Ann. Inst. Statist. Math.
Vol. 48, No. 2, 283–294 (1996)

THE QUASI-LIKELIHOOD ESTIMATION IN REGRESSION*

JONG-WUU WU

Department of Statistics, Tamkang University, Tamsui, Taipei, Taiwan 25137, R.O.C.

(Received March 10, 1994; revised August 17, 1995)

Abstract. I propose a simply method to estimate the regression parameters in quasi-likelihood model. My main approach utilizes the dimension reduction technique to first reduce the dimension of the regressor X to one dimension before solving the quasi-likelihood equations. In addition, the real advantage of using dimension reduction technique is that it provides a good initial estimate for one-step estimator of the regression parameters. Under certain design conditions, the estimators are asymptotically multivariate normal and consistent. Moreover, a Monte Carlo simulation is used to study the practical performance of the procedures, and I also assess the cost of CPU time for computing the estimates.

Key words and phrases: Regression parameter, quasi-likelihood model, link function, Monte Carlo simulation.

1. Introduction

In the general parametric inference approach, one frequently assumes the distribution of $Y \mid X = x$ to belong to a family of distributions depending on the parameters $\beta^T x$ and θ. Here β is a p-dimensional column vector of regression coefficients and θ is a q-dimensional column vector parameters. In the present paper, I consider the following model:

$$E(Y \mid X = x) = g(\beta^T x; \theta) \tag{1.1}$$

and

$$\mathrm{Var}(Y \mid X = x) = \phi V[g(\beta^T x; \theta)]. \tag{1.2}$$

Here I let $g(\cdot;\cdot)$ and $V(\cdot)$ be given functions and ϕ, β, and θ are unknown parameters to be estimated. If the true parameter θ is θ_0, $g(t;\theta_0)$ can be referred to as a inversed link function (see McCullagh and Nelder (1989)), and thus $g(\beta^T x;\theta)$

* This research partially supported by the National Science Council, R.O.C. (Plan No. NSC 82-0208-M-032-023-T).

can be considered as a parametric family of link functions. Under $\theta = \theta_0$, the model of form given by (1.1) and (1.2) were discussed by McCullagh and Nelder (1989) using the quasi-likelihood approach.

Suppose the observed random data are $\{(Y_i, X_i), i = 1, \ldots, n\}$, where Y_i is a real-valued response variable and $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ is a p-dimensional column vector of covariates. Then the quasi-likelihood estimate $(\hat{\beta}, \hat{\theta}_n)$ is a set of solution to the system of equations:

$$\sum_{i=1}^{n} \frac{Y_i - g(\beta^T X_i; \theta)}{V[g(\beta^T X_i; \theta)]} \cdot \frac{\partial g(\beta^T X_i; \theta)}{\partial \beta} = 0 \tag{1.3}$$

and

$$\sum_{i=1}^{n} \frac{Y_i - g(\beta^T X_i; \theta)}{V[g(\beta^T X_i; \theta)]} \cdot \frac{\partial g(\beta^T X_i; \theta)}{\partial \theta} = 0. \tag{1.4}$$

The parameter ϕ is unknown. In the absence of information beyond two moments of Y given $X = x$, there is little alternative to using

$$\hat{\phi}_n = \frac{1}{n - p - q} \sum_{i=1}^{n} \frac{[Y_i - g(\hat{\beta}^T X_i; \hat{\theta}_n)]^2}{V[g(\hat{\beta}^T X_i; \hat{\theta}_n)]}.$$

General asymptotic theory for $\hat{\beta}$, $\hat{\theta}_n$ and $\hat{\phi}_n$ are given in Cheng and Wu (1994).

Usually, there is no explicit solution to (1.3) and (1.4). Numerical method must be frequently employed to determine the solution, though not every numerical procedure can guarantee a solution for the system of equations (1.3) and (1.4) even if the solution exists.

A simple estimation method is proposed in this paper for solving the above problem. My main approach utilizes the dimension reduction technique, (see Powell *et al.* (1989), Duan and Li (1987), and Brillinger (1982)) to first reduce the dimension of the regressor X to one dimension before solving the quasi-likelihood equations. In Subsection 2.1, I discuss the simple estimation procedure and show, under the sufficient conditions, the estimator $\hat{\beta}_*$ based on the simple estimation method is asymptotically multivariate normal. The consistent estimator of the asymptotic covariance matrix of $\hat{\beta}_*$ can be derived. Another approach is to use the quasi-likelihood estimate $\tilde{\beta}$ based on the estimated link using the estimate $\tilde{\theta}_n$ developed in Subsection 2.2. The estimator $\tilde{\beta}$ is also asymptotically multivariate normal. Further, the consistent estimator of the asymptotic covariance matrix of the estimator $\tilde{\beta}$ exists.

In Section 3, I compare the asymptotic properties of the estimators $\hat{\beta}$, $\hat{\beta}_*$ and $\tilde{\beta}$ under linear model. Further, general finite sample properties of the estimates are discussed using simulations. The simulation performances of $\hat{\beta}$, $\hat{\beta}_*$ and $\tilde{\beta}$ are very similar, however, it deserves to be mentioned that in my Monte Carlo simulation, the cost of CPU time for computing $\hat{\beta}_*$ is only about $\frac{1}{7}$ of that required for computing the estimator $\hat{\beta}$, on the average. On the other hand, the cost of CPU time for computing $\tilde{\beta}$ is about $\frac{1}{4}$ of that required for computing $\hat{\beta}$. In general, the computations of $\hat{\beta}_*$ and $\tilde{\beta}$ are much simpler.

2. Estimates of regression coefficients

2.1 *Simpler estimation procedure*

Let $\{(Y_i, X_i), i = 1, 2, \ldots, n\}$ denote a random sample, where Y_i is a real-valued response variable and $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ is a p-dimensional column vector of covariates. Theoretically, the quasi-likelihood estimates $(\hat{\theta}_n, \hat{\beta}_n)$, under the certain conditions given in Cheng and Wu (1994), will converge in probability to a constant vector (θ_0, β_0). However, in practical applications, if the dimension of the parameters (θ, β) is too large, then the solution of the system of equations (1.3) and (1.4) is not easy to derive. To overcome the computation difficulties, here I shall start with the technique which is especially useful for reducing the dimension of the covariate space. My aim is to first estimate some proportion δ_0 of the true regression coefficients β_0. Let $\gamma_0 \delta_0 = \beta_0$ where γ_0 is some scalar constant. If δ_0 is estimable and suppose δ_0 is regarded as known, then one can write $g(\beta^T x; \theta) = g(\gamma z; \theta)$ with $z = \delta_0^T x$ being the reduced one-dimensional covariate.

The ordinary least squares method provides the simplest and most popular estimates of δ_0. It is defined as

$$\hat{\delta}_n = \sum_{i=1}^{n} Y_i (X_i - \bar{X})^T \cdot \left[\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \right]^{-1};$$

see Brillinger (1982) and Duan and Li (1987). The estimator $\hat{\delta}_n$ is $\sqrt{n}$-consistent estimator of δ_0 with unknown γ_0 if I assume the design condition satisfying: $E(\omega^T X \mid \beta_0^T X)$ is linear in $\beta_0^T X$ for all linear combinations $\omega^T X$ of X (see Duan and Li (1987), Li and Duan (1989)). Although this condition is satisfied by random covariates with jointly elliptically symmetric distribution (see Fang *et al.* (1990)). Nevertheless, this condition may not be fulfilled by the usual polynomial regression models. In a technical report, Cheng and Wu (1991) argued that $\hat{\delta}_n$ is not very robust against the violation of the above design condition.

A more robust estimator of δ_0 is

$$\tilde{\delta}_n = \frac{-2}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(\frac{1}{h} \right)^{p+1} \cdot K' \left(\frac{X_i - X_j}{h} \right) \cdot Y_i,$$

which was suggested and studied by Powell *et al.* (1989). Here the kernel $K(\cdot)$ is a weighting function and $h = h_n \to 0$, as $n \to \infty$, is a smoothing parameter that depends on the sample size n. The theoretical conditions of the kernel function and the smoothing parameter were given in Powell *et al.* (1989). Moreover, I conjecture that $\tilde{\delta}_n$ is strongly consistent if $\frac{nh^{p+1}}{\log n} \to \infty$ and the kernel function $K(\cdot)$ satisfies Condition 6 of Prakasa Rao ((1983), p. 183). In addition, other estimation method for the parameter δ_0 such as AD method (see Härdle and Stoker (1989)) and SIR method (see Duan and Li (1991), Li (1991)) may also be considered. But, here, I only consider the estimators $\hat{\delta}_n$ and $\tilde{\delta}_n$.

Let $\tilde{Z}_i = \tilde{\delta}_n^T X_i$, $i = 1, 2, \ldots, n$, and $\xi = (\gamma, \theta^T)^T$. Then I consider the quasi-likelihood estimator $\tilde{\xi}_n = (\tilde{\gamma}_n, \tilde{\theta}_n^T)^T$ to be a set of solution to the system of equations:

$$\sum_{i=1}^{n} \frac{Y_i - g(\gamma \tilde{Z}_i; \theta)}{V[g(\gamma \tilde{Z}_i; \theta)]} \cdot \frac{\partial g(\gamma \tilde{Z}_i; \theta)}{\partial \gamma} = 0 \tag{2.1}$$

and

$$\sum_{i=1}^{n} \frac{Y_i - g(\gamma \tilde{Z}_i; \theta)}{V[g(\gamma \tilde{Z}_i; \theta)]} \cdot \frac{\partial g(\gamma \tilde{Z}_i; \theta)}{\partial \theta} = 0. \tag{2.2}$$

In the following, I shall discuss some properties of the quasi-likelihood estimators $\tilde{\gamma}_n$ and $\tilde{\theta}_n$. Before stating the basic conditions and asymptotic results, I first denote $\zeta = (\xi, \delta)$, $\tilde{\zeta}_n = (\tilde{\xi}_n, \tilde{\delta}_n)$, $\zeta_0 = (\xi_0, \delta_0)$, $\frac{\partial L(\zeta; x, y)}{\partial \xi} = \frac{y - g(\gamma(\delta^T x); \theta)}{V[g(\gamma(\delta^T x); \theta)]} \cdot \frac{\partial g(\gamma(\delta^T x); \theta)}{\partial \xi}$, and $g(\zeta; x) \equiv g(\gamma(\delta^T x); \theta)$ for the ease of presentation. According to the proof and assumptions Conditions (R1)–(R4) of Theorem 2.1 are basically very similar to the classical treatment of the standard M.L.E. given in Serfling ((1980), pp. 144–149), and hence the proof will be shortened in here.

THEOREM 2.1. *Suppose assumptions Conditions* (R1)–(R4) *given in the Appendices are satisfied. Then there exists a sequence of solutions* $\tilde{\xi}_n$ *to the system of equations* (2.1) *and* (2.2) *such that as* $n \to \infty$

(i) $\tilde{\xi}_n \to \xi_0 = (\gamma_0, \theta_0)$, *with probability one,*

and

(ii) $\sqrt{n}(\tilde{\xi}_n - \xi_0) \xrightarrow{d} MVN(0, \Sigma_0)$,

where $\Sigma_0 = V_*^{-1} \Sigma^* V_*^{-1}$, $V_* = -E[\frac{\partial^2 L(\zeta_0; X, Y)}{\partial \xi^2}]$, $\Sigma^* = \mathcal{F}^T \cdot \Sigma_{\delta_0} \cdot \mathcal{F} + \Sigma_{11} + 2\mathcal{F}^T \cdot \Sigma_{12}^T$, $\mathcal{F}^T = E[\frac{\partial^2 L(\zeta_0; X, Y)}{\partial \xi \partial \delta}]$, $\Sigma_{12} = E\{[\frac{\partial L(\zeta_0; X, Y)}{\partial \xi}] \cdot [2(P(X, Y) - E(P(X, Y)))]^T\}$, $\Sigma_{11} = E[\frac{\partial L(\zeta_0; X, Y)}{\partial \xi}] \cdot [\frac{\partial L(\zeta_0; X, Y)}{\partial \xi}]^T$, $\Sigma_{\delta_0} = 4E[P(X, Y) \cdot (P(X, Y))^T] - 4\delta_0 \cdot \delta_0^T$, $P(x, y) = f(x) \cdot [\frac{\partial g(\beta^T x; \theta)}{\partial x}] - [y - g(\beta^T x; \theta)] \cdot \frac{\partial f(x)}{\partial x}$, *and* $f(x)$ *is the probability density function of* X.

PROOF. By (R1), for every j, and ξ in the neighborhood $N(\xi_0)$, Taylor's expansion of $\frac{\partial L(\zeta; x, y)}{\partial \xi_j}$ about the point $\xi = \xi_0$ gives:

$$\frac{\partial L(\xi, \tilde{\delta}_n; x, y)}{\partial \xi_j} - \frac{\partial L(\xi_0, \tilde{\delta}_n; x, y)}{\partial \xi_j} = \left[\frac{\partial^2 L(\xi_0, \tilde{\delta}_n; x, y)}{\partial \xi_j \partial \xi}\right] \cdot (\xi - \xi_0) + \frac{1}{2}(\xi - \xi_0)^T \cdot M_j(\tilde{\xi}; \tilde{\delta}_n, x, y) \cdot (\xi - \xi_0)$$

where, $\tilde{\xi}$ lies in the line segment between ξ and ξ_0 and

$$M_j(\tilde{\xi}, \tilde{\delta}_n; x, y) = \left[\frac{\partial^3 L(\tilde{\xi}; \tilde{\delta}_n, x, y)}{\partial \xi_i \partial \xi_j \partial \xi_k}\right]_{(q+1) \times (q+1)}.$$

Define $A_n = (A_{n1}, \ldots, A_{n(q+1)})^T$, $B_n = (B_{n1}, \ldots, B_{n(q+1)})^T$, $C_n = (C_{n1}, \ldots C_{n(q+1)})^T$ where,

$$A_{nj} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial L(\xi_0, \tilde{\delta}_n; X_i, Y_i)}{\partial \xi_j}, \qquad B_{nj} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 L(\xi_0, \tilde{\delta}_n; X_i, Y_i)}{\partial \xi_j \partial \xi},$$

$$C_{nj} = \frac{1}{n}\sum_{i=1}^{n} M_j(\tilde{\xi}, \tilde{\delta}_n; X_i, Y_i).$$

Define $U_n(\xi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial L(\xi, \tilde{\delta}_n; X_i, Y_i)}{\partial \xi}$, then

$$U_n(\xi) = A_n + B_n \cdot (\xi - \xi_0) + \frac{1}{2}[(\xi - \xi_0)^T C_{n1}(\xi - \xi_0), \ldots, (\xi - \xi_0)^T C_{n(q+1)}(\xi - \xi_0)]^T.$$

(a) $A_n \xrightarrow{w.p.1} 0$, $n \to \infty$

(by (1.1), (R3), $\tilde{\delta}_n \xrightarrow{w.p.1} \delta_0$, and the strong law of large numbers (S.L.L.N)).

(b) $B_{nj} \xrightarrow{w.p.1} V_j(\zeta_0)$, $n \to \infty$, $\forall j$ where $V_j(\zeta_0) = E[\frac{\partial^2 L(\zeta_0; X, Y)}{\partial \xi_j \partial \xi}] < \infty$, $j = 1, 2, \ldots, q+1$

(by (R1), (R3), (R4), $\tilde{\delta}_n \xrightarrow{w.p.1} \delta_0$, and S.L.L.N).

(c) For every j,

$$|C_{nj}| \le \frac{1}{n}\sum_{i=1}^{n} |M_j(\tilde{\xi}, \tilde{\delta}_n; X_i, Y_i)| \le \frac{1}{n}\sum_{i=1}^{n} H_j(X_i, Y_i) \xrightarrow{w.p.1} E[H_j(X, Y)] \equiv C_j, \qquad n \to \infty$$

(by (R2), $\tilde{\delta}_n \xrightarrow{w.p.1} \delta_0$, and S.L.L.N);

and, when n is sufficiently enough, for every j, I has

$$|C_{njks}| \le C_{jks} + 1, \qquad \text{with probability one,}$$

where, $C_{nj} = [C_{njks}]$ and $C_j = [C_{jks}]$, $j, k, s = 1, 2, \ldots, q+1$.

(d) By the asymptotic result of $\tilde{\delta}_n$ is

$$\sqrt{n}(\tilde{\delta}_n - \delta_0) = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}[P(X_i, Y_i) - E(P(X, Y))] + o_p(1)$$

given in Powell *et al.* ((1989), pp. 1410–1412), $\tilde{\delta}_n \xrightarrow{w.p.1} \delta_0$, (R1), (R3), (R4), and S.L.L.N, I has

$$\sqrt{n}A_n \xrightarrow{d} MVN(0, \Sigma^*), \qquad n \to \infty.$$

(e) By (a), (b), and (c), there exists $n_0 = n_0(\eta, \varepsilon)$, such that for all $n > n_0(\eta, \varepsilon)$

$$p\{|A_{nj}| < \eta; |B_{njk} - V_{jk}(\zeta_0)| < \eta; |C_{njks}| < c\} > 1 - \varepsilon,$$

where, c is the finite positive number and larger than $\{\max_{j,k,s} C_{jks}+1\}$; η and ε are two small positive number. Then as in Chanda (1954), I can prove (i).

To proof the result (ii) of Theorem 2.1, I note that for n sufficiently large,

$$-\left\{B_{nj}+\frac{1}{2}(\tilde{\xi}_n-\xi_0)^T C_{nj}\right\}\cdot(\tilde{\xi}_n-\xi_0)=A_{nj}+o_p(n^{-1/2}),$$
$$j=1,2,\ldots,q+1,$$

so, $(\tilde{\xi}_n-\xi_0)=(V_*)^{-1}\cdot A_n+o_p(n^{-1/2})$.

Thus, by (d), $\sqrt{n}(\tilde{\xi}_n-\xi_0)\xrightarrow{d} MVN(0,\Sigma_0)$, $n\to\infty$.

This establishes the assertion (ii) of Theorem 2.1. □

Σ_0 is related in making inferences about β. One may use $\hat{\beta}^*=\tilde{\gamma}_n\tilde{\delta}_n$ to estimate β. To measure the precision of $\hat{\beta}_*$, I need to consider consistent estimator $\hat{\Sigma}_0$, of Σ_0. I define $\hat{\Sigma}_0=(\hat{V}_*)^{-1}\hat{\Sigma}_*(\hat{V}_*)^{-1}$, where $\hat{V}_*=\frac{-1}{n}\sum_{i=1}^n[\frac{\partial^2 L(\tilde{\zeta}_n;X_i,Y_i)}{\partial\xi^2}]$, $\hat{\Sigma}^*=\hat{\mathcal{F}}^T\cdot\hat{\Sigma}_{\delta_0}\cdot\hat{\mathcal{F}}+\hat{\Sigma}_{11}+2\hat{\mathcal{F}}^T\cdot\hat{\Sigma}_{12}^T$, $\hat{\mathcal{F}}^T=\frac{1}{n}\sum_{i=1}^n[\frac{\partial^2 L(\tilde{\zeta}_n;X_i,Y_i)}{\partial\xi\partial\delta}]$, $\hat{\Sigma}_{\delta_0}=4[\frac{1}{n}\sum_{i=1}^n\hat{P}_n(X_i,Y_i)\hat{P}_n^T(X_i,Y_i)-\tilde{\delta}_n\cdot\tilde{\delta}_n^T]$, $\hat{P}_n(x,y)=f_n(x)\cdot\frac{\partial g(\hat{\beta}_*^T x,\tilde{\theta}_n)}{\partial x}-[y-g(\hat{\beta}_*^T x,\tilde{\theta}_n)]\cdot\frac{\partial f_n(x)}{\partial x}$, $f_n(x)=\frac{1}{n}\sum_{i=1}^n(\frac{1}{h})^p\cdot K(\frac{x-X_i}{h})$, $\hat{\Sigma}_{11}=\frac{1}{n}\sum_{i=1}^n[\frac{\partial L(\tilde{\zeta}_n;X_i,Y_i)}{\partial\xi}]\cdot[\frac{\partial L(\tilde{\zeta}_n;X_i,Y_i)}{\partial\xi}]^T$, $\hat{\Sigma}_{12}=\frac{1}{n}\sum_{i=1}^n\{(\frac{\partial L(\tilde{\zeta}_n;X_i,Y_i)}{\partial\xi})\cdot[2(\hat{P}_n(X_i,Y_i)-\bar{P}_n)]^T\}$, and $\bar{P}_n=\frac{1}{n}\sum_{i=1}^n\hat{P}_n(X_i,Y_i)$.

Finally, I also state the asymptotic properties of the estimators $\hat{\beta}_*=\tilde{\gamma}_n\tilde{\delta}_n$.

COROLLARY 2.1. *Suppose the conditions in Theorem* 2.1 *are satisfied, then, as* $n\to\infty$

(i) $\hat{\beta}_*\to\beta_0=\gamma_0\delta_0$, *with probability one,*

and

(ii) $\sqrt{n}(\hat{\beta}_*-\beta_0)\xrightarrow{d} MVN(0,\Sigma_*)$,

where $\Sigma_*=\gamma_0^2\Sigma_{\delta_0}+\delta_0\Sigma_1^*\delta_0^T+\gamma_0(\delta_0\sigma_{12}^T+\sigma_{12}\delta_0^T)$, Σ_1^* *is the* 1×1 *submatrix of*

$$\Sigma_0=\begin{bmatrix}\Sigma_1^* & \Sigma_{12}^0\\ \Sigma_{12}^{0T} & \Sigma_2^0\end{bmatrix},$$

$\sigma_{12}=E[(V_1^*\{\frac{\partial L(\zeta_0;X,Y)}{\partial\xi}\}+\mathcal{F}_1^*\cdot[2(P(X,Y)-E(P(X,Y)))])\cdot[2(P(X,Y)-EP(X,Y))]]$, V_1^* *is the* $q+1$ *row vector of the matrix*

$$V_*^{-1}=\begin{bmatrix}V_1^*\\ V_2^*\end{bmatrix}$$

and $\mathcal{F}_1^*$ *is* p *row vector of the matrix*

$$[V_*^{-1}\cdot\mathcal{F}^T]=\begin{bmatrix}\mathcal{F}_1^*\\ \mathcal{F}_2^*\end{bmatrix}.$$

Thus, the asymptotic variance-covariance matrix Σ_* *can be consistently estimated by* $\hat{\Sigma}_*$, *where* $\hat{\Sigma}_*=\tilde{\gamma}_n^2\cdot\hat{\Sigma}_{\delta_0}+\tilde{\delta}_n\cdot\hat{\Sigma}_1^*\cdot\tilde{\delta}_n^T+\tilde{\gamma}_n(\tilde{\delta}_n\cdot\hat{\sigma}_{12}^T+\hat{\sigma}_{12}\cdot\tilde{\delta}_n^T)$ $\hat{\Sigma}_1^*$ *is the* 1×1 *submatrix of*

$$\hat{\Sigma}_0=\begin{bmatrix}\hat{\Sigma}_1^* & \hat{\Sigma}_{12}^0\\ \hat{\Sigma}_{12}^{0T} & \hat{\Sigma}_2^0\end{bmatrix},$$

$\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n}\{(\hat{V}_1^*[\frac{\partial L(\tilde{\zeta}_n;X_i,Y_i)}{\partial\xi}] + \hat{\mathcal{F}}_1^*[2(\hat{P}_n(X_i,Y_i) - \bar{P}_n)]) \cdot [2(\hat{P}_n(X_i,Y_i) - \bar{P}_n)]\}$, $\hat{V}_1^*$ *is the* $q+1$ *row vector of the matrix*

$$\hat{V}_*^{-1} = \begin{bmatrix} \hat{V}_1^* \\ \hat{V}_2^* \end{bmatrix},$$

and $\hat{\mathcal{F}}_1^*$ *is* p *row vector of the matrix*

$$[\hat{V}_*^{-1}\hat{\mathcal{F}}^T] = \begin{bmatrix} \hat{\mathcal{F}}_1^* \\ \hat{\mathcal{F}}_2^* \end{bmatrix},$$

and $\hat{\Sigma}_{\delta_0}$, $\hat{P}_n(x,y)$, $\bar{P}_n$ *as the above definition.*

2.2 *Estimation procedure based on estimated link model*

I consider the random data in the Subsection 2.1. According to the proof of Theorem 2.1, I have:

$$\tilde{\theta}_n = \theta_0 + \frac{1}{n}\sum_{i=1}^{n} Q_n(X_i,Y_i) + o_p(n^{-1/2}) \tag{2.3}$$

where the vector $Q_n(x,y)$ is the $q \times 1$ subvector of the matrix $V_*^{-1}[\frac{\partial L(\zeta_0;x,y)}{\partial\xi} + E\{\frac{\partial^2 L(\zeta_0;X,Y)}{\partial\xi\partial\delta}\} \cdot [2(P(x,y) - EP(X,Y))]^T]$.

Given the estimator $\tilde{\theta}_n$ derived in Subsection 2.1, I have an estimated link model defined by $g(\cdot;\tilde{\theta}_n)$. Based on the estimated link model, I arrive at the following equation:

$$\sum_{i=1}^{n} \frac{Y_i - g(\beta^T X_i;\tilde{\theta}_n)}{V\{g(\beta^T X_i;\tilde{\theta}_n)\}} \cdot \frac{\partial g(\beta^T X_i;\tilde{\theta}_n)}{\partial\beta} = 0. \tag{2.4}$$

Let the solution of (2.4) be denoted by $\tilde{\beta}$.

The asymptotic theory for the estimator $\tilde{\beta}$ is very similar to that developed in Theorem 2.1. Moreover, the conditions for the asymptotic theory are also very similar to those given in Appendices. I simply replace $\partial^i L(\xi,\delta;x,y)$ by $\partial^i L(\beta,\theta;x,y)$, $\zeta = (\xi,\delta)$ by $\alpha = (\beta,\theta)$, $g(\gamma(\delta^T x);\theta)$ by $g(\beta^T x;\theta)$, and $(\tilde{\xi}_n,\tilde{\delta}_n)$ by $(\tilde{\beta},\tilde{\theta}_n)$. I also replace $\partial\xi = (\partial\gamma,\partial\theta)$ by $\partial\beta$, $\partial\delta$ by $\partial\theta$, and $q+1$ by p in the Conditions (R1)–(R4), but replace $\tilde{\delta}_n$ by $\tilde{\theta}_n$, δ by θ, $\xi = (\gamma,\theta)$ by β, and $[2(P(X,Y) - EP(X,Y))]$ by $Q_n(X,Y)$. Consequently, I have

THEOREM 2.2. *Suppose the above corrected conditions in the Appendices are satisfied. Then there exists a sequence of solutions* $\tilde{\beta}$ *to the equation* (2.4) *such that as* $n \to \infty$

(i) $\tilde{\beta} \to \beta_0$, *with probability one,*

and

(ii) $\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} MVN(0,\tilde{\Sigma}_0)$,

where the covariance matrix $\tilde{\Sigma}_0 = \tilde{V}_*^{-1}\tilde{\Sigma}^*\tilde{V}_*^{-1}$, $\tilde{V}_* = -E\{\frac{\partial^2 L(\alpha_0;X,Y)}{\partial\beta^2}\}$, $\alpha_0 = (\beta_0, \theta_0)$, $\tilde{\Sigma}^* = \tilde{\mathcal{F}}^T \cdot \Sigma_{\theta_0} \cdot \tilde{\mathcal{F}} + \tilde{\Sigma}_{11} + 2\tilde{\mathcal{F}}^T \cdot \tilde{\Sigma}_{12}^T$, $\tilde{\mathcal{F}}^T = E\{\frac{\partial^2 L(\alpha_0;X,Y)}{\partial\beta\partial\theta}\}$, $\Sigma_{\theta_0} = E[Q_n(X,Y)] \cdot [Q_n(X,Y)]^T$, $\tilde{\Sigma}_{11} = E\{\frac{\partial L(\alpha_0;X,Y)}{\partial\beta}\} \cdot \{\frac{\partial L(\alpha_0;X,Y)}{\partial\beta}\}^T$, $\tilde{\Sigma}_{12} = E\{[\frac{\partial L(\alpha_0;X,Y)}{\partial\beta}] \cdot [Q_n(X,Y)]^T\}$.

Similarly, to measure the precision of $\tilde{\beta}$, I note that $\tilde{\Sigma}_0$ can be consistently estimated by $\hat{\tilde{\Sigma}}_0$. Here I define $\hat{\tilde{\Sigma}}_0 = \hat{\tilde{V}}_*^{-1}\hat{\tilde{\Sigma}}^*\hat{\tilde{V}}_*^{-1}$, where $\hat{\tilde{\Sigma}}^* = \hat{\tilde{\mathcal{F}}}^T \cdot \hat{\Sigma}_{\theta_0} \cdot \hat{\tilde{\mathcal{F}}} + \hat{\tilde{\Sigma}}_{11} + 2\hat{\tilde{\mathcal{F}}}^T \cdot \hat{\tilde{\Sigma}}_{12}^T$, $\hat{\tilde{\mathcal{F}}}^T = \frac{1}{n}\sum_{i=1}^n \{\frac{\partial^2 L(\tilde{\beta},\tilde{\theta}_n;X_i,Y_i)}{\partial\beta\partial\theta}\}$, $\hat{\tilde{\Sigma}}_{11} = \frac{1}{n}\sum_{i=1}^n \{\frac{\partial L(\tilde{\beta},\tilde{\theta}_n;X_i,Y_i)}{\partial\beta}\} \cdot \{\frac{\partial L(\tilde{\beta},\tilde{\theta}_n;X_i,Y_i)}{\partial\beta}\}^T$, $\hat{\Sigma}_{\theta_0} = \frac{1}{n}\sum_{i=1}^n \{\hat{Q}_n(X_i,Y_i)\} \cdot \{\hat{Q}_n(X_i,Y_i)\}^T$, $\hat{\tilde{\Sigma}}_{12} = \frac{1}{n}\sum_{i=1}^n \{[\frac{\partial L(\tilde{\beta},\tilde{\theta}_n;X_i,Y_i)}{\partial\beta}] \cdot [\hat{Q}_n(X_i,Y_i)]^T\}$, and $\hat{Q}_n(X_i,Y_i)$ is the $q \times 1$ subvector of $[\frac{-1}{n}\sum_{i=1}^n \{\frac{\partial^2 L(\tilde{\xi}_n,\tilde{\delta}_n;X_i,Y_i)}{\partial\xi^2}\}]^{-1}[\frac{\partial L(\tilde{\xi}_n,\tilde{\delta}_n;X_i,Y_i)}{\partial\xi} + \{\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 L(\tilde{\xi}_n,\tilde{\delta}_n;X_i,Y_i)}{\partial\xi\partial\delta}\} \cdot [2(\hat{P}_n(X_i,Y_i) - \bar{P}_n)]^T]$, $\hat{\tilde{V}}_* = \frac{-1}{n}\sum_{i=1}^n \{\frac{\partial^2 L(\tilde{\beta},\tilde{\theta}_n;X_i,Y_i)}{\partial\beta^2}\}$.

3. Simulation studies and final remarks

In order to compare the finite sample properties of the estimates $\hat{\beta}$, $\hat{\beta}_*$ and $\tilde{\beta}$, a Monte Carlo experiment has been done in which 1000 samples of different size n were generated from different populations for the dependent variable and covariates. The computation of the estimate $\hat{\beta}_*$ involves the selection of the kernel function K and the smoothing parameter h. To stabilize the density estimates, I usually use positive product kernel. Thus I define

$$K(u_1, u_2, \ldots, u_p) = \prod_{i=1}^{p} K_*(u_i)$$

and let K_* be univariate "biweight" kernel

$$K_*(u) = \left(\frac{15}{16}\right)(1-u^2)^2 I(|u| \leq 1);$$

see also Müller (1984). Although my theoretical results do not constrain the choice of bandwidth h, some Monte Carlo experience suggests that reasonable small-sample performance is obtained by setting h in the range of 1 to 2 (one to two mean standard deviations of predictors), (also see Härdle and Stoker (1989)), so I set $h = 1.5\sum_{i=1}^p (\frac{S_i}{p})$ where S_i is the sample standard deviation of the predictor X_i.

I first consider the regressors $X_i = (X_{i1}, X_{i2}, X_{i3})^T$ to have jointly multivariate normal distribution with mean $(0,0,0)$ and the identity covariance matrix. For each given X_i, Y_i was generated from a binomial distribution with the number of trials $N = 50$ and the probability of success $p(X_i)$. For $X_i = (X_{i1}, X_{i2}, X_{i3})^T$, I consider two models for $p(X_i)$: *logit* $\{p(X_i)\} = \theta + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$, and $p(X_i) = \Phi(\theta + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3})$. Secondly, in addition, two models

were also investigated in my experiment. Model 1 considered $Y_i = \theta + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$. The regressors $X_i = (X_{i1}, X_{i2}, X_{i3})^T$ are independent of error ε_i and have jointly multivariate normal distribution with mean $(2, 1, 0.5)$ and variances and covariances: $\sigma_{11}^2 = \sigma_{22}^2 = 9$, $\sigma_{33}^2 = 64$, $\sigma_{12} = \sigma_{13} = 0$, $\sigma_{23} = 10$. Random error ε_i has identical logistic distribution with mean zero and variance= $\pi^2/3$. It is noted that the logistic distribution can be obtained as a mixture of extreme value distribution. Model 2 is basically the same as the Model 1 except that I define $X_{i2} = X_{i1}^2$ and the distribution of (X_{i1}, X_{i3}) is the marginal distribution derived from the multivariate distribution of Model 1. For Model 2, according to $X_{i2} = X_{i1}^2$, hence the dimension reduction technique is only used to (X_2, X_3). So in the system of equations (2.1) and (2.2), I replace $(\tilde{Z}_i = \tilde{\delta}_n^T X_i; \theta)$ by $(\tilde{Z}_{i*} = \tilde{\delta}_n^T X_{i*}, X_{i*} = (X_{i2}, X_{i3})^T; \theta_* = (\theta, \beta_1))$. The true parameter values are $\theta = 1$, and $(\beta_1, \beta_2, \beta_3) = (1, 1, 2)$.

Thus, the Monte Carlo estimates of the means of β_i are given in Tables 1-4. Also, for each case considered, the estimated mean squared errors are also calculated in order to measure the efficiencies of the estimates.

I find the experiment encouraging. Basically speaking, when the sample size is large, the biases of the estimates $\hat{\beta}$, $\hat{\beta}_*$, and $\tilde{\beta}$ are almost negligible. And, the mean squared errors of $\tilde{\beta}_i$ are slightly smaller than the other estimates. This means that comparing with the quasi-likelihood estimates $\hat{\beta}$ and $\hat{\beta}_*$, the overall performance of $\tilde{\beta}$ is more satisfactory.

Moreover, my simulations show that, the cost of CPU time for computing $\hat{\beta}_*$ is only about $\frac{1}{7}$ of that required for computing the estimator $\hat{\beta}$, on the average, and the cost of CPU time for computing $\tilde{\beta}$ is about $\frac{1}{4}$ of that required for computing the estimator $\hat{\beta}$. In general, the computations of the estimators $\tilde{\beta}$ and $\hat{\beta}_*$ are much simpler.

Table 1. Estimated means and empirical mean squared errors (in parentheses) of the estimators for the logit model.

n	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
50	0.9815	0.9914	1.9976
	(6.473)	(6.468)	(7.685)
100	0.9917	0.9954	1.9984
	(4.476)	(4.354)	(4.987)
n	$\hat{\beta}_{1*}$	$\hat{\beta}_{2*}$	$\hat{\beta}_{3*}$
50	0.9923	0.9831	1.9846
	(6.245)	(6.233)	(7.586)
100	1.0061	0.9928	1.9924
	(4.431)	(4.322)	(4.675)
n	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$
50	0.9898	0.9916	1.9914
	(6.234)	(6.156)	(7.575)
100	1.0032	1.0035	2.0075
	(4.385)	(4.321)	(4.647)

Note: The unit in parentheses is 10^{-4}.

Table 2. Estimated means and empirical mean squared errors (in parentheses) of the estimators for the Probit model.

n	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
50	0.9759	0.9673	1.9421
	(4.296)	(4.357)	(5.038)
100	1.0191	0.9978	2.0162
	(3.425)	(3.352)	(4.075)
n	$\hat{\beta}_{1*}$	$\hat{\beta}_{2*}$	$\hat{\beta}_{3*}$
50	0.9876	0.9968	1.9938
	(4.257)	(4.317)	(4.983)
100	1.0082	1.0084	2.0156
	(3.321)	(3.325)	(3.857)
n	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$
50	1.0047	1.0038	2.0081
	(4.237)	(4.287)	(4.973)
100	1.0035	1.0041	2.0075
	(3.319)	(3.323)	(3.785)

Note: The unit in parentheses is 10^{-4}.

Table 3. Estimated means and empirical mean squared errors (in parentheses) of the estimators for the Model 1.

n	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
50	1.0020	0.9828	1.9988
	(9.560)	(11.497)	(1.580)
100	1.0036	1.01828	1.9946
	(4.176)	(4.774)	(0.820)
n	$\hat{\beta}_{1*}$	$\hat{\beta}_{2*}$	$\hat{\beta}_{3*}$
50	1.0068	0.9905	1.9960
	(9.194)	(10.424)	(1.268)
100	1.0002	1.0110	2.0009
	(3.346)	(4.626)	(0.754)
n	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$
50	1.0092	0.9943	1.9940
	(9.011)	(9.887)	(1.112)
100	1.0074	0.9985	2.0040
	(2.931)	(4.552)	(0.721)

Note: The unit in parentheses is 10^{-3}.

Finally, I compare the asymptotic covariance matrices of the estimators under a linear model. Consider the random regressors $X_i = (X_{i1}, X_{i2}, X_{i3})^T$ to have jointly multivariate normal distribution with mean $(0, 0, 0)$ and the identity covariance matrix, and the response variable satisfying $Y_i = \theta + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$ where ε_i are independent of X_i, and $\varepsilon_i \sim N(0, 1)$, $i = 1, 2, \ldots, n$. Suppose the

Table 4. Estimated means and empirical mean squared errors (in parentheses) of the estimators for the Model 2.

n	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
50	1.0020	1.0018	2.0007
	(16.373)	(0.749)	(1.093)
100	0.9829	1.0033	2.0067
	(8.221)	(0.266)	(0.717)
n	$\hat{\beta}_{1*}$	$\hat{\beta}_{2*}$	$\hat{\beta}_{3*}$
50	0.9997	1.0046	1.9996
	(15.179)	(0.663)	(0.991)
100	0.9935	1.0044	2.0048
	(7.831)	(0.262)	(0.632)
n	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$
50	1.0101	0.9986	1.9999
	(14.582)	(0.620)	(0.940)
100	1.0049	0.9987	2.0038
	(7.636)	(0.259)	(0.589)

Note: The unit in parentheses is 10^{-3}.

true parameter values to be $\theta = 1$ and $(\beta_1, \beta_2, \beta_3) = (1, 1, 2)$. Then the asymptotic covariance matrix of $\sqrt{n}(\hat{\beta}_* - \beta)$ is the symmetric matrix A $= [a_{ij}]$ with $a_{11} = 2.0439724$, $a_{12} = -8.8118 \times 10^{-3}$, $a_{13} = -0.0176236$, $a_{22} = 2.0439724$, $a_{23} = -0.0176236$, $a_{33} = 2.017537$; But, the asymptotic covariance matrix of $\sqrt{n}(\tilde{\beta}_* - \beta)$ is the identity matrix where $\tilde{\beta}_* = \tilde{\gamma}_n \hat{\delta}_n$. Moreover, I also can prove that the asymptotic covariance matrices of $\sqrt{n}(\hat{\beta} - \beta)$ and $\sqrt{n}(\tilde{\beta} - \beta)$ are also the identity matrix. Therefore, the estimators $\hat{\beta}$, $\tilde{\beta}$ and $\tilde{\beta}_*$ have the same asymptotic covariance structure and seem to be better than $\hat{\beta}_*$.

Acknowledgements

I would like to express my sincerest thanks to Professor K. F. Cheng for his help and encouragement throughout this work. In addition, I also thank the associate editor and the referee for helpful comments.

Appendices

Assumptions for Theorem 2.1.

(R1) For all ζ in a neighborhood $N(\zeta_0)$ of ζ_0 where $\zeta = (\xi, \delta)$, the derivatives $\frac{\partial^2 L(\zeta;x,y)}{\partial \xi_i \partial \xi_j}$, $\frac{\partial^2 L(\zeta;x,y)}{\partial \xi_i \partial \delta}$ and $\frac{\partial^3 L(\zeta;x,y)}{\partial \xi_i \partial \xi_j \partial \xi_k}$ exist for all (x, y) in the support of (X, Y) and $i, j, k = 1, 2, \ldots, q+1$.

(R2) There exist function $H_i(x, y)$, possibly depending on ζ_0, and $i = 1, 2, \ldots, q+1$, such that $E\{H_i(X, Y)\} < \infty$, and for all ζ in $N(\zeta_0)$, and (x, y) in the support of (X, Y)

$$\left| \frac{\partial^3 L(\zeta; x, y)}{\partial \xi_i \partial \xi_j \partial \xi_k} \right| \leq H_i(x, y), \qquad i, j, k = 1, 2, \ldots, q+1.$$

(R3) The functions $\frac{\partial L(\zeta;x,y)}{\partial \xi_i}$, $\frac{\partial^2 L(\zeta;x,y)}{\partial \xi_i \partial \xi_j}$ and $\frac{\partial^2 L(\zeta;x,y)}{\partial \xi_i \partial \delta}$ are uniformly continuous at ζ_0 for all (x,y) in the support of (X,Y), $i,j = 1,2,\ldots,q+1$.

(R4) $E\{\frac{\partial L(\zeta_0;X,Y)}{\partial \xi}\}\cdot\{\frac{\partial L(\zeta_0;X,Y)}{\partial \xi}\}^T$ is a positive definite matrix, $E\{\frac{\partial^2 L(\zeta_0;X,Y)}{\partial \xi_i \partial \xi_j}\}$ is a nonsingular matrix, and $E\{\frac{\partial^2 L(\zeta_0;X,Y)}{\partial \xi \partial \delta}\} < \infty$.

REFERENCES

Brillinger, D. R. (1982). A generalized linear model with "Gaussian" regressor variables, *A Festchrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday* (eds. P. J. Bickel, K. Doksum and J. L. Hodges), 97–114, Wadsworth International Group, Belmont, California.

Chanda, K. C. (1954). A note on the consistency and maxima of the roots of likelihood equations, *Biometrika*, **41**, 56–61.

Cheng, K. F. and Wu, J. W. (1991). Adjusted least squares estimates for the scaled regression coefficients with censored data, Tech. Report, 90–108, National Central University, Chungli, Taiwan, R.O.C.

Cheng, K. F. and Wu, J. W. (1994). Testing goodness of fit for a parametric family of link function, *J. Amer. Statist. Assoc.*, **89**, 657–664.

Duan, N. and Li, K.-C. (1987). Distribution-free and link-free estimation for the sample selection model, *J. Econometrics*, **35**, 25–35.

Duan, N. and Li, K.-C. (1991). Slicing regression: A link free regression method, *Ann. Statist.*, **19**, 505–530.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*, Chapman and Hall, New York.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivative, *J. Amer. Statist. Assoc.*, **84**, 986–995.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction, *J. Amer. Statist. Assoc.*, **86**, 316–342.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation, *Ann. Statist.*, **17**, 1009–1052.

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

Müller, H.-G. (1984). Smooth optimum kernel, estimators of densities, regression curves and models, *Ann. Statist.*, **12**, 766–774.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients, *Econometrica*, **57**, 1403–1430.

Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation*, Academic Press, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley, New York.