Ann. Inst. Statist. Math.
Vol. 52, No. 2, 343–350 (2000)

ESTIMATING EQUATIONS WITH NUISANCE PARAMETERS: THEORY AND APPLICATIONS*

KE-HAI YUAN[1]** AND ROBERT I. JENNRICH[2]

[1] *Department of Psychology, University of California, Los Angeles, 1282A Franz Hall, Box 951563, Los Angeles, CA 90095-1563, U.S.A.*

[2] *Department of Mathematics, University of California, Los Angeles, CA 90095, U.S.A.*

(Received January 28, 1998; revised November 17, 1998)

Abstract. In a variety of statistical problems the estimate $\hat{\theta}_n$ of a parameter θ is defined as the root of a generalized estimating equation $G_n(\hat{\theta}_n, \hat{\gamma}_n) = 0$ where $\hat{\gamma}_n$ is an estimate of a nuisance parameter γ. We give sufficient conditions for the asymptotic normality of $\hat{\theta}_n$ defined in this way and derive their asymptotic distribution. A circumstance under which the asymptotic distribution of $\hat{\theta}_n$ will not be influenced by that of $\hat{\gamma}_n$ is noted. As an example, we consider a covariance structure analysis in which both the population mean and the population fourth-order moment are nuisance parameters. Applications to pseudo maximum likelihood, generalized least squares with estimated weights, and M-estimation with an estimated scale parameter are discussed briefly.

Key words and phrases: Asymptotic distribution, generalized estimating equation, covariance structure analysis, pseudo maximum likelihood, generalized least squares, equivariant M-estimation.

1. Introduction

Let $G_n(\theta, \gamma)$, $n = 1, 2, \ldots$ be a sequence of p-variate stochastic functions of $\theta \in \Theta \subset R^p$ and $\gamma \in \Gamma \subset R^q$. That is for each θ and γ, $G_n(\theta, \gamma)$ is a random vector. Frequently G_n will have the form

$$G_n(\theta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} g_i(\theta, \gamma), \tag{1.1}$$

where each $g_i(\theta, \gamma)$ is a p-variate random vector. We would like to define an estimate $\hat{\theta}_n$ of θ as a root of $G_n(\hat{\theta}_n, \gamma) = 0$, but we don't know γ which for this purpose we view as a nuisance parameter. Suppose we have an estimate $\hat{\gamma}_n$ of γ and define $\hat{\theta}_n$ as a root of

$$G_n(\hat{\theta}_n, \hat{\gamma}_n) = 0. \tag{1.2}$$

When g_i is the derivative of the logarithm of a density function, Gong and Samaniego (1981) called $\hat{\theta}_n$ a pseudo maximum likelihood estimate of θ. In the context of a generalized linear model for longitudinal data, Liang and Zeger (1986) used an equation of the form (1.2) to estimate mean parameters $\hat{\theta}_n$ given independent estimates $\hat{\gamma}_n$ of some covariance parameters. They called their equation a generalized estimating equation

* This work was supported by National Institute on Drug Abuse Grants DA01070 and DA00017.

** Now at Department of Psychology, University of North Texas, Denton, TX 76203, U.S.A.

(GEE). We will adopt this terminology whenever $\hat{\gamma}_n$ is given and $\hat{\theta}_n$ is defined by an equation of the form (1.2).

Many estimates are GEE estimates: the generalized least squares estimate with estimated weights considered by Carroll *et al.* (1988) and Chapters 2 and 3 of Yuan (1995); the pseudo maximum likelihood estimate considered by Gong and Samaniego (1981) and applied to an elliptical population by Kano *et al.* (1993); the M-estimate with an estimated scale parameter discussed in Bickel (1975) and in Sections 6.5 and 7.7 of Huber (1981); etc. A recent review of applications and historical developments of GEE estimates is given by Liang and Zeger (1995). When γ is a nuisance parameter, Godambe and Thompson (1974) considered how to choose an optimum function g^* that does not depend on γ. Unlike Godambe and Thompson's perspective, we assume that a sequence of functions has already been chosen for estimating θ and g_i involves both θ and γ. We do not restrict ourselves to optimal g_i because as indicated in Yuan and Jennrich (1998) many nonoptimal g_i are of interest.

The γ in $G_n(\theta, \gamma)$ may be a vector of covariance parameters in generalized least squares; may be part of the mean parameters (e.g., the population mean in covariance structure analysis); may be the extra parameters of an elliptical distribution as considered by Kano *et al.* (1993). One may want to minimize an objective function that involves both θ and γ. It may be difficult or for some other reason undesirable to minimize the function simultaneously with respect to θ and γ. If an estimate $\hat{\gamma}_n$ can be obtained easily one may elect to minimize the objective function with respect to θ given $\gamma = \hat{\gamma}_n$ and thereby be led to solving a stationarity equation of the form (1.2) for $\hat{\theta}_n$.

Section 2 discusses the asymptotic distribution of the GEE estimates $\hat{\theta}_n$. Section 3 considers an example from covariance structure analysis in detail. In this example, θ represents the covariance parameters and those of primary interest. The mean vector μ for the population sampled and the weight matrix W used in generalized least squares estimation are treated as nuisance parameters. Other areas of applications are considered briefly in Section 4. These include pseudo maximum likelihood estimation, generalized least squares, and scale equivariant M-estimation.

2. The asymptotic distribution of $\hat{\theta}_n$

For notational convenience let $\delta = (\theta, \gamma)$ and $\mathcal{D} = \Theta \times \Gamma$ be the parameter space of δ. We will use:

ASSUMPTION 1. On $\mathcal{D}$ and with probability one, $\dot{G}_n = dG_n/d\delta$ exists, is continuous, and converges uniformly to a non-stochastic limit J.

This is a uniform strong law of large numbers. It is fairly easy to verify when the stochastic functions g_i are continuously differentiable and identically distributed. If the G_n converge with probability one to a non-stochastic limit G, one can show that $J = \dot{G}$, but we will not use this.

Let J be defined as in Assumption 1 and let A and B be the components of $J(\delta)$ corresponding to θ and γ so $J(\delta) = (A, B)$. Before considering the asymptotic normality of $\hat{\theta}_n$, we need some lemmas.

LEMMA 1. *Let x_n be a sequence of random vectors in $X \subset R^p$ such that $x_n \xrightarrow{P} x_0 \in X$. If f_n is a sequence of continuous stochastic functions from X into R^q such that $f_n \xrightarrow{P} f$ uniformly on a neighborhood of x_0, then $f_n(x_n) \xrightarrow{P} f(x_0)$.*

PROOF. There is a subsequence n_i of n such that $x_{n_i} \to x_0$ a.e. and $\sup_x ||f_{n_i}(x) - f(x)|| \to 0$ a.e. Because the uniform limit of continuous functions is continuous, f is continuous a.e. and

$$f_{n_i}(x_{n_i}) = f_{n_i}(x_{n_i}) - f(x_{n_i}) + f(x_{n_i}) \to f(x_0) \quad \text{a.e.}$$

It follows that $f_n(x_n) \xrightarrow{P} f(x_0)$. See for example Port ((1994), Chapter 40). □

LEMMA 2. *Let x_n be a sequence of random vectors in $X \subset R^p$ such that $x_n \xrightarrow{P} x_0 \in X$. If f_n is a sequence of continuously differentiable stochastic functions from X into R^q such that the Jacobians $\dot{f}_n \xrightarrow{P} \dot{f}$ uniformly on a neighborhood of x_0, then there is a sequence of p by q stochastic matrices C_n such that*

(i) $f_n(x_n) - f_n(x_0) = C_n(x_n - x_0)$,

(ii) $C_n \xrightarrow{P} \dot{f}(x_0)$.

PROOF. It is sufficient to give the proof for scalar valued functions f_n. By the mean value theorem, $f_n(x_n) - f_n(x_0) = \dot{f}_n(\bar{x}_n)(x_n - x_0)$ for some $\bar{x}_n$ on the line from x_0 to x_n. The choice $C_n = \dot{f}_n(\bar{x}_n)$ satisfies (i) and $\bar{x}_n \xrightarrow{P} x_0$. Using Lemma 1, $C_n = \dot{f}_n(\bar{x}_n) \xrightarrow{P} \dot{f}(x_0)$, so C_n satisfies (ii). □

THEOREM 1. *Assume Assumption 1 holds and A is non-singular. If $\hat{\theta}_n \xrightarrow{P} \theta$, $\hat{\gamma}_n$ is $\sqrt{n}$-consistent, and*

$$\sqrt{n}[G_n(\delta) + B(\hat{\gamma}_n - \gamma)] \xrightarrow{\mathcal{L}} N(0, \Pi), \tag{2.1}$$

then $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} N(0, \Omega)$, where $\Omega = A^{-1}\Pi A^{-T}$.

PROOF. Using Lemma 2 we have

$$\sqrt{n}(G_n(\hat{\delta}_n) - G_n(\delta)) = A_n\sqrt{n}(\hat{\theta}_n - \theta) + B_n\sqrt{n}(\hat{\gamma}_n - \gamma),$$

where $A_n \xrightarrow{P} A$ and $B_n \xrightarrow{P} B$. Since $G_n(\hat{\delta}_n) = 0$,

$$\begin{aligned} -A_n\sqrt{n}(\hat{\theta}_n - \theta) =& \sqrt{n}G_n(\delta) + B\sqrt{n}(\hat{\gamma}_n - \gamma) \\ &+ (B_n - B)\sqrt{n}(\hat{\gamma}_n - \gamma) \xrightarrow{\mathcal{L}} N(0, \Pi). \end{aligned} \tag{2.2}$$

The theorem follows from (2.2) and the Slutsky's (1925) theorem. □

Pierce (1982) and Randles (1982) studied the relationship of the asymptotic distribution of $T_n(\hat{\gamma}_n) = T_n(x_1, \ldots, x_n; \hat{\gamma}_n)$ and that of $\hat{\gamma}_n$, where T_n is an explicit function of the x_i and $\hat{\gamma}_n$. Gong and Samaniego (1981) and Parke (1986) studied pseudo maximum likelihood estimation where the estimating equation G_n is the score function. Theorem 1 generalizes the former by choosing $G_n(\theta, \gamma) = \theta - T_n(\gamma)$ and the latter because the estimating function G_n need not be derived from a density function.

Note that (2.1) is satisfied if

$$\sqrt{n}\begin{pmatrix} G_n(\theta, \gamma) \\ \hat{\gamma}_n - \gamma \end{pmatrix} \xrightarrow{\mathcal{L}} N\left(0, \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}\right), \tag{2.3}$$

which is assumed in Gong and Samaniego (1981), Pierce (1982), Randles (1982), and Parke (1986). In such a case,

$$\Pi = V_{11} + BV_{21} + V_{12}B^T + BV_{22}B^T. \tag{2.4}$$

From (2.1) or (2.4) we observe that if $B = 0$, the asymptotic distribution of $\hat{\gamma}_n$ does not influence the asymptotic distribution of $\hat{\theta}_n$ as long as $\hat{\gamma}_n$ is $\sqrt{n}$-consistent for γ, so knowing $\hat{\gamma}_n$ is equivalent to knowing γ. When $B \neq 0$, the asymptotic distribution of $\hat{\gamma}_n$ does influence the asymptotic distribution of $\hat{\theta}_n$, the effect can be seen from the asymptotic variance of $\hat{\theta}_n$. In the context of pseudo maximum likelihood $A = -V_{11}$ and Parke (1986) observed that for the pseudo MLE, $V_{12} = 0$ generally. So

$$\Omega = V_{11}^{-1} + A^{-1}BV_{22}B^TA^{-1}. \tag{2.5}$$

The first term on the right hand side of (2.5) is the inverse of the information matrix corresponding to θ, the second term is nonnegative definite which reflects the cost of not knowing γ. When $V_{12} = 0$ in the context of GEE,

$$\Omega = A^{-1}V_{11}A^{-T} + A^{-1}BV_{22}B^TA^{-T}. \tag{2.6}$$

The second term on the right hand side of (2.6) also reflects the cost of estimating the extra parameter γ in (1.2). For a good estimating function G_n, it usually happens that $A = -V_{11}$ as will be the case in an example in the next section. Since we do not assume any density, there is no formal information matrix in the context of estimation based on GEE. But V_{11} plays the role of an information matrix.

When $\hat{\gamma}_n$ is asymptotically efficient, $V_{12} = 0$ (e.g., Pierce (1982)). Since neither $G_n(\theta, \gamma)$ nor $\hat{\gamma}_n$ need to be efficient in the context of GEE, V_{12} may not be zero. When $V_{12} \neq 0$, it is possible that $BV_{12} + V_{21}B^T + BV_{22}B^T < 0$. Thus $\hat{\theta}_n(\hat{\gamma}_n)$, the estimate based on the estimating equation with an estimated nuisance parameter, can be more efficient than $\hat{\theta}_n(\gamma)$, the estimate based on an estimating equation with knowing the true value γ. Pierce (1982) gave a simple example where $T_n(\hat{\gamma}_n)$ has a smaller asymptotic variance than $T_n(\gamma)$. Our example in the next section shows that a covariance structure parameter estimate using a mean parameter estimate can be more efficient than the one using the true value of the mean parameter.

3. An example

Let $X_1, \ldots, X_n$ be i.i.d. random vectors with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \Sigma(\theta)$. In covariance structure analysis, the interest is in getting a good estimate of θ. Generally, μ is a nuisance parameter. The usual practice (e.g., Bentler (1995)) in covariance structure analysis is to fit $\Sigma(\theta)$ to the sample covariance S by maximum Wishart likelihood assuming $X_i \sim N(\mu, \Sigma(\theta))$. When the X_i are not normal, we may choose some other method, e.g. least squares or generalized least squares. Here we consider the properties of the generalized least squares estimate $\hat{\theta}_n$ by employing the result in Section 2.

For a symmetric matrix A, let $\mathrm{vech}(A)$ be the vector formed by stacking the columns of A leaving out the elements above the diagonals. We denote $Y_i = \mathrm{vech}(X_iX_i^T)$, $\sigma(\theta) = \mathrm{vech}(\Sigma(\theta))$, and $\tau(\mu) = \mathrm{vech}(\mu\mu^T)$. Suppose $\Gamma = \mathrm{Var}(Y_i)$ exists and is nonsingular. Using $E(X_iX_i^T) = \Sigma(\theta) + \mu\mu^T$, we define

$$Q_n(\theta, \mu, W) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sigma(\theta) - \tau(\mu))^TW(Y_i - \sigma(\theta) - \tau(\mu)),$$

where W is any positive definite matrix. For consistent estimates $\hat{\mu}_n$ and $\hat{W}_n$, the estimate $\hat{\theta}_n$ which minimizes $Q_n(\hat{\theta}_n, \hat{\mu}_n, \hat{W}_n)$ is called a generalized least squares estimate of θ. Let $\dot{\sigma}(\theta) = d\sigma/d\theta$ and

$$g_i(\theta, \mu, W) = \dot{\sigma}^T(\theta) W (Y_i - \sigma(\theta) - \tau(\mu)).$$

Then we have

$$G_n(\theta, \mu, W) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta, \mu, W) = \dot{\sigma}^T(\theta) W (\bar{Y} - \sigma(\theta) - \tau(\mu)), \tag{3.1}$$

and it is easily verified that $G_n(\hat{\theta}_n, \hat{\mu}_n, \hat{W}_n) = 0$.

Since μ is a nuisance parameter and we know $\hat{\mu}_n = \bar{X}$ is strongly consistent for μ and satisfies $\sqrt{n}(\bar{X} - \mu) \xrightarrow{\mathcal{L}} N(0, \Sigma(\theta))$, replacing μ in (3.1) by $\bar{X}$, we get $G_n(\theta, \bar{X}, W) = \dot{\sigma}^T(\theta) W (s - \sigma(\theta))$, where $s = \text{vech}(S)$. Since Γ is generally unknown, we need an estimate for $W = \Gamma^{-1}$. Two such estimates for W are the inverse of the sample covariance matrix of the Y_i and the inverse of the matrix formed by the cross products of the fitted residuals considered by Yuan and Bentler (1997). Both of these estimates are $\sqrt{n}$-consistent.

An estimate $\hat{\theta}_n$ which satisfies $G_n(\hat{\theta}_n, \bar{X}, \hat{W}_n) = 0$ is a GEE estimate. To apply our result, we need to check the related assumptions. Assume that $\Sigma(\theta)$ is twice continuously differentiable, then $\dot{G}_n$ exists and is continuous. Since the second-order moment of X_i exists, $\dot{G}_n$ also converges uniformly to a non-stochastic limit on a compact set of δ. It is obvious that $\sqrt{n}[G_n(\delta) + B_\mu(\bar{X} - \mu)] \xrightarrow{\mathcal{L}} N(0, \Pi)$, where B_μ is the submatrix of B corresponding to μ. It is easily verified that B_W, the submatrix of B corresponding to W, is zero. So it follows from Theorem 1 that for either of the weight estimates of W in Yuan and Bentler (1997), $\hat{\theta}_n$ is asymptotically normal. The asymptotic distribution of $\hat{W}_n$ does not influence the asymptotic distribution of $\hat{\theta}_n$ as long as $\hat{W}_n$ is $\sqrt{n}$-consistent for Γ^{-1}. But the asymptotic distribution of $\hat{\mu}_n$ does influence the asymptotic distribution of $\hat{\theta}_n$. Since

$$A = -\dot{\sigma}^T(\theta) W \dot{\sigma}(\theta), \qquad B_\mu = -\dot{\sigma}^T(\theta) W \dot{\tau}(\mu),$$

and

$$V = \begin{pmatrix} \dot{\sigma}^T(\theta) W \dot{\sigma}(\theta) & \dot{\sigma}^T(\theta) W \Delta \\ \Delta^T W \dot{\sigma}(\theta) & \Sigma(\theta) \end{pmatrix},$$

where $\Delta = \text{Cov}(Y_i, X_i)$, using (2.4), we have

$$\Omega = V_{11}^{-1} - A^{-1}\dot{\sigma}^T(\theta) W \{\dot{\tau}(\mu)\Delta^T + \Delta \dot{\tau}^T(\mu) - \dot{\tau}(\mu)\Sigma(\theta)\dot{\tau}^T(\mu)\} W \dot{\sigma}(\theta) A^{-1}.$$

In this example $A = -V_{11}$.

It is interesting to observe that when the X_i obey an elliptical symmetric distribution, $\Delta = \dot{\tau}(\mu)\Sigma(\theta)$ as showed in Yuan and Bentler (1995). This implies that $B_\mu = -V_{12}V_{22}^{-1}$ and Ω simplifies to

$$\Omega = V_{11}^{-1} - A^{-1} B_\mu V_{22} B_\mu^T A^{-1}. \tag{3.2}$$

So $\hat{\theta}_n$ is more efficient than the estimate obtained using the true value μ. Equation (3.2) can be compared with equation (1.3) in Pierce (1982). Here we do not assume that $\hat{\gamma}_n = \bar{X}$ is asymptotically efficient. Indeed $\bar{X}$ is not an efficient estimate of μ when sampled from a multivariate t-distribution.

4. General areas of application

In the last section we demonstrated how to use our results in a specific problem and how to check the relevant assumptions. In this section, we will outline some important areas to which our result can be applied. Since we outline the applications in some general areas, we can not give exact assumptions for these applications. But we will try to check Assumption 1 and give some relevant references whenever possible.

4.1 *Pseudo maximum likelihood*

Let Y_i, $i = 1, \ldots, n$ be independent random vectors with densities $f_i(y_i, \theta, \gamma)$, $i = 1, \ldots, n$. Suppose we have an $\sqrt{n}$-consistent estimate $\hat{\gamma}_n$ for the nuisance parameter γ. Replacing γ by $\hat{\gamma}_n$, and we maximize $\sum_{i=1}^n \ln f_i(y_i, \theta, \hat{\gamma}_n)$ for $\hat{\theta}_n$. Then $\hat{\theta}_n$ is called a pseudo MLE. Corresponding to (1.1),

$$g_i(\theta, \gamma) = \frac{f'_{i\theta}(y_i, \theta, \gamma)}{f_i(y_i, \theta, \gamma)}, \tag{4.1}$$

where $f'_{i\theta}(y_i, \theta, \hat{\gamma}_n) = \partial f_i(y_i, \theta, \hat{\gamma}_n)/\partial\theta$. Assume the g_i in (4.1) satisfy the assumptions of Theorem 1, then $\hat{\theta}_n$ is asymptotically normal. When the Y_i are i.i.d. scalar random variables and both θ and γ are scalar values, Gong and Samaniego (1981) investigated the consistency and asymptotic normality of $\hat{\theta}_n$. They also gave an interesting application of pseudo MLE. When the Y_i are a random sample from an elliptical distribution, Kano *et al.* (1993) considered statistical inference based on pseudo MLE. They discussed several ways of estimating nuisance parameters by some inexpensive methods. Our assumptions are different from those of Gong and Samaniego and Kano *et al.* We only require that Y_i are independent, they can be of different dimensions as is often the case in applications as for example repeated measures.

4.2 *Generalized least squares*

Let $Y_i = \mu_i(\theta) + e_i$, $i = 1, \ldots, n$, where for each i, $\mu_i(\cdot)$ is a vector valued function, $E(e_i) = 0$, and $\mathrm{Var}(e_i) = \Sigma_i(\gamma)$. Several methods of estimating γ by modeling the first two moments of the Y_i were discussed in Chapter 6 of Yuan (1995), all these estimates satisfy $\sqrt{n}$-consistency. When the Y_i are scalar random variables and $\mu_i(\theta)$ are linear in θ, Davidian and Carroll (1987) gave a comprehensive review on how to estimate γ. Let

$$g_i(\theta) = \dot{\mu}_i^T(\theta) W_i(\gamma)(Y_i - \mu_i(\theta))$$

for some weight functions $W_i(\cdot)$. For a $\sqrt{n}$-consistent estimate of γ, the estimate $\hat{\theta}_n$ that satisfies $G_n(\hat{\theta}_n, \hat{\gamma}_n) = 0$ will be called a generalized least squares estimate of θ. Under some specific assumptions on μ_i, W_i, and e_i, the consistency and asymptotic normality of $\hat{\theta}_n$ is rigorously investigated in Chapter 2 of Yuan (1995). Generally $B = 0$ for the generalized least squares estimates. So the asymptotic efficiency of $\hat{\theta}_n$ will be the same for all the $\sqrt{n}$-consistent estimates for γ. When the Y_i are scalar random variables and the $\mu_i(\theta)$ are linear in θ, Carroll *et al.* (1988) investigated the small sample effects of the estimated weights. We suspect that their results can be generalized to the GEE estimates.

4.3 *Scale equivariant M-estimation*

Suppose we have independent random vectors (x_i, y_i) with y_i scalar valued, and we want to fit a model $y_i = x_i^T\theta + e_i$, where $E(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2$. Our interest is in getting a good estimate of θ when possible outliers exist in the data. An equivariant M-estimate $\hat{\theta}_n$ of θ can be defined (Huber (1981), Chapter 7) as

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{y_i - x_i^T\hat{\theta}_n}{\sigma}\right)x_i = 0,$$

when σ is known. We view σ as a nuisance parameter. But in order to get an equivariant estimate for θ, we need an equivariant estimate for σ. Let

$$g_i(\theta, \sigma) = \psi\left(\frac{y_i - x_i^T\theta}{\sigma}\right)x_i.$$

If we have an $\sqrt{n}$-consistent equivariant estimate $\hat{\sigma}_n^2$ of σ^2, then the estimate $\hat{\theta}_n$ which satisfies $G_n(\hat{\theta}_n, \hat{\sigma}_n) = 0$ will be an equivariant GEE estimate. In the simple location case, Huber ((1981), Section 7.7) recommended $\hat{\sigma}_n = \mathrm{MAD}(y_i)$. Bickel (1975) used a similar equivariant $\hat{\sigma}_n$. Yuan (1997) checked the uniform convergence of the G_n with Tukey's biweight ψ. Similar techniques can be used to check for Assumption 1. We generally have $B = 0$ for symmetrically distributed errors, so as in generalized least squares the asymptotic distribution of $\hat{\sigma}_n^2$ will not influence the asymptotic distribution of $\hat{\theta}_n$.

Acknowledgements

We gratefully acknowledge the constructive advice of two referees, which lead to an improved version of the paper.

References

Bentler, P. M. (1995). *EQS Structural Equations Program Manual*, Multivariate Software, Encino, California.

Bickel, P. J. (1975). One step Huber estimates in the linear model, *J. Amer. Statist. Assoc.*, **70**, 428–434.

Carroll, R. J., Wu, C. F. J. and Ruppert, D. (1988). The effect of estimating weights in weighted least squares, *J. Amer. Statist. Assoc.*, **83**, 1045–1054.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation, *J. Amer. Statist. Assoc.*, **82**, 1079–1091.

Godambe, V. P. and Thompson, M. E. (1974). Estimating equations in the presence of a nuisance parameter, *Ann. Statist.*, **2**, 568–571.

Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications, *Ann. Statist.*, **9**, 861–869.

Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.

Kano, Y., Berkane, M. and Bentler, P. M. (1993). Statistical inference based on pseudo-maximum likelihood estimators in elliptical populations, *J. Amer. Statist. Assoc.*, **88**, 135–143.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.

Liang, K. Y. and Zeger, S. L. (1995). Inference based on estimating functions in the presence of nuisance parameters, *Statist. Sci.*, **10**, 158–173.

Parke, W. R. (1986). Pseudo maximum likelihood estimation: the asymptotic distribution, *Ann. Statist.*, **14**, 355–357.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics, *Ann. Statist.*, **10**, 475–478.

Port, S. C. (1994). *Theoretical Probability for Applications*, Wiley, New York.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters, *Ann. Statist.*, **10**, 462–474.

Slutsky, E. (1925). Über stochastische Asymptoten and Grenzwerte, *Metron*, **5**, 1–90.

Yuan, K.-H. (1995). Assymptotics for nonlinear regression models with applications, Ph.D. Thesis, University of California, Los Angeles.

Yuan, K.-H. (1997). A theorem of uniform convergence of stochastic functions with applications, *J. Multivariate Anal.*, **62**, 100–109.

Yuan, K.-H. and Bentler, P. M. (1995). Mean and covariance structure analysis: theoretical and practical improvements, UCLA Statistical Series, No. 194.

Yuan, K.-H. and Bentler, P. M. (1997). Mean and covariance structure analysis: theoretical and practical improvements, *J. Amer. Statist. Assoc.*, **92**, 767–774.

Yuan, K.-H. and Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions, *J. Multivariate Anal.*, **65**, 245–260.