Biometrika (2008), **95**, 2, *pp*. 489–507
doi: 10.1093/biomet/asm094
Advance Access publication 31 January 2008

Diagnostic measures for empirical likelihood of general estimating equations

BY HONGTU ZHU AND JOSEPH G. IBRAHIM

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420, U.S.A.
hzhu@bios.unc.edu ibrahim@bios.unc.edu

NIANSHENG TANG

Department of Statistics, Yunnan University, Kunming, China
nstang@ynu.edu.cn

AND HEPING ZHANG

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut 06520-8034, U.S.A.
heping.zhang@yale.edu

SUMMARY

We develop diagnostic measures for assessing the influence of individual observations when using empirical likelihood with general estimating equations, and we use these measures to construct goodness-of-fit statistics for testing possible misspecification in the estimating equations. Our diagnostics include case-deletion measures, local influence measures and pseudo-residuals. Our goodness-of-fit statistics include the sum of local influence measures and the processes of pseudo-residuals. Simulation studies are conducted to evaluate our methods, and real datasets are analyzed to illustrate the use of our diagnostic measures and goodness-of-fit statistics.

Some key words: Diagnostic measure; Empirical likelihood; Estimating equation; Goodness-of-fit statistic; Resampling method.

1. INTRODUCTION

Diagnostic measures such as residuals and Cook's distance have been widely used to identify influential observations in various regression models, such as generalized linear models (Cox & Snell, 1968; Cook & Weisberg, 1982; Davison & Tsai, 1992; Zhu et al., 2001). Cook (1986) proposed a general approach for assessing the local influence of a minor perturbation to a statistical model, which has been further investigated by many authors, such as Zhu & Lee (2001) and Critchley & Marriott (2004). In addition, classical diagnostic measures including residuals and Cook's distance can be used to construct goodness-of-fit statistics (Stute, 1997; Lin et al., 2002; Zhu & Zhang, 2004). However, little has been done to extend these diagnostic approaches to more general statistical models, and in particular to general estimating equations.

The method of general estimating equations provides a flexible framework for analyzing independent and correlated data (Hansen, 1982; Qin & Lawless, 1994; Owen, 2001; Imbens, 2002). Methods such as the generalized method of moments and empirical likelihood have

been implemented, and statistics have been constructed for testing possible misspecification of the estimating equations when the number of equations exceeds the number of parameters. For instance, Qin & Lawless (1994) applied empirical likelihood to combine the estimating equations, to prove asymptotic efficiency of the empirical likelihood estimators and to present an empirical likelihood ratio statistic as a goodness-of-fit statistic. Empirical likelihood has been used to substitute for a parametric likelihood in many settings, such as survival analysis and time series analysis (Qin & Lawless, 1994; Kitamura, 1997; Owen, 2001; Chen & Cui, 2003).

However, few diagnostic measures have been developed in the context of empirical likelihood, even though lack of robustness to outliers has been noted by Tsao & Zhou (2001). To the best of our knowledge, only three diagnostic measures, namely empirical likelihood displacement, length and shape measures, have been proposed (DiCiccio & Monti, 2001; Lazar, 2005). For instance, the empirical likelihood displacement measures the influence of an individual observation on the estimators based on empirical likelihood. Quantifying the magnitude of those diagnostic measures remains an open problem, and measures such as residuals and local influence have not been formally defined.

2. AN OVERVIEW OF EMPIRICAL LIKELIHOOD FOR COMBINING ESTIMATING EQUATIONS

We assume that $x_1, \ldots, x_n$ are independent observations from a d-variate unknown distribution F with a p-dimensional parameter $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$. Without assuming a parametric form for F, we can still infer about θ using $r(\geqslant p)$ functionally independent estimating functions

$$g(x, \theta) = (g_1(x, \theta), \ldots, g_r(x, \theta))^{\mathrm{T}},$$

which satisfy the unconditional moment condition

$$E_F\{g(x, \theta_0)\} = 0, \quad \text{for } \theta_0 \in \Theta, \tag{1}$$

where E_F denotes the expectation with respect to F. Equation (1) is often referred to as the estimating equations or moment condition model (Hansen, 1982; Qin & Lawless, 1994; Owen, 2001; Imbens, 2002).

Example 1. Let $x_1 = (y_1, x_{1,(2)}), \ldots, x_n = (y_n, x_{n,(2)})$ be independent observations from a d-variate distribution F such that

$$E(y_i|x_{i,(2)}) = \mu(x_{i,(2)}^{\mathrm{T}}\beta), \quad \mathrm{var}(y_i|x_{i,(2)}) = \sigma^2 V\{\mu(x_{i,(2)}^{\mathrm{T}}\beta)\}, \tag{2}$$

for $i = 1, \ldots, n$, where $y_i = x_{i,(1)}$, the expectation is taken with respect to the conditional distribution of y given $x_{(2)}$, $\theta = (\beta, \sigma^2) \in R^p$, $p = d$, and $\mu(\cdot)$ and $V(\cdot)$ are known functions. Following the reasoning described in Chen & Cui (2003), $g(x, \theta)$ can be chosen as

$$(e(x)\mu'(x_{(2)}^{\mathrm{T}}\beta)x_{(2)}^{\mathrm{T}}V\{\mu(x_{(2)}^{\mathrm{T}}\beta)\}^{-1}, [e(x)^2\sigma^{-4}V\{\mu(x_{(2)}^{\mathrm{T}}\beta)\}^{-1} - \sigma^{-2}]\omega(x_{(2)}^{\mathrm{T}}\beta))^{\mathrm{T}}, \tag{3}$$

where $e(x) = y - \mu(x_{(2)}^{\mathrm{T}}\beta)$, $\mu'(x_{(2)}^{\mathrm{T}}\beta) = d\mu(u)/du$ evaluated at $u = x_{(2)}^{\mathrm{T}}\beta$, and $\omega(x_{(2)}^{\mathrm{T}}\beta)$ is an $(r - p)$-dimensional weighted function. The first $(p - 1)$ components of $g(x, \theta)$ are associated with the quasi-score

$$\partial_\beta \int_y^{\mu(x_{(2)}^{\mathrm{T}}\beta)} (y - u)[V\{\mu(u)\}]^{-1}/du$$

and the last $(r - p)$ components of $g(x, \theta)$ are associated with the variance structure in (2), where ∂_β denotes partial differentiation with respect to β. A question of interest is to infer about θ using (3).

The empirical likelihood function is given by

$$L(F) = \prod_{i=1}^{n} dF(x_i) = \prod_{i=1}^{n} p_i$$

(Owen, 2001; Qin & Lawless, 1994), where $p_i = dF(x_i) = \mathrm{pr}(X_i = x_i)$. Moreover, the covariance matrix for F is assumed to be nonsingular. Under condition (1), a profile empirical likelihood ratio function for θ is defined as

$$L_E(\theta) = \sup\left\{\prod_{i=1}^{n} p_i \,|\, p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(x_i, \theta) = 0\right\}.$$

As shown in Qin & Lawless (1994) and Owen (2001), we have

$$L_E(\theta) = \prod_{i=1}^{n} n^{-1}\{1 + t_n(\theta)^{\mathrm{T}} g(x_i, \theta)\}^{-1}$$

and the value of $L_E(\theta)$ can be achieved by

$$p_i(\theta) = n^{-1}\{1 + t_n(\theta)^{\mathrm{T}} g(x_i, \theta)\}^{-1},$$

where $t_n(\theta)$, an $r \times 1$ vector, is the root of $\sum_{i=1}^{n} g(x_i, \theta)\{1 + t^{\mathrm{T}} g(x_i, \theta)\}^{-1} = 0$.

A maximum empirical likelihood estimator of θ, denoted by $\hat{\theta}$, can be obtained by maximizing the empirical loglikelihood function

$$l_E(\theta) = \sum_{i=1}^{n} l_{E,i}(\theta) = -\sum_{i=1}^{n} \log\{1 + t_n(\theta)^{\mathrm{T}} g(x_i, \theta)\},$$

where $l_{E,i}(\theta) = -\log\{1 + t_n(\theta)^{\mathrm{T}} g(x_i, \theta)\}$. Qin & Lawless (1994) established the asymptotic normality of $\hat{\theta}$ and $\hat{t} = \hat{t}_n(\hat{\theta})$: as $n \to \infty$, in distribution,

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\{0, C_\theta\}, \quad \sqrt{n}(\hat{t} - 0) \to N\{0, C_t\},$$

where the covariance matrices C_θ and C_t are given, respectively, by $C_\theta = (S_{21} S_{11}^{-1} S_{12})^{-1}$ and $C_t = S_{11}^{-1} + S_{11}^{-1} S_{12} S_{22\cdot1}^{-1} S_{21} S_{11}^{-1}$. In addition, we define $S_{22\cdot1} = -S_{21} S_{11}^{-1} S_{12}$ and

$$S = S(0, \theta_0) = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & 0 \end{pmatrix} = \begin{pmatrix} E_F(g^{\otimes 2}) & -E_F(\partial_\theta g)^{\mathrm{T}} \\ -E_F(\partial_\theta g) & 0 \end{pmatrix}_{(0,\theta_0)},$$

in which $g = g(x, \theta_0)$, $\partial_\theta g = \partial_\theta g(x, \theta_0)$, ∂_θ denotes partial differentiation with respect to θ and $a^{\otimes 2} = aa^{\mathrm{T}}$ for any vector a.

The maximum empirical likelihood estimator also has some properties under a misspecified model (Chen et al., 2007). Condition (1) is misspecified if $E_F\{g(x, \theta)\} \neq 0$ for all $\theta \in \Theta$. Even though some estimating equations in (1) are misspecified, $(\hat{t}, \hat{\theta})$ can be obtained by optimizing the empirical likelihood function:

$$(\hat{t}, \hat{\theta}) = \arg\max_{\theta} \ \arg\min_{t}\{Q_n(t, \theta)\},$$

where $Q_n(t, \theta) = -n^{-1} \sum_{i=1}^{n} \log\{1 + t^{\mathrm{T}} g(x_i, \theta)\}$. Under some conditions (Chen et al., 2007), $Q_n(t, \theta)$ converges to $Q(t, \theta) = -\int \log\{1 + t^{\mathrm{T}} g(x, \theta)\}/dF(x)$ almost surely (van der Vaart & Wellner, 1996, Ch. 2·4), and $(\hat{t}, \hat{\theta})$ converges to

$$(t_*, \theta_*) = \arg\max_{\theta} \ \arg\min_{t}\{Q(t, \theta)\}$$

in probability. In particular, if condition (1) holds, then $t_* = 0$ and $\theta_* = \theta_0$.

Moreover, empirical likelihood ratio statistics have been developed for testing linear or nonlinear hypotheses about θ and possible misspecification of the estimating equations under condition (1). For instance, $W_E(\theta_0) = 2l_E(\hat{\theta}) - 2l_E(\theta_0)$ has a limiting chi-squared distribution and can be used to test $H_0 : \theta = \theta_0$, while $W_1 = -2l_E(\hat{\theta})$, which converges weakly to a chi-squared distribution with $r - p$ degrees of freedom, can be used to test the assumption under (1).

3. THREE TYPES OF DIAGNOSTIC MEASURE FOR EMPIRICAL LIKELIHOOD OF ESTIMATING EQUATIONS

3·1. *Case-deletion influence measures*

To quantify the effects of deleting the ith observation x_i on $\hat{\theta}$, we define the maximum empirical likelihood estimators of θ for the full sample X and a subsample $X_{[i]}$, in which x_i is deleted from X, respectively, as follows. For the full sample X, we define

$$Q_n(t, \theta) = n^{-1} \sum_{i=1}^{n} \ell_i(t, \theta) = -n^{-1} \sum_{i=1}^{n} \log\{1 + t^{\mathrm{T}} g(x_i, \theta)\},$$

where $\ell_i(t, \theta) = -\log\{1 + t^{\mathrm{T}} g(x_i, \theta)\}$. Thus, we can obtain $\hat{\theta}$ and $\hat{t} = t_n(\hat{\theta})$ by jointly solving the following equations:

$$Q_{1,n}(t, \theta) = \partial_t Q_n(t, \theta) = -n^{-1} \sum_{i=1}^{n} g(x_i, \theta)\{1 + t^{\mathrm{T}} g(x_i, \theta)\}^{-1} = 0,$$

$$Q_{2,n}(t, \theta) = \partial_\theta Q_n(t, \theta) = -n^{-1} \sum_{i=1}^{n} \partial_\theta g(x_i, \theta) t\{1 + t^{\mathrm{T}} g(x_i, \theta)\}^{-1} = 0,$$

where ∂ denotes partial differentiation with respect to a parameter vector, such as t. For instance, $\partial_\theta g(x; \theta)$ is a $p \times r$ matrix with (k, l)th element $\partial_{\theta_k} g_l(x; \theta)$. For the subsample $X_{[i]}$, we define $Q_{n[i]}(t, \theta)$ as $Q_{n[i]}(t, \theta) = -n^{-1} \sum_{j \neq i}^{n} \log\{1 + t^{\mathrm{T}} g(x_j, \theta)\}$. Similarly to $\hat{\theta}$ and $\hat{t}$, we define $\hat{\theta}_{[i]}$ and $\hat{t}_{[i]}$ as the roots of

$$Q_{1,n[i]}(t, \theta) = \partial_t Q_{n[i]}(t, \theta) = 0, \quad Q_{2,n[i]}(t, \theta) = \partial_\theta Q_{n[i]}(t, \theta) = 0.$$

Following the reasoning in Cook & Weisberg (1982, Ch. 2), we introduce two case-deletion measures to quantify the distance between the maximum empirical likelihood estimator of θ with and without the ith observation deleted from the full sample. Cook's distance $\mathrm{ECD}_i(M)$ is given by

$$\mathrm{ECD}_i(M) = (\hat{\theta}_{[i]} - \hat{\theta})^{\mathrm{T}} M (\hat{\theta}_{[i]} - \hat{\theta}), \tag{4}$$

where M is chosen to be a positive definite matrix. For instance, M can be $-\partial_\theta^2 l_E(\theta)$ evaluated at $\hat{\theta}$, where ∂_θ^2 represents the second-order derivative of $l_E(\theta)$ with respect to θ. We use ECD_i to denote $\mathrm{ECD}_i(M)$ with $M = -\partial_\theta^2 l_E(\hat{\theta})$. Similar to the likelihood displacement (Cook, 1986), the empirical likelihood displacement (Lazar, 2005) is defined by

$$\mathrm{ELD}_i = 2\{l_E(\hat{\theta}) - l_E(\hat{\theta}_{[i]})\}. \tag{5}$$

If the values of ECD_i and ELD_i are large, then the ith observation is an influential point.

We can quantify the effects on $\hat{\theta}$ of deleting two or more observations (Cook & Weisberg, 1982). We define $\hat{\theta}_{[I]}$ as the maximum empirical likelihood estimator of θ for a subsample $X_{[I]}$, in which x_i for all $i \in I$ are deleted from X, where I is an index set with m_I observations, and

define

$$\text{ECD}_I(M) = (\hat{\theta}_{[I]} - \hat{\theta})^{\mathrm{T}} M (\hat{\theta}_{[I]} - \hat{\theta}), \quad \text{ELD}_I = 2\{l_E(\hat{\theta}) - l_E(\hat{\theta}_{[I]})\}. \tag{6}$$

However, calculating case-deletion measures requires evaluating $n!/\{m_I!(n - m_I)!\}$ cases, which can be computationally intensive for large m_I.

3·2. *Local influence measures*

We consider the local influence method for a case-weight perturbation $\omega \in R^n$, for which the empirical log-likelihood function $l_E(\theta|\omega)$ is defined by $l_E(\theta|\omega) = \sum_{i=1}^n \omega_i l_{E,i}(\theta)$. In this case, $\omega = \omega^0$, defined to be an $n \times 1$ vector with all elements equal to 1, represents no perturbation to the empirical likelihood, because $l_E(\theta|\omega^0) = l_E(\theta)$. Thus, the empirical likelihood displacement is defined as $\text{LD}_E(\omega) = 2[l_E(\hat{\theta}) - l_E\{\hat{\theta}(\omega)\}]$, where $\hat{\theta}(\omega)$ is the maximum empirical likelihood estimator of θ based on $l_E(\theta|\omega)$. Let $\omega(a) = \omega^0 + ah$ with $\omega(0) = \omega^0$ and $d\omega(a)/da|_{a=0} = h$, where h is a direction in R^n. Thus, the normal curvature of the influence graph $(\omega^{\mathrm{T}}, \text{LD}_E(\omega))^{\mathrm{T}}$ is given by

$$C_h(\omega^0) = h^{\mathrm{T}} H_{\text{LD}_E(\omega^0)} h,$$

where

$$H_{\text{LD}_E(\omega^0)} = -2 \left. \frac{\partial^2 \text{LD}_E\{\hat{\theta}(\omega)\}}{\partial\omega\partial\omega^{\mathrm{T}}} \right|_{\omega^0} = 2\Delta^{\mathrm{T}} \{-\partial_\theta^2 l_E(\theta)\}^{-1} \Delta \Big|_{\omega^0, \hat{\theta}},$$

in which $\Delta = \partial^2_{\theta\omega} \text{LD}_E(\theta, \omega)$ is a $p \times n$ matrix with (k, i)th element given by $\partial_{\theta_k} l_{E,i}(\theta)$.

We consider two local influence measures based on the normal curvature $C_h(\omega^0)$ as follows. Let $\lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant \lambda_{p+1} = \cdots = \lambda_n = 0$ be the ordered eigenvalues of the matrix $H_{\text{LD}_E(\omega^0)}$ and let $\{v_m = (v_{m1}, \ldots, v_{mn})^{\mathrm{T}} : m = 1, \ldots, n\}$ be the associated orthonormal basis, that is, $H_{\text{LD}_E(\omega^0)} v_m = \lambda_m v_m$. Thus, the spectral decomposition of $H_{\text{LD}_E(\omega^0)}$ is given by

$$H_{\text{LD}_E(\omega^0)} = \sum_{m=1}^{n} \lambda_m v_m v_m^{\mathrm{T}}.$$

The most popular local influence measures include v_1, which corresponds to the largest eigenvalue λ_1, as well as $C_{e_i} = \sum_{m=1}^p \lambda_m v_{mi}^2$, where e_i is an $n \times 1$ vector with ith component 1 and 0 otherwise. The v_1 represents the most influential perturbation to the empirical likelihood function, whereas the ith observation x_i with a large C_{e_i} can be regarded as influential.

3·3. *Pseudo-residuals*

The pseudo-residuals are key tools for revealing departures from assumption (1). We define a vector of pseudo-residuals for each observation, given by

$$R_i = (R_{i,1}, \ldots, R_{i,r})^{\mathrm{T}} = g(x_i, \hat{\theta}), \quad \text{for } i = 1, \ldots, n.$$

The R_i can be regarded as a generalization of residuals from a class of parametric models to general estimating equations (Cox & Snell, 1968). The values of the R_i may be used to detect anomalous or influential observations (Cook & Weisberg, 1982). Since $E_F(R_i)$ are close to zero under condition (1), it is worthwhile to inspect R_i against some function of data, which may provide an assessment of the adequacy of the estimating equations in (1). Moreover, test statistics based on pseudo-residuals can be constructed to assess an overall goodness-of-fit, see § 4·2.

We further develop standardized pseudo-residuals. We introduce $(\sigma_1^2, \ldots, \sigma_r^2) = \text{diag}\{E_F(g^{\otimes 2})\}$ and its estimator $(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_r^2)$, which is formally discussed in § 3·4. Then we

define a vector of standardized pseudo-residuals as

$$R_i^s = (R_{i,1}^s, \ldots, R_{i,r}^s)^{\mathrm{T}} = (g_1(x_i, \hat{\theta})/\hat{\sigma}_1, \ldots, g_r(x_i, \hat{\theta})/\hat{\sigma}_r)^{\mathrm{T}}.$$

We may consider x_i as an 'outlier' if $|R_{i,j}^s|$, for some j, exceeds a threshold, such as 3 (Cook & Weisberg, 1982, p. 22).

We also consider an alternative definition of standardized pseudo-residuals. In some statistical problems, such as regression, x_i may have a natural partition, $(x_{i,(1)}^{\mathrm{T}}, x_{i,(2)}^{\mathrm{T}})^{\mathrm{T}}$, and we have

$$E_F\{g(x_i, \theta)|X_{(2)}\} = 0, \quad \text{for all} \quad i = 1, \ldots, n,$$

where $X_{(2)} = (x_{1,(2)}, \ldots, x_{n,(2)})$. We define $(\sigma_{i,1}^2, \ldots, \sigma_{i,r}^2) = \mathrm{diag}[E_F\{g(x_i, \theta)^{\otimes 2}|X_{(2)}\}]$ as a function of $X_{(2)}$ and θ and define its estimator $(\hat{\sigma}_{i,1}^2, \ldots, \hat{\sigma}_{i,r}^2)$, as discussed in § 3·4. Thus, conditional on $X_{(2)}$, the standardized pseudo-residuals for the ith observation are defined by

$$R_{(c)i}^s = (R_{(c)i,1}^s, \ldots, R_{(c)i,r}^s)^{\mathrm{T}} = (g_1(x_i, \hat{\theta})/\hat{\sigma}_{i,1}, \ldots, g_r(x_i, \hat{\theta})/\hat{\sigma}_{i,r})^{\mathrm{T}}. \tag{7}$$

For example, consider a linear regression $y_i = x_{i,(2)}^{\mathrm{T}}\beta + \sigma\epsilon_i$, where $\epsilon_i \sim N(0, 1)$. Thus, x_i has a natural partition $(y_i, x_{i,(2)}^{\mathrm{T}})^{\mathrm{T}}$ and $x_{i,(1)} = y_i$. If $\hat{\beta}$ denotes the ordinary least-squares estimator of β, we choose $R_i = x_{i,(2)}(y_i - x_{i,(2)}^{\mathrm{T}}\hat{\beta})$ and thus

$$R_{(c)i}^s = \frac{y_i - x_{i,(2)}^{\mathrm{T}}\hat{\beta}}{\hat{\sigma}\sqrt{(1 - h_{ii})}}(1, \ldots, 1)^{\mathrm{T}},$$

where $\hat{\sigma}$ is a consistent estimator of σ and $h_{ii} = x_{i,(2)}^{\mathrm{T}}(\sum_{j=1}^n x_{j,(2)}^{\otimes 2})^{-1}x_{i,(2)}$.

3·4. *Stochastic behaviour of diagnostic measures under the correct model*

For diagnostic purposes, it is desirable to obtain a one-step, computationally feasible approximation of $\hat{\theta}_{[i]}$, because exact calculation of $\hat{\theta}_{[i]}$ requires running nested optimization routines for each observation (Owen, 2001, Ch. 12). Moreover, if the number of estimating equations r and the sample size n are relatively large, then calculating $\hat{\theta}_{[i]}$ exactly for each observation can be computationally prohibitive. For instance, for some nonlinear functions $\mu(\cdot)$ and $V(\cdot)$ in Example 1, calculating $\hat{\theta}$ and $\hat{\theta}_{[i]}$ for each observation can be computationally intensive (Chen & Cui, 2003).

We obtain the following theorems, for which the assumptions and detailed proofs can be found in the Appendix.

PROPOSITION 1. *Suppose that Assumptions* A1–A3 *in the Appendix and equation* (1) *are true. Then*

(i) *the one-step approximation for* $\hat{\theta}_{[i]}$ *is*

$$\hat{\theta}_{[i]} = \hat{\theta} + O_p(n^{-1}) = \hat{\theta} - n^{-1}S_{22\cdot1}^{-1}S_{21}S_{11}^{-1}g(x_i, \hat{\theta})\{1 + o_p(1)\};$$

(ii) *the one-step approximation for* $\hat{t}_{[i]}$ *is*

$$\hat{t}_{[i]} - \hat{t} = O_p(n^{-1}) = -n^{-1}(S_{11}^{-1} + S_{11}^{-1}S_{12}S_{22\cdot1}^{-1}S_{21}S_{11}^{-1})g(x_i, \hat{\theta})\{1 + o_p(1)\}.$$

Proposition 1(i) and (ii), respectively, provide the one-step approximations $\hat{\theta}$ and $\hat{t}$, which can be used to reduce the burden of calculating the maximum empirical likelihood estimator for each $X_{[i]}$. The matrices S_{11}, S_{12} and $S_{22\cdot1}$ can be approximated by their corresponding sample means, say $S_{11} \doteq n^{-1}\sum_{i=1}^n g(x_i, \hat{\theta})^{\otimes 2}$. The formulae can be further generalized to measure the effects

of deleting more than one observation on $\hat{\theta}$ and $\hat{t}$. For instance, if the index set is $I = \{i, j\}$, then the one-step approximations for $\hat{\theta}_{[i,j]}$ and $\hat{t}_{[i,j]}$ are, respectively, given by

$$\hat{\theta}_{[i,j]} - \hat{\theta} = O_p(n^{-1}) = -n^{-1} S_{22\cdot1}^{-1} S_{21} S_{11}^{-1} \{g(x_i, \hat{\theta}) + g(x_j, \hat{\theta})\}\{1 + o_p(1)\},$$

$$\hat{t}_{[i,j]} - \hat{t} = O_p(n^{-1}) = -n^{-1} (S_{11}^{-1} + S_{11}^{-1} S_{12} S_{22\cdot1}^{-1} S_{21} S_{11}^{-1}) \{g(x_i, \hat{\theta}) + g(x_j, \hat{\theta})\}\{1 + o_p(1)\}.$$

We examine the properties of pseudo-residuals, such as their expectations. We may then develop both formal and informal diagnostic tools for the examination of the adequacy of estimating equations. However, without additional information about the distribution of x_i, it is not feasible to determine the joint distribution of $g(x_i, \hat{\theta})$. We instead consider the expectations and variances of pseudo-residuals as follows.

PROPOSITION 2. *Suppose that Assumptions* A1–A3 *in the Appendix and equation* (1) *are true. Then*

$$E_F[g_k(x_i, \hat{\theta})] \doteq -n^{-1} E_F\{\partial_\theta g_k(x_i)^{\mathrm{T}} S_{22\cdot1}^{-1} S_{21} S_{11}^{-1} g(x_i)\} - n^{-1} \mathrm{tr}[E_F\{\partial_\theta^2 g_k(x_i)\} S_{22\cdot1}^{-1}]$$

$$\hat{\sigma}_k^2 \doteq \mathrm{var}_F\{g_k(x_i)\} - 2n^{-1} E_F\{g_k(x_i) \partial_\theta g_k(x_i)^{\mathrm{T}} S_{22\cdot1}^{-1} S_{21} S_{11}^{-1} g(x_i)\} - n^{-1} E_F\{\partial_\theta g_k(x_i)^{\mathrm{T}} S_{22\cdot1}^{-1} \partial_\theta g_k(x_i)\},$$

for $k = 1, \ldots, r$, *where* $g(x_i) = g(x_i, \theta_0)$ *and* $g_k(x_i) = g_k(x_i, \theta_0)$. *Furthermore, if we consider the standardized pseudo-residuals in* (7), *then*

$$\hat{\sigma}_{i,k} \doteq \mathrm{var}_F\{g_k(x_i) | X_{(2)}\} - 2n^{-1} E_F[g_k(x_i)\{\partial_\theta g_k(x_i)^{\mathrm{T}} S_{22\cdot1}^{-1} S_{21} S_{11}^{-1} g(x_i) - \mu_{i,k}\} | X_{(2)}] + n^{-1} E_F[\{\partial_\theta g_k(x_i)^{\mathrm{T}} S_{22\cdot1}^{-1} S_{21} S_{11}^{-1} g(x_i)\}^{\otimes 2} | X_{(2)}],$$

for $k = 1, \ldots, r$, *where* $\mu_{i,k} = E_F\{\partial_\theta g_k(x_i)^{\mathrm{T}} S_{22\cdot1}^{-1} S_{21} S_{11}^{-1} g(x_i) | X_{(2)}\}$.

Proposition 2 shows that the bias of $E_F(R_{i,k})$ has order n^{-1} under condition (1), so that $R_{i,k}$ should oscillate around 0. Therefore, if many $|R_{i,k}|$ are significantly greater than zero, then one should question whether or not all estimating functions in (1) are correctly specified. Moreover, all submatrices of S such as S_{11} can be consistently estimated using those of $S_n(\hat{t}, \hat{\theta})$.

We examine the properties of the case-deletion and local influence measures, such as their asymptotic expansions. These properties are useful for understanding the connections among these diagnostic measures.

THEOREM 1. *If Assumptions* A1–A3 *in the Appendix and condition* (1) *are satisfied, then*

$$C_{e_i} = 2\mathrm{ECD}_i\{1 + o_p(1)\} = 2\mathrm{ELD}_i\{1 + o_p(1)\} = -2n^{-1} \Delta_i^{\mathrm{T}} S_{22\cdot1}^{-1} \Delta_i \{1 + o_p(1)\} = O_p(n^{-1}), \tag{8}$$

$$\sum_{i=1}^{n} C_{e_i} = 2\sum_{i=1}^{n} \mathrm{ECD}_i\{1 + o_p(1)\} = 2\sum_{i=1}^{n} \mathrm{ELD}_i\{1 + o_p(1)\} = 2p + o_p(1), \tag{9}$$

where $\Delta_i = \partial_\theta l_{E,i}(x_i; \hat{\theta}) = S_{21} S_{11}^{-1} g(x_i, \hat{\theta}) + o_p(1)$.

Theorem 1 extends the classical diagnostic measure from parametric models to the empirical likelihood for estimating equations (Zhu & Zhang, 2004). Equation (8) gives a geometric interpretation for the two case-deletion measures $2\mathrm{ECD}_i$ and $2\mathrm{ELD}_i$, which are asymptotically equivalent to the normal curvature along the direction e_i. The asymptotic equivalence in (8) can be regarded as reminiscent of the equivalence between the empirical log likelihood ratio and Wald test statistics

within the framework of the empirical likelihood. Since the sum of the C_{e_i}'s is close to $2p$ and the x_i's are independently and identically distributed, each C_{e_i}'s should be near their mean, $2p/n$. Thus, a point with a large C_{e_i} value may be regarded as extremely influential (Cook, 1986; Zhu & Zhang, 2004). Similar arguments hold for ECD_i and ELD_i.

3·5. *Stochastic behaviour of diagnostic measures under misspecified model*

In this section, we evaluate the effects of misspecified estimating equations on the expectations of pseudo-residuals and the sums of the case-deletion and local influence measures. We also quantify the effects of deleting an observation on $\hat{\theta}$ under misspecified estimating equations.

We first define

$$\tilde{S}(t,\theta) = \begin{pmatrix} \tilde{S}_{11} & \tilde{S}_{12} \\ \tilde{S}_{21} & \tilde{S}_{22} \end{pmatrix}_{(t,\theta)}, \qquad \tilde{H}(t,\theta) = \begin{pmatrix} E_F(a^2 g^{\otimes 2}) & E_F\{a^2 g t^{\mathrm{T}} \partial_\theta g\} \\ E_F\{a^2 (\partial_\theta g)^{\mathrm{T}} t g^{\mathrm{T}}\} & E_F[a^2\{(\partial_\theta g)^{\mathrm{T}} t\}^{\otimes 2}] \end{pmatrix}_{(t,\theta)},$$

where $g = g(x,\theta)$, $a^{-1} = 1 + t^{\mathrm{T}} g(x,\theta)$, $\tilde{S}_{11} = E_F(a^2 g^{\otimes 2})$, $\tilde{S}_{12} = E_F(a^2 g t^{\mathrm{T}} \partial_\theta g) - E_F(a \partial_\theta g)$, $\tilde{S}_{21} = \tilde{S}_{21}^{\mathrm{T}}$, and $\tilde{S}_{22} = E_F[a^2\{(\partial_\theta g)^{\mathrm{T}} t\}^{\otimes 2}] - E_F(a t^{\mathrm{T}} \partial_\theta^2 g)$. We use the subscript $*$ to indicate that a vector, or a matrix, and its elements are evaluated at (t_*, θ_*). For instance, $\tilde{H}_* = \tilde{H}(t_*, \theta_*)$ and $\tilde{S}_* = \tilde{S}(t_*, \theta_*)$.

THEOREM 2. *If condition* (1) *is misspecified and Assumption* A4 *in the Appendix is true, then the following hold*:

(i) $E_F\{g_k(x_i,\hat{\theta})\} \doteq E_F\{g_k(x_i,\theta_*)\} = O(1)$ *for some* k;

(ii) $\hat{\theta}_{[i]} - \hat{\theta} = O_p(n^{-1}) = -n^{-1}\tilde{S}_{22\cdot 1*}^{-1}\Delta_{i*}\{1 + o_p(1)\}$,

$$\hat{t}_{[i]} - \hat{t} = O_p(n^{-1}) = -n^{-1}\{a_{i*}\tilde{S}_{11*}^{-1} g(x_i,\theta_*) + \tilde{S}_{11*}^{-1}\tilde{S}_{12*}\tilde{S}_{22\cdot 1*}^{-1}\Delta_{i*}\}\{1 + o_p(1)\}$$

where $a_{i*} = \{1 + t_*^{\mathrm{T}} g(x_i,\theta_*)\}^{-1}$,

$$\Delta_{i*} = \partial_\theta l_{E,i}(x_i;\theta_*) = -a_i\{\partial_\theta t_n(\theta_*) g(x_i,\theta_*) + \partial_\theta g(x_i,\theta_*) t_n(\theta_*)\}$$

and $\tilde{S}_{22\cdot 1*} = \tilde{S}_{22*} - \tilde{S}_{21*}\tilde{S}_{11*}^{-1}\tilde{S}_{12*}$;

(iii) $\sum_{i=1}^n C_{e_i} = 2\sum_{i=1}^n \mathrm{ECD}_i\{1 + o_p(1)\} = 2\sum_{i=1}^n \mathrm{ELD}_i\{1 + o_p(1)\} = \lambda_0 + o_p(1)$, *where*

$$\lambda_0 = \mathrm{tr}\{\tilde{S}_{22\cdot 1*}^{-1}(-\tilde{S}_{21*}\tilde{S}_{11*}^{-1}, I_p)\tilde{H}_*(-\tilde{S}_{21*}\tilde{S}_{11*}^{-1}, I_p)^{\mathrm{T}}\}. \tag{10}$$

Theorem 2 has some implications. Compared with $O(n^{-1})$ under the correct condition (1), Theorem 2(i) shows an order of $O(1)$ for the bias of the pseudo-residuals under the misspecified model. If all pseudo-residuals $g_k(x_i,\theta_*)$ are close to a positive, or negative, scalar for some k, then we may suspect some misspecification in (1). Theorem 2(ii) measures the effects of deleting an observation on the maximum empirical likelihood estimator under the misspecified model. Theorem 2(iii) indicates that $\sum_{i=1}^n C_{e_i}$ converges to λ_0, which is different from $2p$. Since λ_0 reduces to $2p$ under the corrected model, $\sum_{i=1}^n C_{e_i} - 2p$ can be used to check model misspecification; see § 4·1.

4. GOODNESS-OF-FIT STATISTICS

4·1. *Goodness-of-fit statistics based on local influence measures*

The sum of the local inference measure, or the case-deletion measure, for each case i may be used to test model misspecification in (1); see Theorems 1 and 2.

THEOREM 3. *If Assumptions* A1–A3 *in the Appendix and condition* (1) *are satisfied, then we have*

$$L_n(\hat{\theta}) = n^{-1/2}\sum_{i=1}^{n}(nC_{e_i} - 2p) = n^{-1/2}\sum_{i=1}^{n} k(x_i, \theta_0) + o_p(1),$$

where $k(x_i, \theta_0)$ *is defined in the proof. Furthermore, if* $\sigma_L^2 = E_F\{k(x_i, \theta_0)^2\} > 0$, *then* $L_n(\hat{\theta})$ *converges to* $N(0, \sigma_L^2)$ *in distribution.*

Theorem 3 has the following implications. The statistic $L_n(\hat{\theta})$ can be regarded as an information matrix test for the empirical likelihood, because, in parametric models, the sum of the local-influence measure for each case is closely related to White's (1982) information matrix test (Zhu & Zhang, 2004). Moreover, because of the asymptotic equivalence in (9), both $n^{-1/2}\sum_{i=1}^{n}(2n\mathrm{ECD}_i - 2p)$ and $n^{-1/2}\sum_{i=1}^{n}(2n\mathrm{ELD}_i - 2p)$ converge to $N(0, \sigma_L^2)$ in distribution under the correct model (1). Theorem 1 establishes the limiting distribution of $L_n(\hat{\theta})$ under the correct model (1) and gives an explicit formula for σ_L^2, the variance of $L_n(\hat{\theta})$. To estimate σ_L^2, we can calculate $\{k(x_i; \hat{\theta}) : i = 1, \ldots, n\}$ and the sample variance $s_{L,c}^2$ of $\{k(x_i; \hat{\theta}) : i = 1, \ldots, n\}$. Therefore, we can compute the p-value for $\mathcal{L}_n = L_n(\hat{\theta})/s_{L,c}$ using the standard normal distribution. Combining Theorems 1–3, we can obtain

$$L_n(\hat{\theta}) = \sqrt{n}\left(\sum_{i=1}^{n} C_{e_i} - \lambda_0\right) + \sqrt{n}(\lambda_0 - 2p) = \sqrt{n}\{\lambda_0 - 2p + o_p(1)\}.$$

Thus, $L_n(\hat{\theta}) \doteq n^{1/2}c_0$ when $\lambda_0 - 2p = c_0 \neq 0$ for a scalar c_0. Therefore, the statistic $\mathcal{L}_n = L_n(\hat{\theta})/s_{L,c}$ is a useful tool for testing the hypothesis $H_0 : \lambda_0 = 2p$ versus $H_1 : \lambda_0 \neq 2p$.

4·2. *Goodness-of-fit statistics based on pseudo-residuals*

To develop a residual-based test, we assume that z_i is a subcomponent of x_i for each observation and that

$$E_G\{g(x_i, \theta_0)|z_i\} = 0 \text{ for a } \theta_0 \in \Theta \text{ and } i = 1, \ldots, n, \tag{11}$$

where G is the conditional distribution of x given z. Note that (11) is only a sufficient condition for (1). Moreover, (11) arises naturally within the framework of regression, because estimating functions for regression models are primarily based on the conditional distribution of the responses given the covariates (Wooldridge, 1990; Stute, 1997; Lin et al., 2002). For instance, in regression problems, x_i has a natural partition $(x_{i,(1)}^{\mathrm{T}}, x_{i,(2)}^{\mathrm{T}})^{\mathrm{T}}$, where $x_{i,(1)}$ is a vector of responses and $x_{i,(2)}$ is a vector of covariates. In this case, z_i can be any subcomponent of $x_{i,(2)}$ or a function of $x_{i,(2)}$ (Lin et al., 2002; Stute, 1997).

We are interested in testing the following hypotheses:

$$H_0 : (11) \text{ is true} \quad \text{versus} \quad H_1 : (11) \text{ is not true.} \tag{12}$$

Proposition 1 has shown that, under the null hypothesis H_0, a plot of $R_{i,j}$ against z_i should oscillate around 0. This motivates us to combine the pseudo-residuals with z_i to construct several stochastic processes of $z \in [-\infty, \infty]$ as follows:

$$\mathrm{GF}_k(z; \hat{\theta}) = n^{-1/2}\sum_{i=1}^{n} R_{i,k}1(z_i \leqslant z) = n^{-1/2}\sum_{i=1}^{n} g_k(x_i, \hat{\theta})1(z_i \leqslant z), \tag{13}$$

for $k = 1, \ldots, r$, where $1(A)$ denotes the indicator function of an event A.

The pseudo-residual process in (13) is closely related to residual processes in some specific parametric and semiparametric models including generalized estimating equations for longitudinal data (Stute, 1997; Lin et al., 2002). For instance, consider the linear regression model $y_i = x_{i,(2)}^{\mathrm{T}}\beta + \sigma\epsilon_i$ and $g(x_i, \hat{\theta}) = x_{i,(2)}(y_i - x_{i,(2)}^{\mathrm{T}}\hat{\beta})$. Let $x_{i,k(2)}$ be the kth component of $x_{i,(2)}$. If we choose z_i to be the mth component of $x_{i,(2)}$, then

$$\mathrm{GF}_k(z;\hat{\theta}) = n^{-1/2}\sum_{i=1}^{n} x_{i,k(2)}\big(y_i - x_{i,(2)}^{\mathrm{T}}\hat{\beta}\big)1(x_{i,m(2)} \leqslant z).$$

Moreover, when $x_{i,1(2)} = 1$ for $i = 1, \ldots, n$, $\mathrm{GF}_1(z;\hat{\theta})$ reduces to the residual process $n^{-1/2}\sum_{i=1}^{n}(y_i - x_{i,(2)}^{\mathrm{T}}\hat{\beta})1(z_i \leqslant z)$ for the linear regression model (Stute, 1997). Similarly, following the arguments in Lin et al. (2002), we may choose z_i as $x_{i,(2)}^{\mathrm{T}}\hat{\beta}$. As shown below, for each k, $\mathrm{GF}_k(z;\hat{\theta})$ converges weakly to a centred Gaussian process $\mathrm{GF}_k(z)$ under H_0 (van der Vaart & Wellner, 1996). Based on this result, we can construct Kolmogorov–Smirnov test statistics given by

$$\mathrm{KS}_{n,k} = \sup_z |\mathrm{GF}_k(z;\hat{\theta})|, \quad \text{for } k = 1, \ldots, r. \tag{14}$$

We establish the weak convergence of $\mathrm{GF}_k(z;\hat{\theta})$ and $\mathrm{KS}_{n,k}$ under the null hypothesis H_0 as follows.

THEOREM 4. *If Assumptions* A1–A3 *and* A5 *in the Appendix and the null hypothesis* H_0 *in* (12) *are true, then* $\mathrm{GF}_k(z;\hat{\theta})$ *converges weakly to* $\mathrm{GF}_k(z)$ *in the Skorokhod space* $D[-\infty, +\infty]$; *the Skorokhod space is the set of all right-continuous real functions on* $[-\infty, \infty]$ *having limits on the left. Also,* $\mathrm{KS}_{n,k}$ *converges weakly to* $\sup_z |\mathrm{GF}_k(z)|$, *where* $\mathrm{GF}_k(z)$ *is a Gaussian process with zero mean and covariance function*

$$\Sigma_k(z_1, z_2) = E_F[\{g_k(x)1(Z \leqslant z_1) - D_k(z_1)g(x)\}\{g_k(x)1(Z \leqslant z_2) - D_k(z_2)g(x)\}] \tag{15}$$

for any z_1 *and* z_2, $k = 1, \ldots, r$, *where* $g_k(x) = g_k(x, \theta_0)$, $g(x) = g(x, \theta_0)$ *and* $D_k(z) = E_F\{1(Z \leqslant z)\partial_\theta g_k(x)^{\mathrm{T}}\}S_{22\cdot1}^{-1}S_{21}S_{11}^{-1}$.

Theorem 4 formally characterizes the asymptotic distributions of the stochastic processes of interest $\{\mathrm{GF}_k(z;\hat{\theta}) : k = 1, \ldots, r\}$, which form the foundation for using $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$ as test statistics.

Since the null hypothesis H_0 states that all estimating equations in (11) are correctly specified, we need to combine the $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$ and test whether or not any of the $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$ show any patterns beyond random fluctuation. However, because the variances of $\{g_k(x, \hat{\theta}) : k = 1, \ldots, p\}$ may be quite different, we must adjust such differences between the variances of $g_k(x, \hat{\theta})$ before we combine $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$. Thus, based on $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$, we construct another maximum statistic as follows:

$$\mathrm{KS}_n = \max_{1 \leqslant k \leqslant r}\big[\mathrm{KS}_{n,k}\{(S_{n,11})_{k,k}\}^{-1/2}\big], \tag{16}$$

where $(S_{n,11})_{k,k}$ is the (k, k)th element of $S_{n,11}(\hat{t}, \hat{\theta}) = \partial_t^2 Q_n(\hat{t}, \hat{\theta})$. The continuous mapping theorem yields that KS_n converges weakly to $\max_{1 \leqslant k \leqslant r}[\sup_z |\mathrm{GF}_k(z)|\{(S_{11})_{k,k}\}^{-1/2}]$, where $(S_{11})_{k,k}$ is the (k, k)th element of S_{11}. Note that $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$ and KS_n are applicable to the cases with $r \geqslant p$, whereas the empirical likelihood ratio statistic W_1 is limited to the cases with $r > p$.

Following the arguments in Theorem 4, we can establish the power of the statistic KS_n, when model (1) is misspecified.

COROLLARY 1. *If model* (1) *is misspecified and Assumptions* A4 *and* A5 *in the Appendix are true, then, for* $k = 1, \ldots, r$, $\mathrm{GF}_k(z; \hat{\theta}) = \sqrt{n} E_F\{g_k(x, \theta_*) 1(Z \leqslant z)\} + O_p(1)$ *and* KS_n *converges to* ∞ *in probability.*

4·3. *A resampling method*

Two complications exist in applying the asymptotic results in Theorem 4 for testing the hypotheses in (12). First, the limiting distributions $\{\sup_z |\mathrm{GF}_k(z)| : k = 1, \ldots, r\}$ have complicated analytical forms. Thus, we cannot use these limiting distributions directly to calculate the critical value of the test statistics $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$. Secondly, because we test r hypotheses simultaneously, we need to correct properly for multiple comparisons in order to control the familywise error rate. In particular, we need to account for the correlations among all the $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$.

In the following, we devise a resampling method for approximating the p-values of $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$. In particular, in Step 2, we correct for multiple comparisons by accounting for the correlations among the $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$.

Step 1. Generate independent and identically distributed random samples, $\{v_i^{(q)} : i = 1, \ldots, n\}$, from $N(0, 1)$ for $q = 1, \ldots, Q$, where Q is the number of replications, $Q = 1000$, say.

Step 2. Calculate

$$\mathrm{GF}_k(z; \hat{\theta})^{(q)} = n^{-1/2} \sum_{i=1}^{n} v_i^{(q)} \{g_k(x_i, \hat{\theta}) 1(z_i \leqslant z) - \hat{D}_k(z) g(x_i, \hat{\theta})\}$$

for $k = 1, \ldots, r$, where $\hat{D}_k(z) = \{n^{-1} \sum_{i=1}^{n} \partial_\theta g_k(x_i, \hat{\theta})^{\mathrm{T}} 1(z_i \leqslant z)\} S_{n,22\cdot1}^{-1} S_{n,21} S_{n,11}^{-1}$. Note that, conditional on the observed data, $\mathrm{GF}_k(z; \hat{\theta})^{(q)}$ converges weakly to the desired Gaussian process in Theorem 4 as $n \to \infty$ (Kosorok, 2003; van der Vaart & Wellner, 1996).

Step 3. Calculate the test statistics $\mathrm{KS}_{n,k}^{(q)} = \sup_z |\mathrm{GF}_k(z; \hat{\theta})^{(q)}|$ for $k = 1, \ldots, r$ and obtain $\{\mathrm{KS}_{n,k}^{(q)} : q = 1, \ldots, Q; k = 1, \ldots, r\}$.

Step 4. Calculate the p-value of $\mathrm{KS}_{n,k}$ using $\{\mathrm{KS}_{n,k}^{(q)} : q = 1, \ldots, Q\}$ for each k.

Step 5. Calculate $\mathrm{KS}_n = \max_{1 \leqslant k \leqslant r} [\mathrm{KS}_{n,k} \{(S_{n,11})_{k,k}\}^{-1/2}]$ and

$$\mathrm{KS}_n^{(q)} = \max_{1 \leqslant k \leqslant r} \left[\mathrm{KS}_{n,k}^{(q)} \{(S_{n,11})_{k,k}\}^{-1/2}\right] \text{ for } q = 1, \ldots, Q.$$

Step 6. Finally, we compute the p-value of KS_n using $\{\mathrm{KS}_n^{(q)} : q = 1, \ldots, Q\}$; that is,

$$p = Q^{-1} \sum_{q=1}^{Q} 1\left(\mathrm{KS}_n^{(q)} > \mathrm{KS}_n\right). \tag{17}$$

5. SIMULATION STUDIES AND REAL DATA EXAMPLES

5·1. *Preamble*

We conducted Monte Carlo simulations and real data analyses to examine the finite-sample performance of the diagnostic measures and their associated goodness-of-fit statistics. First, we applied all diagnostic measures to an artificial dataset, in which an 'outlier' was added. We expected that most of the diagnostic measures would detect the 'outlier'. Secondly, we evaluated the rates of the Type I and Type II errors for $\mathcal{L}_n$ and KS_n and compared them with W_1 when used as goodness-of-fit statistics.

In all empirical work, we used the following procedures. We first used the modified conjugate gradient method to calculate the maximum empirical likelihood estimator $\hat{\theta}$. We approximated the p-values of $\mathcal{L}_n$ and W_1 based on 799 bootstrap samples, which led to their asymptotic refinements (Owen, 2001). Moreover, we used $Q = 999$ replications in the resampling procedure to calculate the p-values of KS_n and $\{\mathrm{KS}_{n,k} : k = 1, \ldots, r\}$.

5·2. *Simulation studies*

Experiment 1. Following the simulation study in Chen & Cui (2003), we consider the following generalized linear model:

$$y_i = \mu(x_{i,(2)}^{\mathrm{T}}\beta) + \sigma V\{\mu(x_{i,(2)}^{\mathrm{T}}\beta)\}^{1/2}\varepsilon_i, \tag{18}$$

for $i = 1, \ldots, n$, where $\mu(t) = \exp(t)$, $V(t) = t^2$, $x_i = (y_i, x_{i,(2)}^{\mathrm{T}})^{\mathrm{T}}$ and $\varepsilon_i \sim N(0, 1)$ is independent of $x_{i,(2)}$. The $x_{i,(2)}^{\mathrm{T}} = (x_{i,1(2)}, x_{i,2(2)})$ were simulated from an $N(0, I_2)$ distribution and the true value of $\theta^{\mathrm{T}} = (\beta_1, \beta_2, \sigma^2)^{\mathrm{T}}$ was set at $(1{\cdot}0, 1{\cdot}0, 0{\cdot}5)^{\mathrm{T}}$. We set $n = 200$ and changed y_{200} into $y_{200} + 5{\cdot}0$ in order to add an 'outlier'.

We applied the diagnostic measures to detect y_{200} as an influential observation by using the following process. As in Chen & Cui (2003), we considered the estimating functions

$$g(x_i, \theta) = \begin{pmatrix} e(x_{i,(2)}, \beta)\partial_\beta \mu(x_{i,(2)}^{\mathrm{T}}\beta)/V\{\mu(x_{i,(2)}^{\mathrm{T}}\beta)\} \\ e(x_{i,(2)}, \beta)^2/V\{\mu(x_{i,(2)}^{\mathrm{T}}\beta)\} - \sigma^2 \\ e(x_{i,(2)}, \beta)^2 w(x_{i,(2)}, \beta)/V\{\mu(x_{i,(2)}^{\mathrm{T}}\beta)\} - \sigma^2 w(x_{i,(2)}, \beta) \end{pmatrix}, \tag{19}$$

where $e(x_{i,(2)}, \beta) = y_i - \mu(x_{i,(2)}^{\mathrm{T}}\beta)$ and $w(x_{i,(2)}, \beta) = \partial_\beta V\{\mu(x_{i,(2)}^{\mathrm{T}}\beta)\}/V\{\mu(x_{i,(2)}^{\mathrm{T}}\beta)\}$. The maximum empirical likelihood estimator of θ was calculated as $\hat{\theta} = (0{\cdot}952, 0{\cdot}997, 0{\cdot}641)^{\mathrm{T}}$. Figure 1 shows that the 200th observation was classified as the most influential observation by C_{e_i}, ECD_i, $R_{i,3(c)}$ and $R_{i,4(c)}$, but not by v_1, a plot not presented here.

Experiment 2. We assessed the performance of $\mathcal{L}_n$, KS_n and $\{\mathrm{KS}_{n,k}\}_1^r$ as goodness-of-fit statistics by evaluating rates of Type I and II errors associated with each of these statistics. We simulated data from a linear regression model $y_i = x_{i,(2)}\theta + \varepsilon_i$, for $i = 1, \ldots, n$, where $x_i = (y_i, x_{i,(2)})$, $\theta_0 = 1{\cdot}0$, and $x_{i,(2)}$ and ε_i were independently generated from an $N(0, 1)$ distribution. Since $E_F(y_i - x_{i,(2)}\theta) = 0$ and $E_F\{(y_i - x_{i,(2)}\theta)^2 - 1{\cdot}0\} = 0$, we can infer θ by using the estimating functions

$$g(x_i, \theta) = \begin{pmatrix} y_i - x_{i,(2)}\theta + c \\ (y_i - x_{i,(2)}\theta)^2 - 1{\cdot}0 \end{pmatrix},$$

where c is a fixed scalar. In particular, $E_F\{g(x, \theta)\} = 0$ when $c = 0$. We used $n = 100$ and $n = 50$ to obtain simulated datasets. For each simulated dataset, the significance level was set at 0·05. We chose $z_i = x_{i,(2)}$ in calculating $\{\mathrm{KS}_{n,k} : k = 1, 2\}$ and KS_n, and we estimated the rejection rates of the five goodness-of-fit statistics including $\mathcal{L}_n$, W_1, $\{\mathrm{KS}_{n,k} : k = 1, 2\}$ and KS_n, using 500 replications.

Table 1 presents estimates for the rejection rates for the five goodness-of-fit statistics at the 5% significance level. We observed that, except for $\mathrm{KS}_{n,1}$, the Type I errors for the other four statistics were not excessive. Consistent with our expectations, the power for detecting misspecification

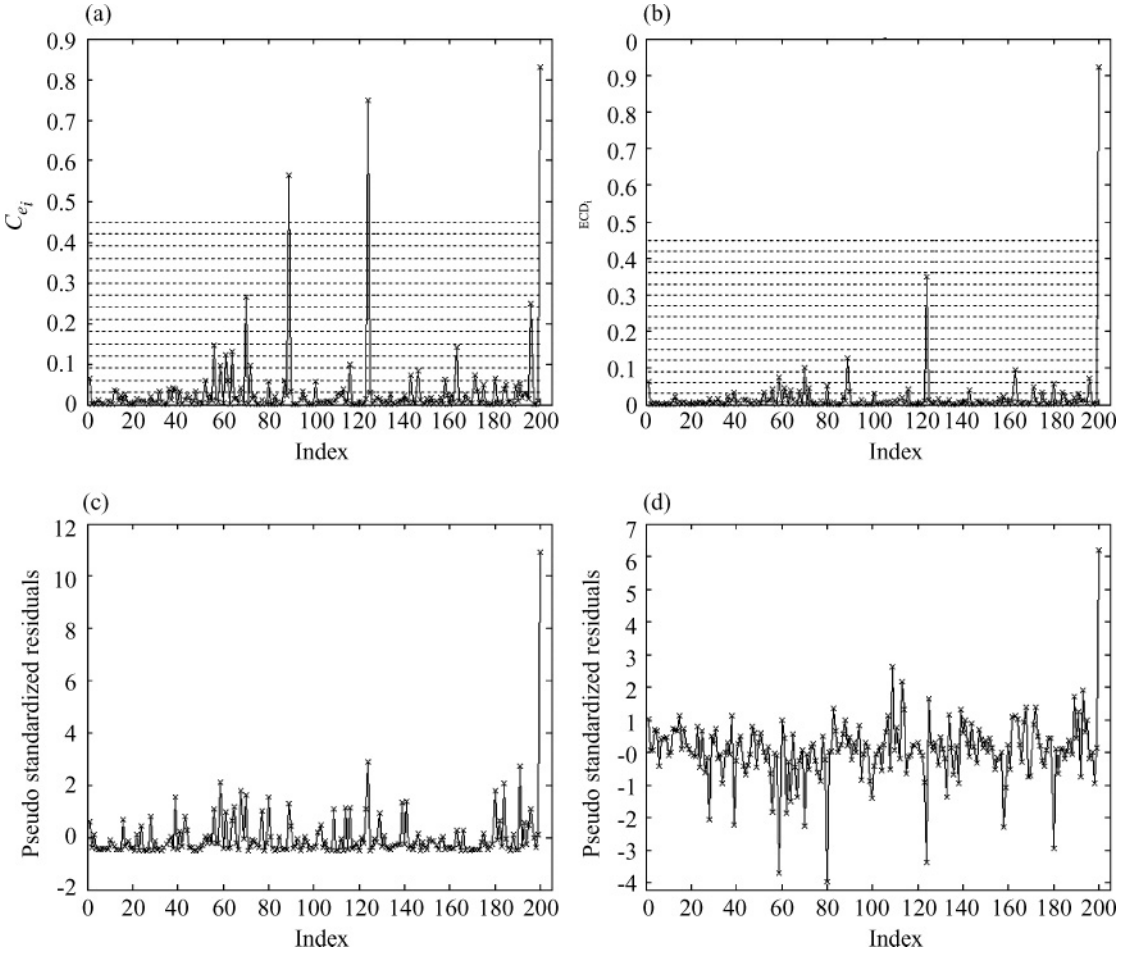

Fig. 1. Results from a simulated dataset. Index plots of diagnostic measures: (a) C_{e_i}; (b) ECD_i; (c) $R_{i,3}$; (d) $R_{i,4}$.

Table 1. *Comparison of the rejection rates for the goodness-of-fit statistics at the* 0·05 *significance level*

	$n = 100$					$n = 50$				
	Bootstrap		Resampling			Bootstrap		Resampling		
c	$\mathcal{L}_n$	W_1	$\mathrm{KS}_{n,1}$	$\mathrm{KS}_{n,2}$	KS_n	$\mathcal{L}_n$	W_1	$\mathrm{KS}_{n,1}$	$\mathrm{KS}_{n,2}$	KS_n
0·0	0·05	0·04	0·01	0·06	0·05	0·03	0·04	0·01	0·04	0·03
0·2	0·08	0·19	0·32	0·13	0·23	0·09	0·08	0·13	0·08	0·11
0·4	0·32	0·70	0·74	0·23	0·68	0·23	0·53	0·41	0·21	0·37
0·5	0·45	0·82	0·85	0·27	0·81	0·36	0·79	0·58	0·25	0·51
0·6	0·60	0·73	0·90	0·34	0·88	0·41	0·84	0·67	0·31	0·60
0·7	0·67	0·76	0·89	0·42	0·87	0·51	0·82	0·77	0·38	0·70
0·8	0·67	0·71	0·91	0·44	0·90	0·47	0·87	0·86	0·40	0·83
1·0	0·61	0·61	0·98	0·66	0·98	0·45	0·78	0·90	0·57	0·90

of the estimating equations increased with the scalar c and the sample size n. However, we also observed peculiar behaviour with W_1 and $\mathcal{L}_n$, because their power for rejecting estimating equations increased with c at the beginning and then decreased. The cause of this behaviour warrants further investigation. Finally, W_1 outperformed $\mathcal{L}_n$ and KS_n outperformed W_1 at both $n = 50$ and $n = 100$. Moreover, compared with the resampling method in § 4·3, the bootstrap method is much more computationally intensive.

5·3. *Steam data*

We considered a dataset from Draper & Smith (1981, p. 205) consisting of 25 observations. Each observation includes the pounds of steam used monthly (y_i), the operating days per month ($x_{i,1(2)}$), and the average atmospheric temperature ($x_{i,2(2)}$). Following Draper & Smith (1981), we considered a linear regression model $y_i = x_{i,(2)}^{\mathrm{T}}\beta + \sigma\varepsilon_i$, where $x_{i,(2)} = (1, x_{i,1(2)}, x_{i,2(2)}, x_{i,2(2)}^2)^{\mathrm{T}}$, $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^{\mathrm{T}}$, and $\varepsilon_i \sim N(0, 1)$. Let $\ell_{n,i}(\theta)$ be the loglikelihood function for the ith observation for the linear regression model. We considered 10 estimating functions

$$g(x_i, \theta) = \begin{pmatrix} (y_i - x_{i,(2)}^{\mathrm{T}}\beta)x_{i,(2)}/\sigma^2 \\ \{1{\cdot}0 - (y_i - x_{i,(2)}^{\mathrm{T}}\beta)^2/\sigma^2\}/2 \\ \{3 - (y_i - x_{i,(2)}^{\mathrm{T}}\beta)^2/\sigma^2\}(y_i - x_{i,(2)}^{\mathrm{T}}\beta)x_{i,(2)}/(2\sigma^2) \\ \{(y_i - x_{i,(2)}^{\mathrm{T}}\beta)^4/\sigma^4 - 4(y_i - x_{i,(2)}^{\mathrm{T}}\beta)^2/\sigma^2 + 1\}/4 \end{pmatrix},$$

where $\theta = (\beta^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$ and $x_i = (y_i, x_{i,(2)}^{\mathrm{T}})^{\mathrm{T}}$. In $g(x_i, \theta)$, the first 5 estimating functions were based on $\partial_\theta \ell_{n,i}(\theta)$, whereas the last 5 were related to the last row of $\partial_\theta^2 \ell_{n,i}(\theta) + \{\partial_\theta \ell_{n,i}(\theta)\}^{\otimes 2}$. We calculated the maximum empirical likelihood estimate of θ as

$$\hat{\theta} = (11{\cdot}982, 0{\cdot}229, -0{\cdot}201, 0{\cdot}001, 0{\cdot}342)^{\mathrm{T}}.$$

Moreover, we chose $z_i = x_{i,(2)}^{\mathrm{T}}\hat{\beta}$.

We applied the diagnostic tools as follows. The standardized residuals and local influence measures indicated that Case 6 was the most influential observation; see Fig. 2(a). We approximated the p-values of the statistics $\mathcal{L}_n$ and W_1 as 0·044 and 0·014, respectively. Using $z_i = x_{i,(2)}^{\mathrm{T}}\hat{\beta}$, we calculated the p-values of $\{\mathrm{KS}_{n,k}\}_1^{10}$ as 0·044, 0·058, 0·010, 0·007, 0·371, 0·158, 0·188, 0·079, 0·044 and 0·296, respectively, and the p-value of KS_n was 0·022; see Fig. 2(c) and (d). Thus, these p-values of $\mathcal{L}_n$, W_1 and KS_n suggest that the steam data might not follow the posited linear relationship between y_i and $x_{i,(2)}$ at nominal level $\alpha = 5\%$. These findings are consistent with previous results (Zhu & Zhang, 2004). However, after deletion of the 6th case, the p-values of the three goodness-of-fit statistics are greater than 0·10, which indicates that these statistics are sensitive to influential observations and outliers.

Furthermore, we calculated two-case case-deletion measures $\mathrm{ECD}_{i,j}$ and $\mathrm{ELD}_{i,j}$. The $\mathrm{ELD}_{i,j}$ indicated that both (4, 6) and (4, 10) were the most influential pairs; see Fig. 3(b). Similar findings were obtained using $\mathrm{ECD}_{i,j}$, not presented here. The inspection of the scatter-plot of $(x_{i,1(2)}, x_{i,2(2)}, y_i)$ revealed that two points, 4 and 10, with small $x_{i,1(2)}$ values were far away from other points in the predictor space; see Fig. 3(a). This result indicated that our diagnostic measure based on single-case deletion can suffer from the well-known masking effect (Lawrance, 1995).

6. DISCUSSION

Many issues still merit further research. One major issue is to generalize the diagnostic measures in § 3 from independently and identically distributed data to weakly correlated data (Kitamura, 1997), longitudinal data (Lin et al., 2002) and survival data (Ibrahim et al., 2001). Another major issue is to develop diagnostic measures for assessing the effects of deleting individual observations on the empirical likelihood confidence region for θ (Lazar, 2005; Owen, 2001). Moreover, it is of interest to generalize our results for the maximum empirical likelihood estimator in § 2 to generalized empirical likelihood estimators (Imbens, 2002).

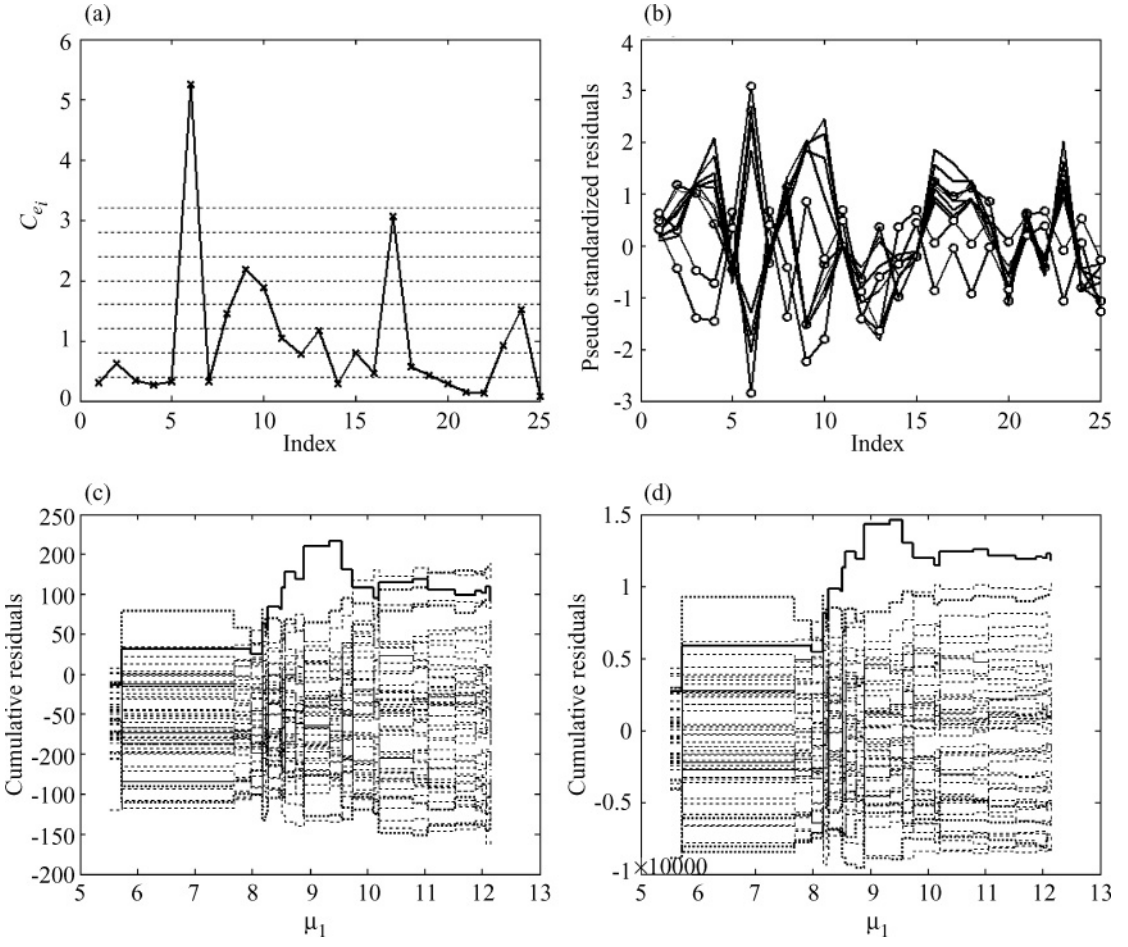

Fig. 2. The steam data: (a) C_{e_i}; (b) $\{R_{i,k}\}_1^{10}$; (c) $\mathrm{KS}_{n,3}$; (d) $\mathrm{KS}_{n,4}$.

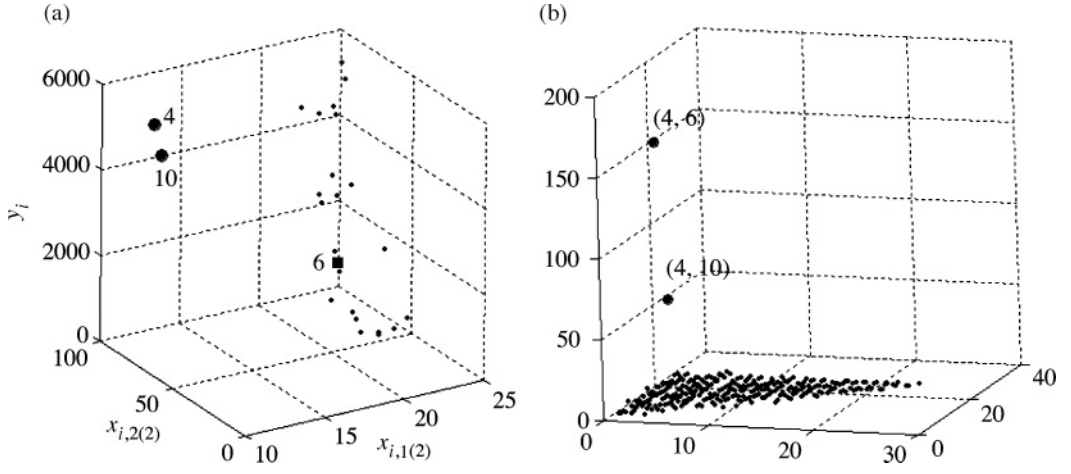

Fig. 3. The steam data: (a) scatterplot of $(x_{i,1(2)}, x_{i,2(2)}, y_i)$; (b) $\mathrm{ELD}_{i,j}$.

ACKNOWLEDGEMENT

This work was supported in part by the U.S. National Science Foundation, the U.S. National Institutes of Health and the National Science Foundation of China. We thank Professor D. M. Titterington, the associate editor and an anonymous referee for valuable suggestions, which greatly helped to improve our presentation.

APPENDIX

Technical details

We define

$$S_n(t,\theta)=\begin{pmatrix}\partial_t Q_{1,n} & (\partial_\theta Q_{1,n})^{\mathrm{T}}\\ (\partial_t Q_{2,n})^{\mathrm{T}} & \partial_\theta Q_{2,n}\end{pmatrix},\qquad S_{n[i]}(t,\theta)=\begin{pmatrix}\partial_t Q_{1,n[i]} & \big(\partial_\theta Q_{1,n[i]}\big)^{\mathrm{T}}\\ \big(\partial_t Q_{2,n[i]}\big)^{\mathrm{T}} & \partial_\theta Q_{2,n[i]}\end{pmatrix},$$

in which all elements are evaluated at (t,θ). We use $\|\cdot\|$ to denote the Euclidean norm of a vector or a matrix.

We make the following assumptions.

Assumption A1. The true value θ_0 of θ is an interior point of $\Theta\subset R^p$. The point 0 is inside the convex hull of the points $g(x_1,\theta),\ldots,g(x_n,\theta)$.

Assumption A2. In a neighbourhood of the true value θ_0, $g(x,\theta)$ has a second-order continuous derivative with respect to θ and $||\partial_\theta g(x,\theta)||$, $||\partial^2_\theta g(x,\theta)||$ and $||g(x,\theta)||^3$ are bounded by some integrable function $g_0(x)$ with $E_F\{g_0(x)\}<\infty$.

Assumption A3. The rank of $E_F\{\partial_\theta g(x,\theta_0)\}$ is p and $E_F\{g(x,\theta_0)g(x,\theta_0)^{\mathrm{T}}\}$ is positive definite.

Assumption A4. (i) As $n\to\infty$, $(\hat t,\hat\theta)$ and $(\hat t_{[i]},\hat\theta_{[i]})$ converge to (t_*,θ_*) in probability.

(ii) The function $\log\{1+t^{\mathrm{T}}g(x,\theta)\}$ is twice continuously differentiable in a neighbourhood of (t_*,θ_*), denoted by $\mathcal{N}$, and

$$\int \sup_{\mathcal{N}}[\|\log\{1+t^{\mathrm{T}}g(x,\theta)\}\|+\|\partial_\theta\log\{1+t^{\mathrm{T}}g(x,\theta)\}\|^2 + \|\partial^2_\theta\log\{1+t^{\mathrm{T}}g(x,\theta)\}\|]dF(x)<\infty.$$

(iii) The matrix $-E_F[\partial^2_\eta\log\{1+t^{\mathrm{T}}g(x,\theta)\}]$ at (t_*,θ_*) is positive definite, where $\eta=(t^{\mathrm{T}},\theta^{\mathrm{T}})^{\mathrm{T}}$.

Assumption A5. The $z_1,\ldots,z_n$ are independent observations from an unknown marginal distribution of F with respect to z, and the distribution of z_i is continuous.

Assumptions A1–A3 have been used to examine the asymptotic properties of the maximum empirical likelihood estimator (Qin & Lawless, 1994; Owen, 2001). Some sufficient conditions for Assumption A.4(i) can be found in Chen et al. (2007). Assumptions A4(ii) and (iii) are standard conditions ensuring a Taylor expansion of $Q_n(t,\theta)$ at (t_*,θ_*).

Proof of Proposition 1. Using Theorem 1 in Qin & Lawless (1994), we can obtain $\hat\theta-\theta_0=O_p(n^{-1/2})$, $\hat t=O_p(n^{-1/2})$, $\hat\theta_{[i]}-\theta_0=O_p(n^{-1/2})$ and $\hat t_{[i]}=O_p(n^{-1/2})$. Thus, $\hat\theta-\hat\theta_{[i]}=O_p(n^{-1/2})$ and $\hat t-\hat t_{[i]}=O_p(n^{-1/2})$. Expanding $Q_{1,n[i]}(\hat t_{[i]},\hat\theta_{[i]})$ and $Q_{2,n[i]}(\hat t_{[i]},\hat\theta_{[i]})$ at $(\hat t,\hat\theta)$ yields

$$0=Q_{1,n[i]}(\hat t,\hat\theta)+[\{\partial_\theta Q_{1,n[i]}(\hat t,\hat\theta)\}^{\mathrm{T}}(\hat\theta_{[i]}-\hat\theta)+\partial_t Q_{1,n[i]}(\hat t,\hat\theta)^{\mathrm{T}}(\hat t_{[i]}-\hat t)]\{1+o_p(1)\},$$

$$0=Q_{2,n[i]}(\hat t,\hat\theta)+[\partial_\theta Q_{2,n[i]}(\hat t,\hat\theta)^{\mathrm{T}}(\hat\theta_{[i]}-\hat\theta)+\partial_t Q_{2,n[i]}(\hat t,\hat\theta)^{\mathrm{T}}(\hat t_{[i]}-\hat t)]\{1+o_p(1)\}.$$

Since $\max_{1\leqslant i\leqslant n}|\hat t^{\mathrm{T}}g(x_i,\hat\theta)|=o_p(1)$, $Q_{2,n}(\hat t,\hat\theta)=0$ and $Q_{1,n}(\hat t,\hat\theta)=0$ (Owen, 2001; pp. 218–21), we obtain $Q_{1,n[i]}(\hat t,\hat\theta)=O_p(n^{-1})$ and $Q_{2,n[i]}(\hat t,\hat\theta)=O_p(n^{-1})$. Furthermore, it follows from Assumption A2

that both $S_n(\hat{t},\hat{\theta})$ and $S_{n[i]}(\hat{t},\hat{\theta})$ converge to $S(t_0,\theta_0)=O(1)$ almost surely, as $n\to\infty$. Thus, we obtain $((\hat{t}_{[i]}-\hat{t})^{\mathrm{T}},(\hat{\theta}_{[i]}-\hat{\theta})^{\mathrm{T}})^{\mathrm{T}}=O_p(n^{-1})$ and

$$\begin{pmatrix}\hat{t}_{[i]}-\hat{t}\\ \hat{\theta}_{[i]}-\hat{\theta}\end{pmatrix}=-n^{-1}S^{-1}\begin{pmatrix}g(x_i,\hat{\theta})\\ \partial_\theta g(x_i,\hat{\theta})\hat{t}\end{pmatrix}\{1+o_p(1)\}=-n^{-1}S^{-1}\begin{pmatrix}g(x_i,\theta_0)\\ 0\end{pmatrix}\{1+o_p(1)\}.$$

Thus, Proposition 1(i) and (ii) immediately follow from the definition of S and the explicit form of S^{-1}. □

Proof of Proposition 2. Let $g_k(x_i)=g_k(x_i,\theta_0)$. Expanding $g_k(x_i,\hat{\theta})$ at θ_0 gives

$$g_k(x_i,\hat{\theta})=g_k(x_i)+\partial_\theta g_k(x_i)^{\mathrm{T}}(\hat{\theta}-\theta_0)+0{\cdot}5(\hat{\theta}-\theta_0)^{\mathrm{T}}\{\partial_\theta^2 g_k(x_i)\}(\hat{\theta}-\theta_0)+O_p(n^{-3/2}).$$

As shown in Qin & Lawless (1994), $\hat{\theta}-\theta_0=-n^{-1}S_{22{\cdot}1}^{-1}S_{21}S_{11}^{-1}\sum_{i=1}^n g(x_i)+o_p(n^{-1/2})$. Therefore, we can obtain

$$E_F\{\partial_\theta g_k(x_i)^{\mathrm{T}}(\hat{\theta}-\theta_0)\}\doteq -n^{-1}E_F\left\{\partial_\theta g_k(x_i)^{\mathrm{T}}S_{22{\cdot}1}^{-1}S_{21}S_{11}^{-1}\sum_{i=1}^n g(x_i)\right\},$$

$$E_F\{(\hat{\theta}-\theta_0)^{\mathrm{T}}\partial_\theta^2 g_k(x_i)(\hat{\theta}-\theta_0)\}\doteq -n^{-1}\mathrm{tr}\big[E_F\{\partial_\theta^2 g_k(x_i)\}S_{22{\cdot}1}^{-1}\big],$$

in which we use $S_{22{\cdot}1}=-S_{21}S_{11}^{-1}S_{21}$. Furthermore, to calculate $\hat{\sigma}_k$, the contribution from the term $(\hat{\theta}-\theta_0)^{\mathrm{T}}\{\partial_\theta^2 g_k(x_i)\}(\hat{\theta}-\theta_0)$ has an order higher than $O_p(n^{-1})$, which is negligible. Thus, we obtain

$$g_k(x_i,\hat{\theta})=g_k(x_i)-n^{-1}\partial_\theta g_k(x_i)^{\mathrm{T}}S_{22{\cdot}1}^{-1}S_{21}S_{11}^{-1}g(x_i)-n^{-1}\partial_\theta g_k(x_i)^{\mathrm{T}}S_{22{\cdot}1}^{-1}S_{21}S_{11}^{-1}\sum_{j=1,j\neq i}^n g(x_j).$$

The proof of Proposition 2 can then be completed after some algebraic manipulations. □

Proof of Theorem 1. For C_{e_i}, we first calculate Δ_i as follows:

$$\Delta_i=\partial_\theta l_{E,i}(x_i;\hat{\theta})=-\partial_\theta t_n(\hat{\theta})g(x_i,\hat{\theta})+o_p(1)=S_{21}S_{11}^{-1}g(x_i,\theta_0)+o_p(1).$$

Thus, $C_{e_i}=-2n^{-1}\Delta_i^{\mathrm{T}}S_{22{\cdot}1}^{-1}\Delta_i+o_p(1)$. Furthermore, for ELD_i, expanding $l_E(\hat{\theta}_{[i]})$ at $\hat{\theta}$, we obtain $l_E(\hat{\theta}_{[i]})-l_E(\hat{\theta})=0{\cdot}5(\hat{\theta}_{[i]}-\hat{\theta})^{\mathrm{T}}\{\partial_\theta^2 l_E(\hat{\theta})\}(\hat{\theta}_{[i]}-\hat{\theta})\{1+o_p(1)\}$. Substituting Proposition 1(i) and (ii) into $l_E(\hat{\theta}_{[i]})$ yields that $\mathrm{ELD}_i=-n^{-1}\Delta_i^{\mathrm{T}}S_{22{\cdot}1}^{-1}\Delta_i\{1+o_p(1)\}$. Similarly to the argument for ELD_i, we can derive the asymptotic expansion for ECD_i. It follows from $C_{e_i}=-2n^{-1}\Delta_i^{\mathrm{T}}S_{22{\cdot}1}^{-1}\Delta_i+o_p(1)$ that $\sum_{i=1}^n C_{e_i}=2p+o_p(1)$. Similarly to C_{e_i}, we can prove that the sums of ELD_i and of ECD_i are close to $2p$. □

Proof of Theorem 2. We can follow the reasoning in Propositions 1 and 2 and Theorem 1 to prove Theorem 2(i)–(iii). For instance, similarly to the argument for Proposition 2(i), we can obtain

$$\hat{\eta}_{[i]}-\hat{\eta}=n^{-1}S(\theta_*,t_*)^{-1}\{1+{t_*}^{\mathrm{T}}g(x_i,\theta_*)\}^{-1}\begin{pmatrix}-g(x_i,\theta_*)\\ -\partial_\theta g(x_i,\theta_*)^{\mathrm{T}}t_*\end{pmatrix}\{1+o_p(1)\},$$

where $\hat{\eta}_{[i]}^{\mathrm{T}}=(\hat{t}_{[i]}^{\mathrm{T}},\hat{\theta}_{[i]}^{\mathrm{T}})$. Using Assumption A4(ii), we thus prove Theorem 2(ii). □

Proof of Theorem 3. We can obtain

$$\partial_\theta^2 l_E(\hat{\theta})=\partial_\theta Q_{2,n}-\partial_\theta t_n(\hat{\theta})\partial_t Q_{1,n}\partial_\theta t_n(\hat{\theta})^{\mathrm{T}}=\sum_{i=1}^n\big[\partial_\theta^2\ell_i(\hat{t},\hat{\theta})-\partial_\theta t_n(\hat{\theta})\partial_t^2\ell_i(\hat{t},\hat{\theta})\{\partial_\theta t_n(\hat{\theta})\}^{\mathrm{T}}\big].$$

It follows from the definition of C_{e_i} that

$$L_n(\hat{\theta})=n^{-1/2}2\sum_{i=1}^n\mathrm{tr}\big(\{-n^{-1}\partial_\theta^2 l_E(\hat{\theta})\}^{-1}\big[\Delta_i\Delta_i^{\mathrm{T}}+\partial_\theta^2\ell_i(\hat{t},\hat{\theta})-\partial_\theta t_n(\hat{\theta})\partial_t^2\ell_i(\hat{t},\hat{\theta})\{\partial_\theta t_n(\hat{\theta})\}^{\mathrm{T}}\big]\big).$$

Let $g_i = g(x_i, \hat{\theta})$. Since Δ_i is a function of $\hat{t}$, we expand Δ_i at $t_0 = 0$ to obtain

$$\Delta_i = \partial_\theta t_n(\hat{\theta}) g_i + \partial_\theta g_i \hat{t} - \hat{t}^{\mathrm{T}} g_i \partial_\theta t_n(\hat{\theta}) g_i + O_p(1)||\hat{t}||^2.$$

Similarly to Δ_i, we use a Taylor expansion to obtain

$$\partial_t^2 \ell_i(\hat{t}, \hat{\theta}) = (1 - 2\hat{t}^{\mathrm{T}} g_i) g_i g_i^{\mathrm{T}} + O_p(1)||\hat{t}||^2, \quad \partial_\theta^2 \ell_i(\hat{t}, \hat{\theta}) = -\partial_\theta(\partial_\theta g_i)(I_p \otimes \hat{t}) + O_p(1)||\hat{t}||^2,$$

where $\otimes$ denotes the usual Kronecker product. Combining the above results, we obtain

$$\begin{aligned}\Delta_i \Delta_i^{\mathrm{T}} &+ \partial_\theta^2 \ell_i(\hat{t}, \hat{\theta}) - \partial_\theta t_n(\hat{\theta}) \partial_t^2 \ell_i(\hat{t}, \hat{\theta}) \{\partial_\theta t_n(\hat{\theta})\}^{\mathrm{T}} \\ &= \partial_\theta t_n(\hat{\theta}) g_i \hat{t}^{\mathrm{T}} (\partial_\theta g_i)^{\mathrm{T}} + \partial_\theta g_i \hat{t} g_i^{\mathrm{T}} \{\partial_\theta t_n(\hat{\theta})\}^{\mathrm{T}} - \partial_\theta(\partial_\theta g_i)(I_p \otimes \hat{t}) + O_p(1)||\hat{t}||^2.\end{aligned}$$

Thus,

$$\begin{aligned}L_n(\hat{\theta}) = &\left[n^{-1} \sum_{i=1}^n 4 g_i^{\mathrm{T}} \{\partial_\theta t_n(\hat{\theta})\}^{\mathrm{T}} \{ - n^{-1} \partial_\theta^2 l_E(\hat{\theta})\}^{-1} \partial_\theta g_i \right] n^{1/2} \hat{t} \\ &- 2\mathrm{tr} \left[\{-n^{-1} \partial_\theta^2 l_E(\hat{\theta})\}^{-1} \left\{ n^{-1} \sum_{i=1}^n \partial_\theta(\partial_\theta g_i) \right\} \sqrt{n}(I_p \otimes \hat{t}) \right].\end{aligned}$$

Furthermore, let $A_1 = E_F[\partial_\theta\{\partial_\theta g(x; \theta_0)\}]$ and $A_2 = E_F\{g(x; \theta_0)^{\mathrm{T}} S_{11}^{-1} S_{12} (S_{21} S_{11}^{-1} S_{12})^{-1} \partial_\theta g(x; \theta_0)\}$. We have $n^{-1} \sum_{i=1}^n \partial_\theta(\partial_\theta g_i) = A_1 + o_p(1)$ and

$$n^{-1} \sum_{i=1}^n 4 g_i^{\mathrm{T}} \{\partial_\theta t_n(\hat{\theta})\}^{\mathrm{T}} \{-n^{-1} \partial_\theta^2 l_E(\hat{\theta})\}^{-1} \partial_\theta g_i = -4A_2 + o_p(1).$$

Therefore, $L_n(\hat{\theta}) = -4A_2 n^{1/2} \hat{t} - 2\,\mathrm{tr}\{(S_{21} S_{11}^{-1} S_{12})^{-1} A_1 \sqrt{n}(I_p \otimes \hat{t})\} + o_p(1)$. Since $\sqrt{n}\hat{t} = -A_3 n^{-1/2} \sum_{i=1}^n g(x_i, \theta_0)\{1 + o_p(1)\}$, we obtain $L_n(\hat{\theta}) = n^{-1/2} \sum_{i=1}^n k(x_i; \theta_0) + o_p(1)$, where $A_3 = S_{11}^{-1} + S_{11}^{-1} S_{12} S_{22\cdot 1}^{-1} S_{21} S_{11}^{-1}$ and

$$k(x_i; \theta_0) = 2\mathrm{tr}\left[\left(S_{21} S_{11}^{-1} S_{12}\right)^{-1} A_1 \{I_p \otimes A_3 g(x_i; \theta_0)\} \right] + 4A_2 A_3 g(x_i; \theta_0).$$

It follows from the Lindeberg–Feller theorem that $n^{-1/2} \sum_{i=1}^n k(x_i, \theta_0) \to N(0, \sigma^2)$ in distribution. □

Proof of Theorem 4 *and Corollary* 1. Proofs of Theorem 4 and Corollary 1 follow from the standard empirical process theory (van der Vaart & Wellner, 1996, Ch. 2). The detailed proofs can be found in a supplementary report, which is available upon request. □

REFERENCES

CHEN, S. X. & CUI, H. J. (2003). An extended empirical likelihood for generalized linear models. *Statist. Sinica* **13**, 69–81.

CHEN, X. H., HONG, H. & SHUM, M. (2007). Nonparametric likelihood ratio tests between parametric and moment condition models. *J. Economet.* **141**, 109–40.

COOK, R. D. (1986). Assessment of local influence (with Discussion). *J. R. Statist. Soc.* B **48**, 133–69.

COOK, R. D. & WEISBERG, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

COX, D. R. & SNELL, E. J. (1968). A general definition of residuals (with Discussion). *J. R. Statist. Soc.* B **30**, 248–75.

CRITCHLEY, F. & MARRIOTT, P. (2004). Data-informed influence analysis. *Biometrika* **91**, 125–40.

DAVISON, A. C. & TSAI, C. L. (1992). Regression model diagnostics. *Int. Statist. Rev.* **60**, 337–55.

DICICCIO, T. J. & MONTI, A. C. (2001). Approximations to the profile empirical likelihood function for a scalar parameter in the context of *M*-estimation. *Biometrika* **88**, 337–51.

DRAPER, N. R. & SMITH, H. (1981). *Applied Regression Analysis*, 2nd ed. New York: John Wiley.

HANSEN, L. P. (1982). Large sample properties of generalised method of moments estimators. *Econometrica* **50**, 1029–54.

IBRAHIM, J. G., CHEN, M. H. & SINHA, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.

IMBENS, G. W. (2002). Generalized method of moments and empirical likelihood. *J. Bus. Econ. Statist.* **20**, 493–506.

KITAMURA, Y. (1997). Empirical likelihood methods with weakly dependent processes. *Ann. Statist.* **25**, 2084–102.
KOSOROK, M. R. (2003). Bootstraps of sums of independent but not identically distributed stochastic processes. *J. Mult. Anal.* **84**, 299–318.
LAWRANCE, A. J. (1995). Deletion influence and masking in regression. *J. R. Statist. Soc.* B **57**, 181–9.
LAZAR, N. (2005). Assessing the effect of individual data points on inference from empirical likelihood. *J. Comp. Graph. Statist.* **14**, 626–42.
LIN, D. Y., WEI, L. J. & YING, Z. L. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* **58**, 1–12.
OWEN, A. (2001). *Empirical Likelihood*. New York: Chapman and Hall.
QIN, J. & LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist*. **22**, 300–25.
STUTE, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25**, 613–41.
TSAO, M. & ZHOU, J. (2001). On the robustness of empirical likelihood ratio confidence intervals for location. *Can. J. Statist.* **29**, 129–40.
VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*. New York: Springer.
WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
WOOLDRIDGE, J. M. (1990). A unified approach to robust, regression-based specification tests. *Economet. Theory* **6**, 17–43.
ZHU, H. T. & LEE, S. Y. (2001). Local influence for incomplete data models. *J. R. Statist. Soc.* B **63**, 111–26.
ZHU, H. T., LEE, S. Y., WEI, B. C. & ZHOU, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika* **88**, 727–37.
ZHU, H. T. & ZHANG, H. P. (2004). A diagnostic procedure based on local influence. *Biometrika* **91**, 579–89.

[*Received August* 2006. *Revised August* 2007]