34}$$

From (17), (33) and (34) it follows that there exists a positive constant c_4 such that, when n is sufficiently large,

$$\inf_{\beta\in\partial A_n} H(\beta) = \inf_{\beta\in\partial A_n} (T_1+T_4-T_5) \ge c_4\nu_n^2 \log\log n \quad \text{a.s..} \tag{35}$$

Note that $H(\beta_0)=0$, and, from Lemma 6, $H(\beta)$ is strictly convex on $\beta\in B_n \supset A_n$ for almost surely every sample sequence $\{y_1,\ldots,y_n\}$ when n is sufficiently large. This and (35) imply that the minimizer $\hat\beta_n$ of $H(\beta)$ over B_n must be a stationary point of $H(\beta)$, i.e, a solution of the likelihood equation (3), and an interior point of A_n. By our assumptions, this minimizer is the MLE and $||\hat\beta_n-\beta_0|| \le \nu_n(n^{-1}\log\log n)^{1/2}$ a.s.. Since ν_n can be chosen to be divergent as slowly as possible, it follows that (16) holds.

Now we proceed to show that there exists a constant $c>0$ such that

$$\limsup_{n\to\infty} \frac{||\hat\beta_n-\beta_0||}{\sqrt{n^{-1}\log\log n}} = c \quad \text{a.s..} \tag{36}$$

Suppose (36) is not true. By (16), this implies that $||\hat\beta_n-\beta_0|| = o(\sqrt{n^{-1}\log\log n})$ a.s.. Now let $H(\beta)=H(\beta,n)$ and $T_i=T_i(\beta,n)$ $(i=1,\ldots,5)$ to indicate their dependence on n. Following the same line for proving (17) and (33) one can

show that $T_4(\hat{\beta}_n, n) = o(1)\log\log n$ a.s. and $T_5(\hat{\beta}_n, n) = o(1)\log\log n$ a.s.. Further, by Lemma 1, $T_1(\hat{\beta}_n, n) = \sum_{k=1}^n K(\mathbf{x}_k^t\hat{\beta}_n, \mathbf{x}_k^t\beta_0) = \sum_{k=1}^n D(\mathbf{x}_k^t\hat{\beta}_n; \mathbf{x}_k^t\beta_0, h(\mathbf{x}_k^t\beta_0))$ and accordingly

$$\begin{aligned}|T_1(\hat{\beta}_n, n)| &\le \sum_{k=1}^n c_1 m_k[\mathbf{x}_k^t(\hat{\beta}_n - \beta_0)]^2 \\ &= c_1(\hat{\beta}_n - \beta_0)^t X_n^t M_n X_n(\hat{\beta}_n - \beta_0) = o(1)\log\log n \quad \text{a.s.}\end{aligned}$$

by (C.11) and the assumption that $||\hat{\beta}_n - \beta_0|| = o(\sqrt{n^{-1}\log\log n})$ a.s.. Therefore,

$$|H(\hat{\beta}_n, n)| = |T_1(\hat{\beta}_n, n) + T_4(\hat{\beta}_n, n) - T_5(\hat{\beta}_n, n)| = o(1)\log\log n \quad \text{a.s.}. \tag{37}$$

On the other hand, from Lemma 4 there exists a sequence $n_i \uparrow \infty$ such that

$$\lim_{i\to\infty} \frac{\sum_{k=1}^{n_i} m_k h'(\mathbf{x}_k^t\beta_0)\pi_{0k}^{-1}(1-\pi_{0k})^{-1}(y_k - \pi_{0k})x_{k1}}{\{2I_{n_i}(\beta_0)(1,1)\log\log I_{n_i}(\beta_0)(1,1)\}^{\frac{1}{2}}} = 1 \quad \text{a.s.}. \tag{38}$$

Now define a $p \times 1$ vector $\tilde{\beta}_{n_i}$ with the first component being

$$\tilde{\beta}_{n_i}(1) = [d_1/(4d_0d_4c_1)]\sqrt{\frac{2\log\log I_{n_i}(\beta_0)(1,1)}{I_{n_i}(\beta_0)(1,1)}} + \beta_{01},$$

and $\tilde{\beta}_{n_i}(j) = \beta_{0j}$ $(j = 2, \ldots, p)$. By (38), (C.7) and (C.9), there exists an $i' > 0$ such that $T_4(\tilde{\beta}_{n_i}, n_i) \le -[d_1/(4d_0d_4c_1)]\log\log n_i$ a.s. for $i > i'$. Following the same line of proving (17), one can show that $|T_5(\tilde{\beta}_{n_i}, n_i)| \le [d_1/(28d_0d_4c_1)]\log\log n_i$ a.s. when $i > i''$ for some $i'' > 0$. Further, by Lemma 1, (C.7), (C.9) and (C.11),

$$\begin{aligned}0 \le T_1(\tilde{\beta}_{n_i}, n_i) &= \sum_{k=1}^{n_i} K(\mathbf{x}_k^t\tilde{\beta}_{n_i}, \mathbf{x}_k^t\beta_0) = \sum_{k=1}^{n_i} D(\mathbf{x}_k^t\tilde{\beta}_{n_i}; \mathbf{x}_k^t\beta_0, h(\mathbf{x}_k^t\beta_0)) \\ &\le \sum_{k=1}^{n_i} c_1 m_k[\mathbf{x}_k^t(\tilde{\beta}_{n_i} - \beta_0)]^2 = c_1(\tilde{\beta}_{n_i} - \beta_0)^t X_{n_i}^t M_{n_i} X_{n_i}(\tilde{\beta}_{n_i} - \beta_0) \\ &\le c_1\lambda_p(X_{n_i}^t M_{n_i} X_{n_i})||\tilde{\beta}_{n_i} - \beta_0||^2 \le c_1 d_4 n_i \frac{d_1^2}{8d_0^2d_4^2c_1^2}\frac{d_0}{d_1 n_i}\log\log I_{n_i}(\beta_0)(1,1) \\ &= \frac{d_1}{8d_0d_4c_1}\log\log I_{n_i}(\beta_0)(1,1) \le \frac{d_1}{7d_0d_4c_1}\log\log n_i,\end{aligned}$$

when i is sufficiently large. Combining the above results for T_4, T_5 and T_1, we have

$$H(\tilde{\beta}_{n_i}, n) = T_1(\tilde{\beta}_{n_i}, n) + T_4(\tilde{\beta}_{n_i}, n) - T_5(\tilde{\beta}_{n_i}, n) \le -\frac{d_1}{14d_0d_4c_1}\log\log n_i \quad \text{a.s.},$$

when i is sufficiently large. It follows that $H(\tilde{\beta}_{n_i}, n) < H(\hat{\beta}_n, n)$ a.s. by (37), which contradicts the fact that $\hat{\beta}_n$ is the MLE minimizing $H(\beta, n)$ over B_n. This suffices to prove (7).

Proof of Theorem 2. It is sufficient to prove (9) only for the full model. Namely, we only need to prove

$$0 \le \sum_{k=1}^{n} \{\rho(h(\mathbf{x}_k^t \beta_0); y_k, m_k) - \rho(h(\mathbf{x}_k^t \hat{\beta}_n); y_k, m_k)\} = O(\log\log n) \quad \text{a.s..} \tag{39}$$

The inequality part of (39) is obvious because $\hat{\beta}_n$ is the MLE. The equality part of (39) is equivalent to $H(\hat{\beta}_n, n) = O(\log\log n)$ a.s.. From the proof of Theorem 1 we know $H(\hat{\beta}_n, n) = T_1(\hat{\beta}_n, n) + T_4(\hat{\beta}_n, n) - T_5(\hat{\beta}_n, n)$. Following the same line for proving (37) and noting $||\hat{\beta}_n - \beta_0|| = O(\sqrt{n^{-1}\log\log n})$ a.s., one can show that $H(\hat{\beta}_n, n) = O(\log\log n)$ a.s.. Therefore, (39) holds.

Proof of Theorem 3. By Theorem 1 we have $||\hat{\beta}_n - \beta_0|| = O(\sqrt{n^{-1}\log\log n})$ a.s.. Thus $\hat{\beta}_n \in A_0$ a.s., where A_0 is defined in (C.14). Now we introduce a $p \times 1$ vector $\hat{\beta}_n^*(\alpha)$ which is obtained by augmenting $\hat{\beta}_n(\alpha)$ with $p - p_0$ zeros in such a way that the sub-vector of $\hat{\beta}_n^*(\alpha)$ indexed by α equals $\hat{\beta}_n(\alpha)$. Clearly, $\hat{\beta}_n^*(\alpha) \notin A_0$ for any incorrect model $\alpha \in \mathcal{A}_w$. Hence by (C.14) and the assumption $\ell(\hat{\beta}_n|\mathbf{Y}_n, X_n) = \sup_{\beta \in A_0} \ell(\beta|\mathbf{Y}_n, X_n)$ we have, for any $\alpha \in \mathcal{A}_w$,

$$\ell(\hat{\beta}_n|\mathbf{Y}_n, X_n) - \ell(\hat{\beta}_n^*(\alpha)|\mathbf{Y}_n, X_n) \ge d_6 n \quad \text{a.s. when } n \ge n_0.$$

By Theorem 2 we know $\ell(\hat{\beta}_n|\mathbf{Y}_n, X_n) - \ell(\beta_0|\mathbf{Y}_n, X_n) = O(\log\log n)$ a.s.. Therefore,

$$\liminf_{n\to\infty} n^{-1}\{\ell(\beta_0|\mathbf{Y}_n, X_n) - \ell(\hat{\beta}_n^*(\alpha)|\mathbf{Y}_n, X_n)\} > 0 \quad \text{a.s..}$$

Noting that $\ell(\hat{\beta}_n^*(\alpha)|\mathbf{Y}_n, X_n) = \ell(\hat{\beta}_n(\alpha)|\mathbf{Y}_n, X_{n\alpha})$, we see that (10) and hence (11) hold.

5. Discussion

In this paper we study a set of penalized likelihood based model selection criteria for generalized linear models with binary or proportional responses. We assume that all explanatory variables that affect the response variable Y are available for selection. In this situation, a binomial distribution is appropriate for modeling Y. In practice, some variables affecting Y may not be observed thus a binomial distribution modeling Y may not be valid. Consequently, one may introduce an over-dispersion parameter to the regression model and use a beta-binomial distribution to model Y to account for the effects of those lurking variables. It should not be very difficult to extend the asymptotic results derived in this paper to this situation.

When the focus is only on the four link functions listed in the paper, the results of Lemma 6 can be strengthened in that the negative Hessian matrix $-[c\partial^2\ell/(\partial\beta\partial\beta^t)]$ is positive definite for any β, and accordingly $H(\beta)$ is strictly convex at every β. This can be proved by following Wedderburn (1976) and some intricate calculus.

It is worth mentioning that executing a model selection criterion by a computationally feasible procedure is as important as finding a desirable model selection criterion, especially when there are large number of candidate models for selection. However, this is beyond the scope of this paper. We refer to Qian (1999) and Qian and Field (2002b) for a Markov chain Monte Carlo selection procedure which is both feasible and consistent.

Appendix

Proof of Lemma 1. It is easy to see that $D'_t(t;s,y) = m[(1-y)/(1-h(t)) - y/h(t)]h'(t) - m[(1-y)/(1-h(s)) - y/h(s)]h'(s)$, and $D''_t(t;s,y) = m(1-y)[h'(t)^2/(1-h(t))^2 + h''(t)/(1-h(t))] + my[h'(t)^2/h(t)^2 - h''(t)/h(t)]$. Condition (C.1) implies that $\lim_{t\to-\infty} h(t) = 0$, $\lim_{t\to\infty} h(t) = 1$. It also implies that $h(s)-h(t) = \int_t^s h'(x)dx$ and $h'(s)-h'(t) = \int_t^s h''(x)dx$. Thus, by the uniform continuity of $h'(t)$ and $h''(t)$, it follows that $\lim_{t\to\pm\infty} h'(t) = 0$ and $\lim_{t\to\pm\infty} h''(t) = 0$. Now it is easy to see that $\lim_{t\to-\infty}[h'(t)^2/(1-h(t))^2 + h''(t)/(1-h(t))] = 0$ and $\lim_{t\to+\infty}[h'(t)^2/h(t)^2 - h''(t)/h(t)] = 0$. By this and (C.3), if $c_1 = \max\{\sup_t |h'(t)^2/(1-h(t))^2 + h''(t)/(1-h(t))|, \sup_t |h'(t)^2/h(t)^2 - h''(t)/h(t)|\}$, one has $c_1 < \infty$. Therefore, $|D''_t(t;s,y)| \le c_1 m$ for any t, s and y. Now define $F_1(t) = D(t;s,y) - c_1 m(t-s)^2$ and $F_2(t) = D(t;s,y) + c_1 m(t-s)^2$. Clearly, $F_1(s) = F'_1(s) = F_2(s) = F'_2(s) = 0$ and $F''_1(t) = D''_t(t;s,y) - 2c_1 m < 0$, $F''_2(t) = D''_t(t;s,y) + 2c_1 m > 0$. Therefore $F_1(t) \le 0$ and $F_2(t) \ge 0$ and (R.1) follows. Statement (R.2) follows from the fact that $D(s;s,h(s)) = D'_t(s;s,h(s)) = 0$ and $D''_t(s;s,h(s)) = mh'(s)^2/(h(s)(1-h(s))) \ge 0$.

Proof of Lemma 2. First suppose that (C.1), (C.2) and (C.4) hold. It is easy to see that $K(t,s) = -mh(s)\log(h(t)/h(s)) - m(1-h(s))\log[(1-h(t))/(1-h(s))]$ and $K'_t(t,s) = m[(1-h(s))/(1-h(t)) - h(s)/h(t)]h'(t)$. Further,

$$
\begin{aligned}
K''_t(t,s) &= m\left[\frac{1-h(s)}{(1-h(t))^2} + \frac{h(s)}{h(t)^2}\right]h'(t)^2 + m\left[\frac{1-h(s)}{1-h(t)} - \frac{h(s)}{h(t)}\right]h''(t) \\
&= \frac{mh'(s)^2}{h(s)(1-h(s))}\left\{\left[\frac{h(s)(1-h(s))^2}{(1-h(t))^2} + \frac{(1-h(s))h(s)^2}{h(t)^2}\right]\frac{h'(t)^2}{h'(s)^2}\right. \\
&\quad \left. + \left[\frac{h(s)(1-h(s))^2}{1-h(t)} - \frac{h(s)^2(1-h(s))}{h(t)}\right]\frac{h''(t)}{h'(s)^2}\right\} \qquad (40) \\
&= m(1-h(s))\left[\frac{h'(t)^2}{(1-h(t))^2} + \frac{h''(t)}{1-h(t)}\right] + mh(s)\left[\frac{h'(t)^2}{h(t)^2} - \frac{h''(t)}{h(t)}\right]. \qquad (41)
\end{aligned}
$$

We now consider three cases according to the position of s in relation to t_0 in (C.2) and (C.4): (i) $|s| \le 1.1t_0$, (ii) $s > 1.1t_0$ and (iii) $s < -1.1t_0$. In case (i), $K_t''(t,s) = mh'(s)^2/(h(s)(1-h(s)))\{u(t,s)+v(t,s)\} = mh'(s)^2/(h(s)(1-h(s)))\{[u(t,s) - u(s,s)] + [v(t,s) - v(s,s)] + 1\}$ by (40), the definitions of $u(t,s)$ and $v(t,s)$, and the fact that $u(s,s) = 1$ and $v(s,s) = 0$. By (C.1), $h(t)$, $h'(t)$ and $h''(t)$ are uniformly continuous on $[-1.2t_0, 1.2t_0]$. This implies uniform continuity of $u(t,s)$ and $v(t,s)$ with respect to $t \in [-1.2t_0, 1.2t_0]$ when $s \in [-1.1t_0, 1.1t_0]$. Thus, there exists a positive constant Δ_1 such that $|u(t,s) - u(s,s)| < 1/4$ and $|v(t,s) - v(s,s)| < 1/4$ when $|t-s| < \Delta_1$ and $s \in [-1.1t_0, 1.1t_0]$. Therefore $K_t''(t,s) > (1/2)mh'(s)^2/(h(s)(1-h(s)))$ when $|t-s| < \Delta_1$ and $s \in [-1.1t_0, 1.1t_0]$. In case (ii) where $s > 1.1t_0$, there exists $\Delta_2 > 0$ such that $t > t_0$ when $|t - s| < \Delta_2$. Assuming this, by (C.2), the second term of (41) is non-negative, and by Condition (C.4) the first term of (41) is not smaller than $c_2' m(1-h(s))$, where $c_2' = \inf_{t>t_0}\{h'(t)^2/(1-h(t))^2 + h''(t)/(1-h(t))\}$. Therefore, $K_t''(t,s) \ge c_2' mh(s)(1-h(s))$ when $|t-s| < \Delta_2$ and $s > 1.1t_0$. Similarly, it can be shown that in case (iii) there exist $\Delta_3 > 0$ and $c_2'' > 0$ such that $K_t''(t,s) \ge c_2'' mh(s)(1-h(s))$ when $|t-s| < \Delta_3$ and $s < -1.1t_0$. Define $\Delta = \min\{\Delta_1, \Delta_2, \Delta_3\}$ and $c_2 = (1/2)\min\{1/2, c_2', c_2''\}$. From the three cases discussed above and the fact $K(s,s) = K_t'(s,s) = 0$, Lemma 2 follows under (C.1), (C.2) and (C.4).

Now suppose that (C.1) and (C.5) hold. If $|s| \le \max(s_0, 1.1t_0)$, Lemma 2 can be proved following the same lines as in case (i) above. If $|s| > \max(s_0, 1.1t_0)$, Lemma 2 is obvious when (C.5) holds.

Proof of Lemma 3. It is easy to find that $R_t'(t,s) = h'(t)/h(t) + h'(t)/(1-h(t)) - h'(s)/(h(s)(1-h(s))$ and $R_t''(t,s) = h'(t)^2/(1-h(t))^2 + h''(t)/(1-h(t)) - [h'(t)^2/h(t)^2 - h''(t)/h(t)]$. As shown in Lemma 1, (C.1) implies that $\lim_{t\to-\infty} h(t) = 0$, $\lim_{t\to+\infty} h(t) = 1$, $\lim_{t\to\pm\infty} h'(t) = 0$, and $\lim_{t\to\pm\infty} h''(t) = 0$. These results ensure that $\sup_{t\le t_0} |h'(t)^2/(1-h(t))^2 + h''(t)/(1-h(t))| < \infty$ and $\sup_{t\ge -t_0} |h'(t)^2/h(t)^2 - h''(t)/h(t)| < \infty$. This and (C.3) imply that $|R_t''(t,s)| \le c_3$ and, accordingly, $|R(t,s)| \le c_3(t-s)^2$ for some constant c_3 not depending on t and s.

Now by (C.1) and the Mean Value Theorem,

$$\begin{aligned} R(t_1,s) - R(t_2,s) &= \log\frac{h(t_1)}{1-h(t_1)} - \log\frac{h(t_2)}{1-h(t_2)} - \frac{h'(s)}{h(s)(1-h(s))}(t_1-t_2) \\ &= \left[\frac{h'(t^*)}{h(t^*)(1-h(t^*))} - \frac{h'(s)}{h(s)(1-h(s))}\right](t_1-t_2) \\ &= [R_t'(t^*,s) - R_t'(s,s)](t_1-t_2) = R_t''(t^{**},s)(t^*-s)(t_1-t_2), \end{aligned}$$

where t^* is some value in between t_1 and t_2, and t^{**} is in between t^* and s. It is easy to see that $|R(t_1,s) - R(t_2,s)| \le c_3(|t_1-s| + |t_2-s|)|t_1-t_2|$.

Proof of Lemma 4. The proof of (14) is an application of Lemma 5. Knowing that $m_i y_i$ follows a $\mathrm{Bin}(m_i, \pi_{0i})$ distribution and writing $Z_{ij} = m_i h'(\mathbf{x}_i^t\beta_0)\pi_{0i}^{-1}(1-\pi_{0i})^{-1}(y_i-\pi_{0i})x_{ij}$, it can be verified that $EZ_{ij} = 0$, $EZ_{ij}^2 = [m_i h'(\mathbf{x}_i^t\beta_0)^2/(\pi_{0i}(1-\pi_{0i}))]x_{ij}^2$, and $\sum_{i=1}^n EZ_{ij}^2 = I_n(\beta_0)(j,j) \to \infty$ by (C.7). Further, by (C.7) and (C.8) and the inequality $\lambda_1\{I_n(\beta_0)\} \le I_n(\beta_0)(j,j) \le \lambda_p\{I_n(\beta_0)\}$, it can be shown that

$$\begin{aligned} Z_{nj}^2 &\le \frac{m_n^2 h'(\mathbf{x}_n^t\beta_0)^2}{\pi_{0n}^2(1-\pi_{0n})^2}x_{nj}^2 \le \lambda_p\{I_n(\beta_0)\}\frac{m_n^2 h'(\mathbf{x}_n^t\beta_0)^2}{\pi_{0n}^2(1-\pi_{0n})^2}\mathbf{x}_n^t\mathbf{x}_n\lambda_p\{I_n(\beta_0)\}^{-1} \\ &\le \lambda_p\{I_n(\beta_0)\}\frac{m_n^2 h'(\mathbf{x}_n^t\beta_0)^2}{\pi_{0n}^2(1-\pi_{0n})^2}\mathbf{x}_n^t I_n(\beta_0)^{-1}\mathbf{x}_n \le \lambda_p\{I_n(\beta_0)\}\delta_n^2 \\ &\le \frac{\lambda_p\{I_n(\beta_0)\}}{\lambda_1\{I_n(\beta_0)\}}\frac{I_n(\beta_0)(j,j)}{\log\log I_n(\beta_0)(j,j)}\delta_n^2 \log\log\lambda_p\{I_n(\beta_0)\} \\ &= o\left(\frac{I_n(\beta_0)(j,j)}{\log\log I_n(\beta_0)(j,j)}\right). \end{aligned}$$

Therefore, $\{Z_{ij}\}$ satisfies all conditions in Lemma 5 and (14) follows. Then result (15) follows from (14) and (C.9).

Proof of Lemma 6. By the definitions of $H(\beta)$ and the Fisher information $I_n(\beta)$, it is easy to see that

$$\frac{\partial^2 H(\beta)}{\partial\beta\partial\beta^t} = -\frac{\partial^2\ell}{\partial\beta\partial\beta^t} = T_6 + I_n(\beta) = T_6 + (I_n(\beta) - I_n(\beta_0)) + I_n(\beta_0), \tag{42}$$

where

$$T_6 = \sum_{k=1}^n m_k\left[\frac{(1-2\pi_k)h'(\mathbf{x}_k^t\beta)^2}{\pi_k^2(1-\pi_k)^2} - \frac{h''(\mathbf{x}_k^t\beta)}{\pi_k(1-\pi_k)}\right](y_k-\pi_k)\mathbf{x}_k\mathbf{x}_k^t.$$

Let $\chi(s) = (1-2h(s))h'(s)^2/(h(s)^2(1-h(s))^2) - h''(s)/(h(s)(1-h(s)))$ and rewrite T_6 as

$$\begin{aligned} T_6 &= \sum_{k=1}^n m_k\chi(\mathbf{x}_k^t\beta)(y_k-\pi_{0k})\mathbf{x}_k\mathbf{x}_k^t - \sum_{k=1}^n m_k\chi(\mathbf{x}_k^t\beta)[h(\mathbf{x}_k^t\beta) - h(\mathbf{x}_k^t\beta_0)]\mathbf{x}_k\mathbf{x}_k^t \\ &\overset{\text{denoted}}{=} T_7 - T_8. \end{aligned} \tag{43}$$

Note that

$$\chi(s) = \left\{-\frac{h'(s)^2}{(1-h(s))^2} - \frac{h''(s)}{(1-h(s))}\right\} + \frac{h'(s)^2}{h(s)^2} - \frac{h''(s)}{h(s)}.$$

In the proof of Lemma 1 we have seen that $\lim_{s\to\pm\infty} h'(s) = 0$ implies $|h'(s)| \le c_5$ for some constant c_5, and that $|\chi(s)| \le 2c_1$ under (C.1) and (C.3). From this,

(21), and the definition of $\{\nu_n\}$, it follows that for any $k \le n$,

$$\begin{aligned}|\chi(\mathbf{x}_k^t\beta)(h(\mathbf{x}_k^t\beta)-h(\mathbf{x}_k^t\beta_0))|I(\beta\in B_n) &\le c|\mathbf{x}_k^t(\beta-\beta_0)|I(\beta\in B_n)\\ &\le c||\mathbf{x}_k||||\beta-\beta_0||I(\beta\in B_n) \le cm_k^{-1/2}n^{1/2}\xi_n||\beta-\beta_0||I(\beta\in B_n)\\ &\le cn^{1/2}\xi_n\nu_n^2(n^{-1}\log\log n)^{1/2} = c\xi_n\nu_n^2(\log\log n)^{1/2} = o(1). \qquad (44)\end{aligned}$$

By (44) and (C.11), it is easy to see that $|T_8|I(\beta\in B_n) = o(n)$.

Now let

$$T_7(i,j) = \sum_{k=1}^n m_k\chi(\mathbf{x}_k^t\beta)(y_k-\pi_{0k})x_{ki}x_{kj} \overset{\text{denoted}}{=} \sum_{k=1}^n W_{kij}, \quad i,j=1,\ldots,p.$$

It is easy to see that $EW_{kij} = 0$ and $EW_{kij}^2 = m_k\pi_{0k}(1-\pi_{0k})\chi(\mathbf{x}_k^t\beta)^2(x_{ki}x_{kj})^2 \le cm_k(x_{ki}x_{kj})^2$. Thus according to Condition (C.12) $D_{nij} = \sum_{k=1}^n EW_{kij}^2 \le O(n)$. If $\lim_{n\to\infty} D_{nij} < \infty$, we have $T_7(i,j) = o(a_n)$ a.s. for any sequence $a_n \uparrow \infty$, according to a strong law of large numbers given by Theorem 6.6 of Petrov(1995, p.209), implying $T_7(i,j) = O(1)$ a.s.. If $\lim_{n\to\infty} D_{nij} = +\infty$, by Theorem 6.17 of Petrov(1995, p.222), we have $T_7(i,j) = o(D_{nij}^{2/3}) = o(n^{2/3})$ a.s..

The preceding results about T_7 and T_8 show that $T_6 I(\beta \in B_n) = o(n)$ a.s.. Knowing that $I_n(\beta) = \sum_{i=1}^n [m_i h'(\mathbf{x}_i^t\beta)^2/\pi_i(1-\pi_i)]\mathbf{x}_i\mathbf{x}_i^t$, one can show that $|I_n(\beta) - I_n(\beta_0)| = o(n)$ for $\beta \in B_n$. This is due to (C.11) and the following facts. First, $h'(s)^2/h(s)(1-h(s))$ has bounded first order derivative by (C.1), (C.3) and (C.6). Second, $\max_{1\le k\le n}|\mathbf{x}_k^t\beta - \mathbf{x}_k^t\beta_0|I(\beta\in B_n) = o(1)$ by (44).

From the results for T_6 and $I_n(\beta) - I_n(\beta_0)$, (C.7), (C.9), and (42), it follows that $\partial^2 H(\beta)/\partial\beta\partial\beta^t$ is positive definite and of order $O(n)$ on $\beta \in B_n$, and for almost surely all sample sequences $\{y_1,\ldots,y_n\}$ when n is sufficiently large. The lemma is proved.

Testing Conditions (C.1) to (C.6) for the logistic link. Here $g_1(\pi) = \log(\pi/(1-\pi))$ and the inverse link is $h_1(t) = e^t(1+e^t)^{-1}$. Obviously, $\lim_{t\to-\infty} h_1(t) = 0$, $\lim_{t\to+\infty} h_1(t) = 1$, $h_1'(t) = e^t(1+e^t)^{-2}$, $h_1''(t) = e^t(1-e^t)(1+e^t)^{-3}$, and $h_1'''(t) = (e^t - 4e^{2t} + e^{3t})(1+e^t)^{-4}$. It is easy to see that $|h_1''(t)| \le 1$ and $|h_1'''(t)| \le 6$. Hence (C.1) is satisfied. It is also easy to see that (C.2) is satisfied for any $t_0 > 0$. Now it can be shown that

$$\frac{h_1'(t)^2}{(1-h_1(t))^2} + \frac{h_1''(t)}{1-h_1(t)} = \frac{h_1'(t)^2}{h_1(t)^2} - \frac{h_1''(t)}{h_1(t)} = \frac{e^t}{(1+e^t)^2} = h_1'(t). \qquad (45)$$

From this, $\sup_{t>t_0>0}|h_1'(t)^2/(1-h_1(t))^2 + h_1''(t)/(1-h_1(t))| = \sup_{t>t_0>0}|h_1'(t)^2/h_1(t)^2 - h_1''(t)/h_1(t)| = \sup_{t>t_0>0} h_1'(t) = h_1'(t_0) < \infty$. So (C.3) is satisfied. But

(C.4) is not satisfied because $\inf_{t>t_0>0} h_1'(t) = \inf_{t<t_0<0} h_1'(t) = 0$. Concerning (C.5), it can be verified that

$$u(t,s) + v(t,s) = e^{t-s}\left(\frac{1+e^s}{1+e^t}\right)^2 \geq \begin{cases} e^{s-t}, & \text{if } t \geq s, \\ e^{t-s}, & \text{if } t < s. \end{cases} \tag{46}$$

So $u(t,s) + v(t,s) \geq e^{-\Delta_0}$ if $|t-s| \leq \Delta_0$. Therefore (C.5) is satisfied. Finally, since $h_1'(s)h_1''(s)/h_1(s)(1-h_1(s)) = h_1''(s)$, (C.6) is satisfied.

Testing Conditions (C.1) to (C.6) for the probit link. Here $g_2(\pi) = \Phi^{-1}(\pi)$ and the inverse probit link is $h_2(t) = \Phi(t) = \int_{-\infty}^{t}(2\pi)^{-1/2}e^{-s^2/2}ds$. It is easy to see that $h_2'(t) = (2\pi)^{-1/2}e^{-t^2 2}$, $h_2''(t) = -(t/\sqrt{2\pi})e^{-t^2 2}$ and $h_2'''(t) = ((t^2-1)/\sqrt{2\pi})e^{-t^2/2}$. Since $|h_2''(t)| \leq (2\pi)^{-1/2}e^{-1/2}$ and $-(2\pi)^{-1/2} \leq h_2'''(t) \leq (2/\sqrt{2\pi})e^{-3/2}$, it follows that both $h_2'(t)$ and $h_2''(t)$ are uniformly continuous, so (C.1) is satisfied. Condition (C.2) is clearly satisfied for any $t_0 > 0$. By repeatedly applying l'Hospital's rule, it can be shown that

$$\lim_{t\to+\infty} \frac{h_2'(t)^2}{(1-h_2(t))^2} + \frac{h_2''(t)}{1-h_2(t)} = \lim_{t\to+\infty} \frac{e^{-t^2} - te^{-\frac{1}{2}t^2}\int_t^{+\infty} e^{-\frac{1}{2}s^2}ds}{[\int_t^{+\infty} e^{-\frac{1}{2}s^2}ds]^2} = 1,$$

$$\lim_{t\to-\infty} \frac{h_2'(t)^2}{h_2(t)^2} - \frac{h_2''(t)}{h_2(t)} = \lim_{t\to-\infty} \frac{e^{-t^2} + te^{-\frac{1}{2}t^2}\int_{-\infty}^{t} e^{-\frac{1}{2}s^2}ds}{[\int_{-\infty}^{t} e^{-\frac{1}{2}s^2}ds]^2} = 1.$$

This suggests that (C.3) and (C.4) hold if t_0 is taken to be sufficiently large. By applying l'Hospital's rule, one can show that

$$\lim_{t\to\pm\infty} \frac{h_2'(t)h_2''(t)}{h_2(t)(1-h_2(t))} = \lim_{t\to\pm\infty} \frac{te^{-t^2}}{\int_{-\infty}^{t} e^{-\frac{1}{2}s^2}ds \int_t^{+\infty} e^{-\frac{1}{2}s^2}ds} = 0.$$

Hence (C.6) holds. Now we proceed to prove that (C.5) does not hold for $h_2(t)$. Let $t = s + \Delta$. By repeatedly applying l'Hospital's rule, one can show that

$$\lim_{s\to+\infty} \frac{1-h_2(s)}{1-h_2(s+\Delta)} \cdot \frac{h_2'(s+\Delta)}{h_2'(s)} = 1 \quad \text{for any } \Delta, \tag{47}$$

$$\lim_{s\to+\infty} (1-h_2(s)) \cdot \frac{h_2'(s+\Delta)^2}{h_2'(s)^2} = 0 \quad \text{for any } \Delta, \tag{48}$$

$$\lim_{s\to+\infty} \frac{(1-h_2(s))^2}{1-h_2(s+\Delta)} \cdot \frac{h_2''(s+\Delta)}{h_2'(s)^2} = -1 \quad \text{for any } \Delta, \tag{49}$$

$$\lim_{s\to+\infty} (1-h_2(s)) \cdot \frac{h_2''(s+\Delta)}{h_2'(s)^2} = \begin{cases} -\infty, & \text{if } \Delta < 0 \\ -1, & \text{if } \Delta = 0 \\ 0, & \text{if } \Delta > 0. \end{cases} \tag{50}$$

From (47) and (48) we see $\lim_{s\to+\infty} u(s+\Delta, s) = 1$ for any Δ. From (49) and (50) we have

$$\lim_{s\to+\infty} v(s+\Delta, s) = \begin{cases} +\infty, & \text{if } \Delta < 0 \\ 0, & \text{if } \Delta = 0 \\ -1, & \text{if } \Delta > 0. \end{cases}$$

Therefore, $\lim_{s\to+\infty} u(s+\Delta, s) + v(s+\Delta, s) = 0$ if $\Delta > 0$. Hence (C.5) does not hold.

Testing Conditions (C.1) to (C.6) for the complementary log-log link. Here $g_3(\pi) = \log\{-\log(1-\pi)\}$ and the inverse complementary log-log link is $h_3(t) = 1 - e^{-e^t}$. Also $h_3'(t) = e^{t-e^t}$, $h_3''(t) = (1-e^t)e^{t-e^t}$ and $h_3'''(t) = (1-3e^t+e^{2t})e^{t-e^t}$. It is easy to see that $\lim_{t\to\pm\infty} h_3'(t) = \lim_{t\to\pm\infty} h_3''(t) = \lim_{t\to\pm\infty} h_3'''(t) = 0$, implying that (C.1) is satisfied. Condition (C.2) apparently holds for any $t_0 > 0$. Applying l'Hospital's rule, one can show that

$$\lim_{t\to\pm\infty} \frac{h_3'(t)h_3''(t)}{h_3(t)(1-h_3(t))} = \lim_{t\to\pm\infty} \frac{(1-e^t)e^{2t-e^t}}{1-e^{-e^t}} = 0,$$

which implies (C.6). Note that

$$\frac{h_3'(t)^2}{(1-h_3(t))^2} + \frac{h_3''(t)}{1-h_3(t)} = e^t \to +\infty \quad \text{as } t \to +\infty \tag{51}$$

so (C.3) does not hold. By applying l'Hospital's rule,

$$\lim_{t\to-\infty} \left\{ \frac{h_3'(t)^2}{h_3(t)^2} - \frac{h_3''(t)}{h_3(t)} \right\} = \lim_{t\to-\infty} \left\{ \frac{e^{2t-2e^t}}{(1-e^{-e^t})^2} - \frac{(1-e^t)e^{t-e^t}}{1-e^{-e^t}} \right\} = 0.$$

This suggests (C.4) does not hold. But this and (51) suggest that (C.3) would be satisfied if one considers only those t values bounded from above by a finite value. To see whether (C.5) is satisfied, write $t = s + \Delta$. It can be shown that

$$u(s+\Delta, s) = (1-e^{-e^s})e^{2\Delta} + \frac{(1-e^{-e^s})^2}{(1-e^{-e^{s+\Delta}})^2} e^{2\Delta+(1-2e^\Delta)e^s}, \tag{52}$$

$$v(s+\Delta, s) = (1-e^{-e^s})(e^{\Delta-s} - e^{2\Delta}) - \frac{(1-e^{-e^s})^2}{1-e^{-e^{s+\Delta}}} e^{e^s(1-e^\Delta)}(e^{\Delta-s} - e^{2\Delta}) \tag{53}$$

$$= (e^\Delta - e^{2\Delta+s}) \frac{(1-e^{-e^s})(1-e^{e^s(1-e^\Delta)})}{e^s(1-e^{-e^{s+\Delta}})}. \tag{54}$$

By applying l'Hospital's rule, one can show from (52) that

$$\lim_{s\to-\infty} u(s+\Delta, s) = 1 \quad \text{and} \quad \lim_{s\to+\infty} u(s+\Delta, s) = \begin{cases} +\infty, & \text{if } \Delta < -\log 2 \\ \frac{1}{2}, & \text{if } \Delta = -\log 2 \\ e^{2\Delta}, & \text{if } \Delta > -\log 2. \end{cases} \tag{55}$$

Similarly, from (53) and (54), respectively, we have

$$\lim_{s\to-\infty} v(s+\Delta,s) = e^{\Delta}-1 \quad \text{and} \quad \lim_{s\to+\infty} v(s+\Delta,s) = \begin{cases} +\infty & \text{if } \Delta<0 \\ 0, & \text{if } \Delta=0 \\ -e^{2\Delta}, & \text{if } \Delta>0. \end{cases} \tag{56}$$

From(55) and (56) we have $\lim_{s\to-\infty}\{u(s+\Delta,s)+v(s+\Delta,s)\} = e^{\Delta}$ and

$$\lim_{s\to+\infty}\{u(s+\Delta,s)+v(s+\Delta,s)\} = \begin{cases} +\infty, & \text{if } \Delta<0 \\ 1, & \text{if } \Delta=0 \\ 0, & \text{if } \Delta>0. \end{cases} \tag{57}$$

From (57) it is easy to see that (C.5) does not hold for any $\Delta_0>0$. On the other hand, by using the inequality $e^u-1-u\ge 0$ one can show that

$$\begin{aligned} u(s+\Delta,s)+v(s+\Delta,s) &= (1-e^{-e^s})e^{-s}e^{\Delta} \\ &+\frac{(1-e^{-e^s})^2}{(1-e^{-e^{s+\Delta}})^2}e^{\Delta-s+(1-e^{\Delta})e^s}[e^{-e^{s+\Delta}}+e^{s+\Delta}-1] \ge (1-e^{-e^s})e^{-s}e^{\Delta}. \end{aligned}$$

Using $e^u-1-u\ge 0$ again, one can see that $(1-e^{-e^s})e^{-s}$ is a decreasing function. Thus, there exists an $s_0>0$ such that $u(s+\Delta,s)+v(s+\Delta,s)\ge (1-e^{-e^{-s_0}})e^{s_0+\Delta}$ when $s<-s_0$. Therefore, $\inf_{|t-s|\ge\Delta_0, s<-s_0}\{u(t,s)+v(t,s)\}\ge (1-e^{-e^{-s_0}})e^{s_0-\Delta_0}>0$, suggesting that (C.5) would hold if we focus on $h_3(t)<1-\delta'$ for certain δ' only.

Testing Conditions (C.1) to (C.6) for the log-log link. Here the link is $g_4(\pi) = -\log\{-\log\pi\}$ and its inverse is $h_4(t)=e^{-e^{-t}}$. Therefore, $h_4'(t) = e^{-t-e^{-t}}$, $h_4''(t)=(e^{-t}-1)e^{-t-e^{-t}}$ and $h_4'''(t)=(1-3e^{-t}+e^{-2t})e^{-t-e^{-t}}$. It is easy to see that $\lim_{t\to\pm\infty}h_4'(t)=\lim_{t\to\pm\infty}h_4''(t)=\lim_{t\to\pm\infty}h_4'''(t)=0$, implying that (C.1) is satisfied. Condition (C.2) apparently holds for any $t_0>0$. Applying l'Hospital's rule, one can show that

$$\lim_{t\to\pm\infty}\frac{h_4'(t)h_4''(t)}{h_4(t)(1-h_4(t))} = \lim_{t\to\pm\infty}\frac{(e^{-t}-1)e^{-2t-e^{-t}}}{1-e^{-e^{-t}}} = 0,$$

which implies (C.6). Note that

$$\frac{h_4'(t)^2}{h_4(t)^2}-\frac{h_4''(t)}{h_4(t)} = e^{-t}\to+\infty \quad \text{as } t\to-\infty, \tag{58}$$

so (C.3) does not hold. By applying l'Hospital's rule,

$$\lim_{t\to+\infty}\left\{\frac{h_4'(t)^2}{(1-h_4(t))^2}+\frac{h_4''(t)}{1-h_4(t)}\right\} = \lim_{t\to+\infty}\left\{\frac{e^{-2t-2e^{-t}}}{(1-e^{-e^{-t}})^2}+\frac{(e^{-t}-1)e^{-t-e^{-t}}}{1-e^{-e^{-t}}}\right\} = 0.$$

This suggests (C.4) does not hold. But this and (58) suggest that (C.3) would be satisfied if one considers only those t values bounded from below by a finite value. To see whether (C.5) is satisfied, write $t = s + \Delta$. It can be shown that

$$u(s+\Delta,s)=(1-e^{-e^{-s}})e^{-2\Delta}+\frac{(1-e^{-e^{-s}})^2}{(1-e^{-e^{-s-\Delta}})^2}e^{-2\Delta+(1-2e^{-\Delta})e^{-s}}, \tag{59}$$

$$v(s+\Delta,s)=(1-e^{-e^{-s}})(e^{s-\Delta}-e^{-2\Delta})-\frac{(1-e^{-e^{-s}})^2}{1-e^{-e^{-s-\Delta}}}e^{e^{-s}(1-e^{-\Delta})}(e^{s-\Delta}-e^{-2\Delta}) \tag{60}$$

$$=(e^{-\Delta}-e^{-2\Delta-s})\frac{(1-e^{-e^{-s}})(1-e^{e^{-s}(1-e^{-\Delta})})}{e^{-s}(1-e^{-e^{-s-\Delta}})}. \tag{61}$$

By applying l'Hospital's rule, one can show from (59) that

$$\lim_{s\to+\infty}u(s+\Delta,s)=1 \quad \text{and} \quad \lim_{s\to-\infty}u(s+\Delta,s)=\begin{cases}e^{-2\Delta}, & \text{if } \Delta<\log 2\\ \frac{1}{2}, & \text{if } \Delta=\log 2\\ +\infty, & \text{if } \Delta>\log 2.\end{cases} \tag{62}$$

Similarly, from (60) and (61), respectively, we have

$$\lim_{s\to+\infty}v(s+\Delta,s)=e^{-\Delta}-1 \text{ and } \lim_{s\to-\infty}v(s+\Delta,s)=\begin{cases}-e^{-2\Delta}, & \text{if } \Delta<0\\ 0, & \text{if } \Delta=0\\ +\infty, & \text{if } \Delta>0.\end{cases} \tag{63}$$

From(62) and (63) we have $\lim_{s\to+\infty}\{u(s+\Delta,s)+v(s+\Delta,s)\}=e^{-\Delta}$ and

$$\lim_{s\to-\infty}\{u(s+\Delta,s)+v(s+\Delta,s)\}=\begin{cases}0, & \text{if } \Delta<0\\ 1, & \text{if } \Delta=0\\ +\infty, & \text{if } \Delta>0.\end{cases} \tag{64}$$

From (64) it is easy to see that (C.5) does not hold for any $\Delta_0 > 0$. On the other hand, by using the inequality $e^u - 1 - u \geq 0$ one can show that

$$\begin{aligned}&u(s+\Delta,s)+v(s+\Delta,s)\\ &=(1-e^{-e^{-s}})e^{s}e^{-\Delta}+\frac{(1-e^{-e^{-s}})^2}{(1-e^{-e^{-s-\Delta}})^2}e^{s-\Delta+(1-e^{-\Delta})e^{-s}}[e^{-e^{-s-\Delta}}+e^{-s-\Delta}-1]\\ &\geq(1-e^{-e^{-s}})e^{s}e^{-\Delta}.\end{aligned}$$

Using $e^u-1-u\geq 0$ again, one can see that $(1-e^{-e^{-s}})e^s$ is an increasing function. Thus, there exists an $s_0 > 0$ such that $u(s+\Delta,s)+v(s+\Delta,s)\geq(1-e^{-e^{-s_0}})e^{s_0-\Delta}$ when $s > s_0$. Therefore, $\inf_{|t-s|\geq\Delta_0,s>s_0}\{u(t,s)+v(t,s)\}\geq(1-e^{-e^{-s_0}})e^{s_0-\Delta_0}>0$, suggesting that (C.5) would hold if we focus on $h_4(t) > \delta''$ for certain δ'' only.

Example. A link function yielding local MLEs. Let $h_0(t) = \varpi^{-1}\arctan t + 0.5 + qt^{-2}\sin^2 t$ with $\varpi = 3.14159\cdots$, $q = 0.1$, and $-\infty < t < +\infty$. Our link

function is defined as the inverse of $h_0(t)$. The function $h_0(t)$ and its first three derivatives are plotted in Figure 2. Later we show that $h_0(t)$ defines a strictly increasing cumulative distribution function and satisfies (C.1), (C.3), (C.5) and (C.6), but not (C.2) and (C.4). Therefore, the main results of this paper still apply for $h_0(t)$.

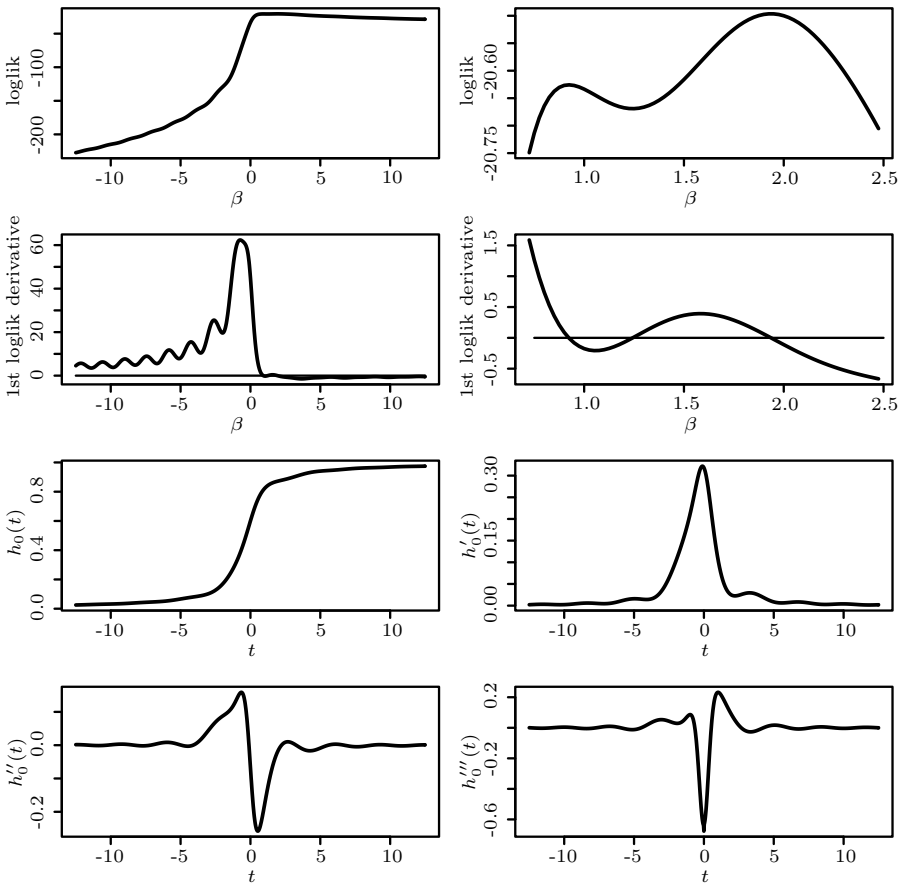

Figure 2. In row 1 the left plot is for $\ell(\beta)$ vs. β while its enlargement near the local maxima is given by the right plot. In row 2 the left plot is for $\ell'(\beta)$ vs. β while its enlargement near the stationary points is given by the right plot. The other four plots are for the inverse link function $h_0(t)$ and its first three derivatives.

Now suppose the response variable $Y = Z/m$ is related to a covariate x through $h_0^{-1}(\pi) = \beta x$, and that we have two observations $(z_1, m_1, x_1) = (6, 10, 1)$ and $(z_2, m_2, x_2) = (47, 50, 2)$. Then the log-likelihood function, ignoring an irrel-

evant constant, is $\ell(\beta) = 6\log h_0(\beta) + 4\log(1-h_0(\beta)) + 47\log h_0(2\beta) + 3\log(1-h_0(2\beta))$; the likelihood equation is $d\ell/d\beta = (10h_0'(\beta)/h_0(\beta)(1-h_0(\beta)))(0.6-h_0(\beta)) + (100h_0'(2\beta)/h_0(2\beta)(1-h_0(2\beta)))(0.94-h_0(2\beta)) = 0$. By plotting $d\ell/d\beta$ and $\ell(\beta)$ against β (see Figure 2) and applying the Newton-Raphson algorithm, we find there are two local maximizers of $\ell(\beta)$, namely $\hat{\beta}_1 = 0.923$ and $\hat{\beta}_2 = 1.936$, which are the local MLEs of β; there is one local minimizer $\hat{\beta}_3 = 1.243$. The log-likelihoods at these β points are -20.63, -20.49 and -20.67 respectively.

In order to test (C.1) to (C.6) for $h_0(t)$, we need the following inequalities based on Taylor expansions for $\sin t$ and $\cos t$.

1. When $t > 0$, (1a) $\sin t \geq -1$, (1b) $\sin t < t$, (1c) $\sin t > t - t^3/6$, (1d) $\sin t < t - t^3/6 + t^5/120$ and (1e) $\sin t > t - t^3/6 + t^5/120 - t^7/5040$.
2. When $t < 0$, (2a) $\sin t \geq -1$, (2b) $\sin t > t$, (2c) $\sin t < t - t^3/6$, (2d) $\sin t > t - t^3/6 + t^5/120$ and (2e) $\sin t < t - t^3/6 + t^5/120 - t^7/5040$.
3. For any $t \neq 0$, (3a) $|\cos t| \leq 1$, (3b) $\cos t > 1 - t^2/2$, (3c) $\cos t < 1 - t^2/2 + t^4/24$ and (3d) $\cos t > 1 - t^2/2 + t^4/24 - t^6/720$.

We see that $h_0'(t) = \varpi^{-1}(1+t^2)^{-1} + qt^{-2}\sin 2t + qt^{-3}(\cos 2t - 1)$. We proceed to prove $h_0'(t) > 0$. When $t < -\sqrt{2.5}$, using inequalities (2a) and (3a), $h_0'(t) \geq \varpi^{-1}(1+t^2)^{-1} - qt^{-2} \geq 0.44t^{-2}(1+t^2)^{-1} > 0$. When $-\sqrt{2.5} \leq t \leq 0$, using (2d) and (3c), $h_0'(t) \geq \varpi^{-1}(1+t^2)^{-1} - (2/3)qt(1-(2/5)t^2) > 0$. When $0 < t \leq 1.32$, using (1c) and (3d), $h_0'(t) \geq \varpi^{-1}(1+t^2)^{-1} - (2/3)qt - (4/45)qt^3 > 0.0076$. When $1.32 < t \leq 1.62$, using (1e) and (3d), $h_0'(t) \geq \varpi^{-1}(1+t^2)^{-1} - (2/3)qt + (8/315)qt^3(7-t^2) > 0.0053$. When $t > 1.62$, using (1a) and (3a), $h_0'(t) \geq \varpi^{-1}(1+t^2)^{-1} - qt^{-2} - 2qt^{-3} \geq (1+t^2)^{-1}t^{-2}[(\varpi^{-1}-q)t^2 - 2qt - (1+2/1.62)q] > 0$, because $(\varpi^{-1}-q)t^2 - 2qt - (1+2/1.62)q$ has two roots at -0.65 and 1.57. Now it is easy to see that $h_0'(t)$ is positive.

The second and third derivatives of $h_0(t)$ are found to be $h_0''(t) = -2\varpi^{-1}t(1+t^2)^{-2} - qt^{-4}[3\cos 2t - 3 + 4t\sin 2t - 2t^2\cos 2t]$ and $h_0'''(t) = \varpi^{-1}(6t^2-2)(1+t^2)^{-3} + qt^{-5}[12\cos 2t - 12 + 18t\sin 2t - 12t^2\cos 2t - 4t^3\sin 2t]$. By repeatedly applying l'Hospital's rule, it can be shown that $\lim_{t\to-\infty} h_0(t) = 0$, $\lim_{t\to+\infty} h_0(t) = 1$ and $\lim_{t\to 0} h_0(t) = 1/2 + q$; $\lim_{t\to\pm\infty} h_0'(t) = \lim_{t\to\pm\infty} h_0''(t) = \lim_{t\to\pm\infty} h_0'''(t) = 0$, $\lim_{t\to 0} h_0'(t) = \varpi^{-1}$, $\lim_{t\to 0} h_0''(t) = -2/3q$ and $\lim_{t\to 0} h_0'''(t) = -2\varpi^{-1}$. Therefore, (C.1) is satisfied for $h_0(t)$.

Using l'Hospital's rule, one can also show that $\lim_{t\to+\infty} t^2(1/2 - \varpi^{-1}\arctan t) = +\infty$ and $\lim_{t\to-\infty} t^2(1/2 + \varpi^{-1}\arctan t) = +\infty$. Using these results, and those in the previous paragraph, one can show that

$$\lim_{t\to+\infty} \frac{h_0'(t)^2}{(1-h_0(t))^2} + \frac{h_0''(t)}{1-h_0(t)} = 0, \quad \lim_{t\to 0} \frac{h_0'(t)^2}{(1-h_0(t))^2} + \frac{h_0''(t)}{1-h_0(t)} = \frac{4\varpi^{-2}}{(1-2q)^2} - \frac{4q}{3-6q},$$

$$\lim_{t\to-\infty} \frac{h_0'(t)^2}{h_0(t)^2} - \frac{h_0''(t)}{h_0(t)} = 0 \qquad \text{and} \qquad \lim_{t\to 0} \frac{h_0'(t)^2}{h_0(t)^2} - \frac{h_0''(t)}{h_0(t)} = \frac{4\varpi^{-2}}{(1+2q)^2} + \frac{4q}{3+6q}.$$

Therefore (C.3) holds and (C.4) does not hold for $h_0(t)$. The condition (C.6) can also be deduced from the above results.

It is easy to see that $h_0''(t)$ is positive when $\cos 2t = 1$ and t is sufficiently large. Thus (C.2) does not hold for $h_0(t)$.

In order to verify (C.5) for $h_0(t)$, write $t = s + \Delta$ and tentatively assume $|\Delta| \le 0.2$. It is easy to show that, uniformly on $|\Delta| \le 0.2$,

$$\lim_{s\to\pm\infty} \frac{1-h_0(s)}{1-h_0(s+\Delta)} = \frac{s(1/2-\varpi^{-1}\arctan s) - qs^{-1}\sin^2 s}{s(\frac{1}{2}-\varpi^{-1}\arctan(s+\Delta)) - qs(s+\Delta)^{-2}\sin^2(s+\Delta)} = 1, \quad (65)$$

$$\lim_{s\to\pm\infty} \frac{h_0(s)}{h_0(s+\Delta)} = \frac{s(\frac{1}{2}+\varpi^{-1}\arctan s) + qs^{-1}\sin^2 s}{s(\frac{1}{2}+\varpi^{-1}\arctan(s+\Delta)) + qs(s+\Delta)^{-2}\sin^2(s+\Delta)} = 1, \quad (66)$$

because $\lim_{s\to+\infty} s(1/2-\varpi^{-1}\arctan s) = \varpi^{-1}$, $\lim_{s\to-\infty} s(\frac{1}{2}+\varpi^{-1}\arctan s) = \varpi^{-1}$ and $\lim_{s\to\pm\infty} \varpi^{-1}s[\arctan(s+\Delta)-\arctan s] = 0$ uniformly on $|\Delta| \le 0.2$. It can also be shown that $s^2[\arctan(s-0.2)-\arctan s] \le s^2[\arctan(s+\Delta)-\arctan s] \le s^2[\arctan(s+0.2)-\arctan s]$, $\lim_{s\to\pm\infty} s^2[\arctan(s\pm 0.2)-\arctan s] = \pm 0.2$ and $|\sin^2(s+\Delta) - \sin^2 s| \le |\Delta|$. Thus, uniformly on $|\Delta| \le 0.2$,

$$\begin{aligned}
&\limsup_{s\to\pm\infty} s^2|h_0(s+\Delta) - h_0(s)| \\
&\le \limsup_{s\to\pm\infty} \varpi^{-1}s^2|\arctan(s+\Delta)-\arctan s| \\
&\quad + \limsup_{s\to\pm\infty} qs^2(s+\Delta)^{-2}|\sin^2(s+\Delta)-\sin^2 s| + \limsup_{s\to\pm\infty} q|s^2(s+\Delta)^{-2}-1|\sin^2 s \\
&\le 0.2\varpi^{-1} + q|\Delta| \le 0.2(\varpi^{-1}+q). \qquad (67)
\end{aligned}$$

Now we can write $h_0'(s+\Delta)/h_0'(s) = 1 + (A+B+C)/D$, where

$$\begin{aligned}
A &= \frac{-\varpi^{-1}s^2(2s\Delta+\Delta^2)}{[1+(s+\Delta)^2](1+s^2)} \ge \frac{-0.4\varpi^{-1}|s|^3 - 0.04\varpi^{-1}s^2}{[1+(|s|-0.2)^2](1+s^2)}, \\
B &= \frac{qs^2[\sin 2(s+\Delta)-\sin 2s] - q[2s\Delta+\Delta^2]\sin 2s}{(s+\Delta)^2} \ge \frac{-q(0.4s^2+0.4|s|+0.04)}{(|s|-0.2)^2}, \\
C &= -2qs^2(s+\Delta)^{-3}\sin^2(s+\Delta) + 2qs^{-1}\sin^2 s \ge -2qs^2(s-0.2)^{-3} - 2q|s|^{-1}, \\
D &= \varpi^{-1}s^2(1+s^2)^{-1} + q\sin 2s - 2qs^{-1}\sin^2 s \quad \text{and} \quad \liminf_{|s|\to\infty} D = \varpi^{-1} - q.
\end{aligned}$$

From the above properties of A, B, C and D we have

$$\liminf_{|s|\to\infty} \frac{h_0'(s+\Delta)}{h_0'(s)} \ge 1 - 0.4q(\varpi^{-1}-q)^{-1} \text{ uniformly on } |\Delta| \le 0.2. \quad (68)$$

By (65),(66), (68), $\lim_{s\to+\infty} h_0(s) = 1$, and $\lim_{s\to-\infty} 1-h_0(s) = 1$, it follows that

$$\liminf_{|s|\to\infty} u(s+\Delta, s) \ge (1-0.4q(\varpi^{-1}-q)^{-1})^2 \doteq 0.667 \text{ uniformly on } |\Delta| \le 0.2. \quad (69)$$

Now consider $h_0''(s+\Delta)$. We have $|h_0''(s+\Delta)| \le 2\varpi^{-1}(|s|+0.2)(1+(|s|-0.2)^2)^{-2} + 2q(|s|-0.2)^{-2}[3(|s|-0.2)^{-2}+2(|s|-0.2)^{-1}+1]$. Thus $\limsup_{|s|\to\infty} s^2|h_0''(s+\Delta)| \le 2q$ uniformly on $|\Delta| \le 0.2$. Accordingly, since $s^2h_0'(s) = D$ and $\liminf_{|s|\to\infty} D = \varpi^{-1}-q$,

$$\limsup_{|s|\to\infty} \frac{s^2|h_0''(s+\Delta)|}{[s^2h_0'(s)]^2} \le \frac{2q}{(\varpi^{-1}-q)^2} \text{ uniformly on } |\Delta| \le 0.2. \tag{70}$$

From (65), (66), (67) and (70), it can be seen that, uniformly on $|\Delta| \le 0.2$,

$$\begin{aligned}
&\limsup_{|s|\to\infty} |v(s+\Delta, s)| \\
&= \limsup_{|s|\to\infty} \left| \frac{(1-h_0(s))h_0(s)}{(1-h_0(s+\Delta))h_0(s+\Delta)} \right| s^2|h_0(s+\Delta)-h_0(s)| \frac{s^2|h_0''(s+\Delta)|}{|s^2h_0'(s)|^2} \\
&\le \frac{0.4q(\varpi^{-1}+q)}{(\varpi^{-1}-q)^2} \doteq 0.351.
\end{aligned} \tag{71}$$

By (69) and (71), we have $\liminf_{|s|\to\infty}\{u(s+\Delta, s)+v(s+\Delta, s)\} \ge 0.316$ uniformly on $|\Delta| \le 0.2$, hence (C.5) holds.

Acknowledgement

We would like to thank an associate editor and the referee for comments and suggestions that improved the presentation of this paper. The first draft of this paper was prepared when he was at La Trobe University, Australia. His research is supported by a grant from the Australian Research Council. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Internat. Symp. on Information Theory* (Edited by B. N. Petrov and F. Csáki), 267-281. Akadémia Kiadó, Budapest.

Chow, Y. S. and Teicher, H. (1997). *Probability Theory*, 3rd Edition. Springer, New York.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13**, 342-368.

George, E. I. (2002). The variable selection problem. In *Statistics in the 21st Century* (Edited by A. E. Raftery, M. A. Tanner and M. T. Wells), 350-358. Chapman and Hall, London.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. 2nd Edition. Springer, New York.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd Edition, Chapman and Hall, London.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford University Press.

Qian, G. (1999). Computations and analysis in robust regression model selection using stochastic complexity. *Comput. Statist.* **14**, 293-314.

Qian, G. and Field, C. (2002a). Law of iterated logarithm and consistent model selection criterion in logistic regression. *SStatist. Probab. Lett.* **56**, 101-112.

Qian, G. and Field, C. (2002b). Using MCMC for logistic regression model selection involving large number of candidate models. In *Selected Proceedings of the* 4*th International Conference on Monte Carlo & Quasi-Monte Carlo Methods in Scientific Computing* (Edited by K. T. Fang, F. J. Hickernell and H. Niederreiter), 460-474. Springer, Hong Kong.

Qian, G. and Künsch, H. (1998). Some notes on Rissanen's stochastic complexity. *IEEE Trans. on Inform. Theory* **44**, 782-786.

Rao, C. R. and Wu, Y. (2001). On model selection (with discussion). In *Model Selection* (Edited by P. Lahiri), 1-64. IMS Lecture Notes – Monograph Series 38, Institute of Mathematical Statistics, Beachwood, Ohio.

Rao, C. R. and Zhao, L. C. (1992). Linear representation of M-estimates in linear models. *Canad. J. Statist.* **20**, 359-368.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Co., Singapore.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory* **42**, 40-47.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, **63**, 27-32.

Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia.

E-mail: g.qian@ms.unimelb.edu.au

Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada.

E-mail: wuyh@mathstat.yorku.ca

(Received February 2004; accepted August 2005)