PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 32, Fasc. 1 (2012), pp. 11–24

NORMAL MAXIMUM LIKELIHOOD, WEIGHTED LEAST SQUARES, AND RIDGE REGRESSION ESTIMATES

BY

CHRISTOPHER S. WITHERS (LOWER HUTT)
AND SARALEES NADARAJAH (MANCHESTER)

Abstract. There have been many papers published (in almost every statistics related journal) suggesting that normal maximum likelihood is superior or inferior to weighted least squares and other approaches. In this note, we show that the three main estimation methods (normal maximum likelihood, weighted least squares and ridge regression) all have the same asymptotic covariance and that there is no gain in efficiency among them. We also show how the bias of these estimators can be reduced and conduct a simulation study to illustrate the magnitude of bias reduction.

2000 AMS Mathematics Subject Classification: Primary: 62J05; Secondary: 62F05, 62G20, 62J07.

Key words and phrases: Bias reduction, normal maximum likelihood, regression, weighted least squares.

1. INTRODUCTION

Weighted least squares, normal maximum likelihood and ridge regression are popular methods for fitting generalized linear models among others. See Jiang [8] for a most excellent account.

There have been many studies in the literature comparing the above methods and others. Many papers have claimed that maximum likelihood can be more efficient or less efficient than least squares or weighted least squares. Here, we discuss seven such papers. There are many other papers, appearing in almost every statistics and related journal.

Hausman and Wise [7] claim in their Section 10.3 entitled *Relative efficiencies of weight least squares versus maximum likelihood estimates* that the latter is more efficient. They state that "the gain in efficiency from using maximum likelihood instead of weighted least squares is small" in some cases and "the relative efficiency of maximum likelihood becomes substantial" in other cases. The basic model used by Hausman and Wise [7] is: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is the dependent vector, $\mathbf{X}$ is a design matrix, $\boldsymbol{\beta}$ is a parameter vector, and $\boldsymbol{\epsilon}$ contains independent normal errors.

Based on numerical results, Bonny et al. [4] state: "we introduced a fast maximum likelihood (ML) algorithm that is unbiased and more accurate than a weighted least-squares fit on signal logarithm." The basic model used by Bonny et al. [4] is: $I = M + b$, where I is the "signal in a voxel of a magnitude image," M is a parameter, and b is the noise assumed to follow the Rice–Nakagami distribution.

Mehrotra et al. [9] say: "The maximum likelihood equations are solved iteratively using an EM-like procedure. It is observed that these estimates have smaller mean squared error than [. . .] iterative weighted least-squares estimates." The basic model considered by them is: $\mathbf{Y} = \boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is assumed to have a multivariate normal distribution.

Olsen et al. [10] state: "Simulated identification experiments show that the maximum likelihood method performs better than a weighted least squares method." The basic model considered by them is: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is assumed to contain independent normal errors.

Abdi [1] says: "weighted least squares often performs better" than maximum likelihood estimates. The basic model considered here is: $y = a + bx + \epsilon$ with ϵ assumed to follow no particular distribution.

Emrich and Urfer [6] state again without any justification: "Normal mixtures are applied in interval mapping to model the segregation of genotypes following Mendel's Law in successive generations of crossing. Standard methods use least squares or maximum likelihood estimates. Theoretically, maximum likelihood is known to result in more efficient estimates than least squares. In the interval mapping literature, some authors state that both methods yield equivalent results, whereas other authors emphasize the higher efficiency of maximum likelihood." The basic model used by Emrich and Urfer [6] is: $y = ax + dz + \epsilon$, where y is the dependent variable, (x, z) are independent variables, (a, d) are parameters, and ϵ is assumed to be normally distributed.

Candy et al. [5] say that the maximum likelihood estimate "was superior to the other estimators considered, including that obtained using inverse probability weighted least squares." The basic model considered here is: $y = g(\boldsymbol{\theta}) + \epsilon$, where ϵ is assumed to be normally distributed and $g(\cdot)$ is a non-linear function of some parameters but that can be approximated by a linear function.

We feel that such papers may lead to misunderstandings. The aim of this note is to show:

(i) all three estimators (weighted least squares, normal maximum likelihood and ridge regression) have the same asymptotic covariance matrix;

(ii) (asymptotically) iteration yields no first-order gain in efficiency for these estimators; and

(iii) a correction can be used to reduce bias to $O(n^{-2})$ without changing the asymptotic covariance of the estimators, where n is the sample size.

One could apply Withers [12] to reduce the bias further, for example, to $O(n^{-3})$.

The last of the three aims is the major contribution of this note. The first two are well known to some extent for a variety of models, including the ones considered in this note. However, as explained, there are several papers claiming discrepancies among asymptotic performances of the three estimators. So, we feel that it is important that the first two aims are established even if they are known for some models.

We consider the following model: $\mathbf{Y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i$, $1 \leqslant i \leqslant n$ on $\mathbb{R}^p$, where $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are independent and identically distributed with zero mean and finite moments of all orders but whose distribution is otherwise unknown. The models considered by each of the seven papers (Hausman and Wise [7], Bonny et al. [4], Olsen et al. [10], Abdi [1], Emrich and Urfer [6]) are particular cases of this model. Furthermore, many of the generalized linear models and all of the general linear models can be expressed in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$.

We have considered the model $\mathbf{Y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i$ because it is the simplest and the most traditional one. A future work is to see how the three aims can be considered and/or extended for more general models.

We estimate the unknown $\boldsymbol{\beta}$ in $\mathbb{R}^r$ say, assuming

$$\mathbf{X} \text{ is a known } p \times r \text{ matrix of rank } r \leqslant p. \tag{1.1}$$

Let F denote the distribution of $\mathbf{Y} = \mathbf{Y}_1$. Set

$$\boldsymbol{\mu} = \boldsymbol{\mu}(F) = \mathbb{E}\,(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta},$$

$$\mu_{i_1,\ldots,i_s} = \mathbb{E}[(\mathbf{Y} - \boldsymbol{\mu})_{i_1} \ldots (\mathbf{Y} - \boldsymbol{\mu})_{i_s}] = \mathbb{E}[e_{i_1} \ldots e_{i_s}], \qquad s \geqslant 2,\ 1 \leqslant i_1, \ldots, i_s \leqslant p,$$

$$\mathbf{V} = \mathbf{V}(F) = \operatorname{covar} \mathbf{Y} = \operatorname{covar} \mathbf{e}_1 = (\mu_{ab}).$$

We assume that $\mathbf{V} > \mathbf{0}$ (positive definite). Let $\widehat{F}$ denote the empirical distribution function of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Set

$$\overline{\mathbf{Y}} = n^{-1} \sum_{i=1}^{n} \mathbf{Y}_i, \qquad \widehat{\mathbf{V}} = \mathbf{V}(\widehat{F}) = n^{-1} \sum_{i=1}^{n} (\mathbf{Y}_i - \overline{\mathbf{Y}})(\mathbf{Y}_i - \overline{\mathbf{Y}})'.$$

If the covariance $\mathbf{V}$ were known, then the best linear unbiased estimate of $\boldsymbol{\beta}$ is $g(\overline{\mathbf{Y}}, \mathbf{V})$, where

$$\begin{gathered} g\,(\boldsymbol{\mu}, \mathbf{V}) = \mathbf{D}^{-1}\mathbf{N} = \mathbf{L}\boldsymbol{\mu}, \quad \mathbf{D} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}, \\ \mathbf{N} = \mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\mu}, \quad \mathbf{L} = \mathbf{D}^{-1}\mathbf{X}'\mathbf{V}^{-1}. \end{gathered} \tag{1.2}$$

We consider the class of estimates

$$\widehat{\boldsymbol{\beta}} = g(\overline{\mathbf{Y}},\ \widetilde{\mathbf{V}} + \varepsilon_n \widehat{\mathbf{k}}), \tag{1.3}$$

where $\widetilde{\mathbf{V}}$ is any estimate of $\mathbf{V}$ of the form

$$\widetilde{\mathbf{V}} = w(\overline{\mathbf{Y}},\ \widehat{\mathbf{V}}) \tag{1.4}$$

satisfying

$$\begin{aligned} &w\left(\boldsymbol{\mu},\ \mathbf{V}\right) = \mathbf{V}, \\ &\partial w\left(\boldsymbol{\mu}, \mathbf{V}\right)/\partial \mu_j \equiv \mathbf{0}, \quad \partial\left\{w\left(\boldsymbol{\mu}, \mathbf{V}\right) - \mathbf{V}\right\}/\partial V_{ij} \equiv \mathbf{0}, \quad \varepsilon_n \to 0 \end{aligned} \tag{1.5}$$

as $n \to \infty$, and $\widehat{\mathbf{k}} = \mathbf{k}(\widehat{F}) \geqslant \mathbf{0}$ is any smooth functional in $\mathbb{R}^{p\times p}$. This includes the ordinary weighted least squares estimate ($\widetilde{\mathbf{V}} = \widehat{\mathbf{V}},\ \widehat{\mathbf{k}} = \mathbf{0}$), the ridge regression estimate ($\widetilde{\mathbf{V}} = \widehat{\mathbf{V}}$, $\widehat{\mathbf{k}}$ proportional to $\mathbf{I}_p$), and, we shall show, the normal maximum likelihood estimate, that is, the maximum likelihood estimate computed as if F were $N_p(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$. If $\mathbf{k}(F)$ is invariant to a change in mean, and F is symmetric about $\boldsymbol{\mu}$, then the distribution of $\widehat{\boldsymbol{\beta}}$ is symmetric about $\boldsymbol{\beta}$.

We prove the following:

THEOREM 1.1. *Under* (1.1)–(1.5),

$$\operatorname{covar}(\widehat{\boldsymbol{\beta}}) = n^{-1}\mathbf{D}^{-1} + O(n^{-2}) \tag{1.6}$$

and

$$\mathbb{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = n^{-1}\mathbf{C} + O(n^{-2} + n^{-1}\varepsilon_n) \approx n^{-1}\mathbf{C}, \tag{1.7}$$

where

$$C_i = -L_{ia}\, K_{bc}\, \mu_{abc}, \quad \mathbf{K} = \mathbf{I}_p - \mathbf{YL}. \tag{1.8}$$

Also $\widehat{\boldsymbol{\beta}} - n^{-1}\mathbf{C}(\widehat{F})$ *has bias* $O(n^{-2})$, *where* $\mathbf{C}(F) = \mathbf{C}$.

The $O(\cdot)$ term for matrix or vector equations like (1.6) and (1.7) should be interpreted elementwise. The convention in (1.8) and below is that repeated indices are implicitly summed over their range $1, \ldots, p$. Note that $\mathbf{K}$ has rank $p - r$. This theorem suggests that there is no advantage in going to the trouble of iterating the normal maximum likelihood estimation equations rather than simply using the weighted least squares estimate or a ridge regression estimate if $\det\mathbf{V}$ is near zero. We prove the theorem in Section 2, together with an analogous result for the case $\varepsilon_n \equiv 1$.

If $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are independent and normally distributed, then no iteration is needed to find the maximum likelihood estimates. This is because they are the same as the least squares estimates. Iterations may be needed for non-normal errors.

In Section 3, we show that the normal maximum likelihood estimate satisfies (1.5). Section 4 gives a q-sample extension for the weighted least squares estimate, and shows that the bias can *increase* with q. The proofs use the techniques of Withers [11], [13] based on von Mises derivatives, although no direct use of these derivatives is needed here. Section 5 performs a simulation study to assess the performances $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}} - n^{-1}\mathbf{C}(\widehat{F})$ with respect to bias. A range of distributions is considered for the error terms.

2. MAIN RESULTS

Here, we first prove Theorem 1.1, and then give an analogous result for the case $\varepsilon_n = 1$.

P r o o f o f T h e o r e m 1.1. First suppose $\widehat{\mathbf{k}} \equiv \mathbf{0}$. Let us apply the formulas for the asymptotic covariance $n^{-1}\mathbf{a}_{21}$ and asymptotic bias $n^{-1}\mathbf{C}$ given by Corollary 3.1 of Withers [11] with $s = p$, $k = p + p^2$, $\mathbf{f}(\mathbf{Y}_1) = \{\mathbf{Y}_1,\ \mathbf{Y}_1\mathbf{Y}_1'\} = \mathbf{Z}$, say, $\mathbf{H} = \mathbf{Y}\big(\mathbb{E}(\mathbf{Z})\big) = \mathbf{Y}(F) = \mathbf{Y}\big(\mathbf{W}(F)\big)\boldsymbol{\mu}(F)$, where $\mathbf{L}(\mathbf{V}) = \mathbf{L}$ of (1.2), and $\mathbf{W}(F) = w\big(\boldsymbol{\mu}(F), \mathbf{V}(F)\big)$. Therefore, $T(\widehat{F}) = \widehat{\boldsymbol{\beta}}$. Then, using the convention in Withers [11],

$$\mathbf{a}_{21} = \big[1^2\big] = \mathrm{l}^a \boldsymbol{\nu}^{ab} \mathrm{l}^{b'}, \tag{2.1}$$

$$2\mathbf{C} = [11] = \mathrm{l}^{ab} \boldsymbol{\nu}^{ab}, \tag{2.2}$$

where

$$\boldsymbol{\nu}^{ab} = \mathbb{E}\big[\big(\mathbf{Z} - \mathbb{E}(\mathbf{Z})\big)_a \big(\mathbf{Z} - \mathbb{E}(\mathbf{Z})\big)_b\big],$$

$$\mathrm{l}^a = \partial \mathbf{H} / \partial \big(\mathbb{E}(\mathbf{Z})\big)_a, \quad \mathrm{l}^{ab} = \partial^2 \mathbf{H} / \partial \big(\mathbb{E}(\mathbf{Z})\big)_a \partial \big(\mathbb{E}(\mathbf{Z})\big)_b.$$

Set $\mathbf{M} = \mathbb{E}(\mathbf{Y}\mathbf{Y}') = \boldsymbol{\mu}\boldsymbol{\mu}' + \mathbf{V}$, so $\mathbf{H} = \mathbf{H}(\boldsymbol{\mu}, \mathbf{M})$. Put $\mathbf{H}_{\cdot a} = \partial \mathbf{H} / \partial \mu_a$, $\mathbf{H}_{\cdot bc} = \partial \mathbf{H} / \partial M_{bc}$, and similarly define the partial derivatives $\mathbf{H}_{\cdot a,b}$, $\mathbf{H}_{\cdot a,bc}$, $\mathbf{H}_{\cdot ab,cd}$. Similarly, viewing functions of $\boldsymbol{\mu}$ and $\mathbf{V} = \mathbf{M} - \boldsymbol{\mu}\boldsymbol{\mu}'$ as functions of $\boldsymbol{\mu}$ and $\mathbf{M}$, we can define $\mathbf{L}_{\cdot a}$, $\mathbf{D}_{\cdot a}$, and so on as partial derivatives with respect to the components of $\boldsymbol{\mu}$ and $\mathbf{M}$.

Then, by (2.1) and (2.2), we get

$$\mathbf{a}_{21} = \mathbf{H}_{\cdot a}\, \boldsymbol{\mu}_{ab}\, \mathbf{H}_{\cdot b}' + \overline{\mathbf{O}} + \mathbf{H}_{\cdot ab}\, \boldsymbol{\mu}_{ab,cd}\, \mathbf{H}_{\cdot cd}' = \mathbf{a}_{21}^1 + \mathbf{a}_{21}^2 + \mathbf{a}_{21}^3, \tag{2.3}$$

say, where

$$\mathbf{O} = \mathbf{H}_{\cdot a}\, \boldsymbol{\mu}_{a,bc}\, \mathbf{H}_{\cdot bc}', \quad \boldsymbol{\mu}_{a,bc} = \mathrm{covar}(Y_{1a}, Y_{1b}Y_{1c}),$$

$$\boldsymbol{\mu}_{ab,cd} = \mathrm{covar}(Y_{1a}Y_{1b}, Y_{1c}Y_{1d}),$$

and, for any $p \times p$ $\mathbf{O}$, $\overline{\mathbf{O}} = \mathbf{O} + \mathbf{O}'$. Also

$$2\mathbf{C} = \mathbf{H}_{\cdot a,b}\, \boldsymbol{\mu}_{ab} + 2\mathbf{H}_{\cdot a,bc}\, \boldsymbol{\mu}_{abc} + \mathbf{H}_{\cdot ab,cd}\, \boldsymbol{\mu}_{abcd} = \mathbf{C}_1 + 2\mathbf{C}_2 + \mathbf{C}_3, \tag{2.4}$$

say. Now

$$(\mathbf{I}_p - \mathbf{Y}\mathbf{L})\, \boldsymbol{\mu} = \mathbf{0}, \quad \text{so } \mathbf{K}\boldsymbol{\mu} = \mathbf{0}, \tag{2.5}$$

$\mathbf{L}_{\cdot a} = -\mathbf{L}\,\mathbf{W}_{\cdot a}\,\mathbf{K}$ and $\mathbf{H}_{\cdot a} = \mathbf{L}_{\cdot a}\,\boldsymbol{\mu} + \mathbf{L}\boldsymbol{\mu}_{\cdot a} = \mathbf{L}\boldsymbol{\mu}_{\cdot a}$. Consequently, the first term in (2.3) is $\mathbf{a}_{21}^1 = \mathbf{L}\boldsymbol{\mu}_{\cdot a}\,\boldsymbol{\mu}_{ab}\,\boldsymbol{\mu}_{\cdot b}^{'}\,\mathbf{L}^{'}$. Moreover,

$$(\boldsymbol{\mu}_{\cdot a})_i = \delta_{ia},$$

where

$$\delta_{ia} = \begin{cases} 1 & \text{if } i = a, \\ 0 & \text{if } i \neq a. \end{cases}$$

Thus, $(\boldsymbol{\mu}_{\cdot a}\,\boldsymbol{\mu}_{ab}\,\boldsymbol{\mu}_{\cdot b})_{ij}^{'} = V_{ij}$ and $\mathbf{a}_{21}^1 = \mathbf{L}\mathbf{V}\mathbf{L}^{'} = \mathbf{D}^{-1}$. Similarly, $\mathbf{H}_{\cdot ab} = \mathbf{L}_{\cdot ab}\boldsymbol{\mu} + \mathbf{L}\boldsymbol{\mu}_{\cdot ab} = \mathbf{0}$ since $\mathbf{L}_{\cdot ab} = -\mathbf{L}\mathbf{W}_{\cdot ab}\mathbf{K}$ and $\boldsymbol{\mu}_{\cdot ab} = \mathbf{0}$. So, $\mathbf{a}_{21} = \mathbf{a}_{21}^1 = \mathbf{D}^{-1}$, which proves (1.6).

Also

$$\mathbf{K}_{\cdot a} = -\mathbf{K}^{'}\mathbf{W}_{\cdot a}\mathbf{K},$$

$$\mathbf{L}_{\cdot a,b} = -\mathbf{L}\mathbf{W}_{\cdot a,b}\mathbf{K} + \sum_{a,b} \mathbf{L}\mathbf{W}_{\cdot a}\mathbf{K}^{'}\mathbf{W}_{\cdot b}\mathbf{K},$$

where

$$\sum_{a,b} f(a,b) = f(a,b) + f(b,a),$$

and

$$\mathbf{H}_{\cdot a,b} = \mathbf{L}_{\cdot a,b}\boldsymbol{\mu} + \mathbf{L}_{\cdot a}\,\boldsymbol{\mu}_{\cdot b} + \mathbf{L}_{\cdot b}\boldsymbol{\mu}_{\cdot a} = \sum_{a,b} \mathbf{L}_{\cdot a}\boldsymbol{\mu}_{\cdot b}.$$

Consequently, we have

$$\mathbf{C}_1 = 2\mathbf{L}_{\cdot a}\,\boldsymbol{\mu}_{\cdot b}V_{ab} = -2\mathbf{L}\mathbf{V}_{\cdot a}\,\mathbf{K}\boldsymbol{\mu}_{\cdot b}\,V_{ab} = 2\mathbf{L}\boldsymbol{\mu}\,\boldsymbol{\mu}_{\cdot a}^{'}\mathbf{K}\boldsymbol{\mu}_{\cdot b}\,V_{ab}$$

since $\mathbf{W}_{\cdot a} = \mathbf{V}_{\cdot a} = -\boldsymbol{\mu}_{\cdot a}\boldsymbol{\mu}^{'} - \boldsymbol{\mu}\boldsymbol{\mu}_{\cdot a}^{'}$, and $\mathbf{K}\boldsymbol{\mu} = \mathbf{0}$. Also $\boldsymbol{\mu}_{\cdot a}^{'}\mathbf{K}\boldsymbol{\mu}_{\cdot b}V_{ab} = K_{ab}\,V_{ab}$ $= \mathrm{tr}\mathbf{K}\mathbf{V} = 0$, so $\mathbf{C}_1 = \mathbf{0}$.

Now

$$\mathbf{L}_{\cdot ab,cd} = -\mathbf{L}\mathbf{W}_{\cdot ab,cd}\,\mathbf{K} + \sum_{ab,cd} \mathbf{L}\,\mathbf{W}_{\cdot ab}\mathbf{K}'\,\mathbf{W}_{\cdot cd}\,\mathbf{K},$$

where

$$\sum_{ab,cd} f(a,b,c,d) = f(a,b,c,d) + f(c,d,a,b).$$

Thus, since $\boldsymbol{\mu}_{\cdot ab} = \mathbf{0}$, we have $\mathbf{H}_{\cdot ab,cd} = \mathbf{L}_{\cdot ab,cd}\,\boldsymbol{\mu} = \mathbf{0}$, and so $\mathbf{C}_3 = \mathbf{0}$. Moreover, $\mathbf{L}_{\cdot bc} = -\mathbf{L}\mathbf{W}_{\cdot bc}\mathbf{K}$ and

$$\mathbf{L}_{\cdot a,bc} = -\mathbf{L}\mathbf{W}_{\cdot a,bc}\mathbf{K} + \sum_{a,bc} \mathbf{L}\mathbf{W}_{\cdot a}\mathbf{K}^{'}\mathbf{W}_{\cdot bc}\,\mathbf{K},$$

where

$$\sum_{a,bc} f(a,b,c) = f(a,b,c) + f(c,a,b) + f(b,c,a),$$

and

$$\mathbf{H}_{\cdot a,bc} = \mathbf{L}_{\cdot a,bc}\boldsymbol{\mu} + \mathbf{L}_{\cdot a}\,\boldsymbol{\mu}_{\cdot bc} + \mathbf{L}_{\cdot bc}\boldsymbol{\mu}_{\cdot a} = -\mathbf{L}\mathbf{W}_{\cdot bc}\,\mathbf{K}\boldsymbol{\mu}_{\cdot a}.$$

Moreover, $(\mathbf{W}_{\cdot bc})_{jk} = (\mathbf{V}_{\cdot bc})_{jk} = \delta_{bj}\,\delta_{ck}$, and we get

$$(\mathbf{H}_{\cdot a,bc})_i = -L_{ij}\,\delta_{bj}\,\delta_{ck}\,K_{kc}\,\delta_{la} = -L_{ib}\,K_{ca}.$$

Hence, by (2.4), Theorem 1.1 is proved when $\widehat{\mathbf{k}} = \mathbf{0}$.

For general $\widehat{\mathbf{k}} = \mathbf{k}(\widehat{F})$, we can expand the functional

$$\widetilde{\boldsymbol{\beta}} = T_{(n)}(\widehat{F}) = g\big(\boldsymbol{\mu}(\widehat{F}), \mathbf{V}(\widehat{F}) + \varepsilon_n\,\mathbf{k}(\widehat{F})\big)$$

as

$$T(\widehat{F}) + \sum_{i=1}^{\infty}\,\varepsilon_n^i\,T_i(\widehat{F}),$$

where $T_i(F) = \mathbf{0}$ since $T_{(n)}(F) = \boldsymbol{\beta}$. Also $\mathbb{E}\big(T_i(\widehat{F})\big) = O(n^{-1})$, so (1.7) holds. This completes the proof of Theorem 1.1. ∎

We now cover the case (1.3) with $\varepsilon_n \equiv 1$ and $\widehat{\mathbf{k}}$ of the form $\widehat{\mathbf{k}} = k(\overline{\mathbf{Y}}, \widehat{\mathbf{V}})$. This is equivalent to assuming (1.3) and (1.4) with $\varepsilon_n = 0$.

THEOREM 2.1. *Suppose that* $\widehat{\boldsymbol{\theta}}_W = g(\overline{\mathbf{Y}}, \widehat{\mathbf{W}})$, *where* $\widehat{\mathbf{W}} = w(\overline{\mathbf{Y}}, \widehat{\mathbf{V}})$. *Set*

$$\begin{gathered}\mathbf{W} = w\left(\boldsymbol{\mu}, \mathbf{V}\right), \quad \mathbf{D}_{\mathbf{W}} = \mathbf{X}^{'}\mathbf{W}^{-1}\mathbf{X}, \\ \mathbf{L}_{\mathbf{W}} = \mathbf{D}_{\mathbf{W}}^{-1}\mathbf{X}^{'}\mathbf{W}^{-1}, \quad \mathbf{K}_{\mathbf{W}} = \mathbf{W}^{-1}\left(\mathbf{I}_p - \mathbf{X}\mathbf{L}_{\mathbf{W}}\right).\end{gathered} \tag{2.6}$$

Then $\widehat{\boldsymbol{\beta}}_{\mathbf{W}}$ *has covariance* $n^{-1}\mathbf{L}_{\mathbf{W}}\,\mathbf{V}\mathbf{L}_{\mathbf{W}}^{'} + O(n^{-2})$ *and bias* $n^{-1}\mathbf{C}_{\mathbf{W}} + O(n^{-2})$, *where*

$$\begin{gathered}\mathbf{C}_{\mathbf{W}} = -\mathbf{L}_{\mathbf{W}}\mathbf{W}_{\cdot a}\mathbf{K}_{\mathbf{W}}\,\boldsymbol{\mu}_{\cdot b}\,\boldsymbol{\mu}_{ab} - \mathbf{L}_{\mathbf{W}}\,\mathbf{W}_{\cdot bc}\,\mathbf{K}_{\mathbf{W}}\,\boldsymbol{\mu}_{\cdot a}\,\boldsymbol{\mu}_{abc}, \\ (\boldsymbol{\mu}_{\cdot b})_i = \delta_{ib}, \\ \mathbf{W}_{\cdot b} = \partial w(\boldsymbol{\mu}, \mathbf{M} - \boldsymbol{\mu}\boldsymbol{\mu}^{'})/\partial\mu_b \quad \textit{at } \mathbf{M} = \mathbb{E}(\mathbf{Y}\mathbf{Y}^{'}), \\ \mathbf{W}_{\cdot ab} = \partial w(\boldsymbol{\mu}, \mathbf{V})/\,\partial V_{ab}.\end{gathered}$$

P r o o f. This is similar to that of Theorem 1.1. ∎

3. THE NORMAL MAXIMUM LIKELIHOOD ESTIMATION

The normal maximum likelihood estimates for $\boldsymbol{\beta}$, $\mathbf{V}$ are $\widehat{\boldsymbol{\beta}} = g(\overline{\mathbf{Y}}, \widetilde{\mathbf{V}})$, $\widetilde{\mathbf{V}} = \widehat{\mathbf{V}} + (\overline{\mathbf{Y}} - \widehat{\boldsymbol{\mu}})(\overline{\mathbf{Y}} - \widehat{\boldsymbol{\mu}})^{'}$, where $\widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. Starting with $\widetilde{\mathbf{V}} = \widehat{\mathbf{V}}$, one iterates these equations until satisfactory convergence is obtained. So, $\widetilde{\mathbf{V}} = \mathbf{W}(\widehat{F})$, where $\mathbf{W}(F)$ is defined in terms of $\boldsymbol{\mu}$, $\mathbf{V}$ by the implicit equation

$$\mathbf{W} = \mathbf{V} + (\mathbf{I}_p - \mathbf{X}\mathbf{L})\,\boldsymbol{\mu}\boldsymbol{\mu}^{'}\,(\mathbf{I}_p - \mathbf{X}\mathbf{L})$$

at $\mathbf{L} = \mathbf{L_W}$ of (2.6). Differentiating, we have

$$\mathbf{W}_{\cdot a} = \mathbf{V}_{\cdot a} - \overline{\mathbf{P}}_a + \overline{\mathbf{Q}}_a, \quad \text{where } \mathbf{P}_a = \mathbf{X}\mathbf{L}_{\cdot a}\, \boldsymbol{\mu}\boldsymbol{\mu}'(\mathbf{I}_p - \mathbf{X}\mathbf{L})' = \mathbf{0}$$

by (2.5) and

$$\mathbf{Q}_a = (\mathbf{I}_p - \mathbf{X}\mathbf{L})\boldsymbol{\mu}_{\cdot a}\, \boldsymbol{\mu}'(\mathbf{I}_p - \mathbf{X}\mathbf{L})' = \mathbf{0}$$

by (2.5). Similarly, $\mathbf{W}_{\cdot bc} = \mathbf{V}_{\cdot bc} - \overline{\mathbf{P}}_{bc}$, where $\mathbf{P}_{bc} = \mathbf{X}\mathbf{L}_{\cdot bc}\, \boldsymbol{\mu}\boldsymbol{\mu}'(\mathbf{I}_p - \mathbf{X}\mathbf{L})' = \mathbf{0}$. So, $\mathbf{W} = \mathbf{V}$, $\mathbf{W}_{\cdot a} = \mathbf{V}_{\cdot a}$ and $\mathbf{W}_{\cdot bc} = \mathbf{V}_{\cdot bc}$. This proves that the maximum likelihood estimate for $\mathbf{V}$ satisfies (1.5).

4. THE WEIGHTED LEAST SQUARES ESTIMATE FOR SEVERAL SAMPLES

Suppose we observe q random samples, the jth sample being from a distribution F_j on $\mathbb{R}^{p_j}$ with mean $\boldsymbol{\mu}_j = \mathbf{X}_j\boldsymbol{\beta}$, where $\mathbf{X}_j$ is a known $p_j \times r$ matrix and $\boldsymbol{\beta}$ an unknown r vector. Let $\mu^j_{i_1,\ldots,i_s}$ denote $\mu_{i_1,\ldots,i_s}$ for F_j. Set

$$\mathbf{V}_j = (\mu^j_{ab}), \quad \mathbf{D}_j = \mathbf{X}_j'\, \mathbf{V}_j^{-1}\mathbf{X}_j, \quad \mathbf{N}_j = \mathbf{X}_j'\, \mathbf{V}_j^{-1}\, \boldsymbol{\mu}_j, \quad \mathbf{L}^j = \mathbf{D}_j'\, \mathbf{X}_j'\, \mathbf{V}_j^{-1}.$$

If $\{\mathbf{V}_j\}$ were known, the best linear unbiased estimate of $\boldsymbol{\beta}$ (and the maximum likelihood estimate for $F = \{F_j\}$ normal) is $g(\overline{\mathbf{Y}}, \mathbf{V})$, where $\overline{\mathbf{Y}} = \{\overline{\mathbf{Y}}_j\}$, $\mathbf{V} = \mathbf{V}(F) = \{\mathbf{V}_j\}$, $\boldsymbol{\mu} = \boldsymbol{\mu}(F) = \{\boldsymbol{\mu}_j\}$, $g(\boldsymbol{\mu}, \mathbf{V}) = \mathbf{D}^{-1}\mathbf{N}$, $\mathbf{D} = \sum_{j=1}^q \lambda_j \mathbf{D}_j$, $\mathbf{N} = \sum_{j=1}^q \lambda_j \mathbf{N}_j$, $\lambda_j = n_0/n_j$ and n_0 is the minimum sample size. We assume that each $\mathbf{V}_j > \mathbf{0}$ and $\mathbf{D} > \mathbf{0}$, that is, $\mathbf{D}$ has rank r, $\mathbf{V}_j$ has rank p_j, $j = 1, \ldots, q$. Put $\widetilde{F} = \{\widehat{F}_j\}$, the set of sample distributions. The weighted least squares estimate is

$$\widehat{\boldsymbol{\beta}} = T(\widehat{F}) = g(\overline{\mathbf{Y}}, \widehat{\mathbf{V}}), \tag{4.1}$$

where $\widehat{\mathbf{V}} = \{\widehat{\mathbf{V}}_j\}$, $\widehat{\mathbf{V}}_j = \mathbf{V}(\widehat{F}_j)$, the jth sample covariance, and

$$T(F) = g\big(\boldsymbol{\mu}(F), \mathbf{V}(F)\big) = \mathbf{D}(F)^{-1}\mathbf{N}(F) \quad \text{for } \mathbf{D}(F) = \mathbf{D},\ \mathbf{N}(F) = \mathbf{N}.$$

THEOREM 4.1. *We have*

$$\operatorname{covar}(\widehat{\boldsymbol{\beta}}) = n_0^{-1}\mathbf{D}^{-1} + O(n_0^{-2}) \tag{4.2}$$

and

$$\mathbb{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = n_0^{-1}\mathbf{C} + O(n_0^{-2}), \tag{4.3}$$

where

$$C_i = -\sum_{j=1}^q \lambda_j\, L^j_{ia} K^j_{bc}\, \mu^j_{abc}$$

with $K^j_{bc} = \lambda_j\, V^{bc}_j - \lambda_j^2\, J^j_{bc}$, $(V^{bc}_j) = \mathbf{V}_j^{-1}$ *and* $\mathbf{J}^j = \mathbf{V}_j^{-1}\mathbf{Y}_j\mathbf{L}^j$.

P r o o f. By equations (2.10), (3.4) and Example 2.1 of Withers [13], (4.2) and (4.3) hold with $\mathbf{D}^{-1}$ replaced by

$$\mathbf{a}_{21} = \left[1^2\right] = \sum_{j=1}^{q} \lambda_j [1^2]_j, \quad 2\mathbf{C} = [11] = \sum_{j=1}^{q} \lambda_j [11]_j,$$

where

$$\left[1^2\right]_j = \partial \mathbf{H} / \partial \big(\mathbb{E}(\mathbf{Z}_j)\big)_a \, \boldsymbol{\nu}_j^{ab} \, \partial \mathbf{H}' / \partial \big(\mathbb{E}(\mathbf{Z}_j)\big)_b,$$

$$[11]_j = \partial^2 \mathbf{H} / \partial \big(\mathbb{E}(\mathbf{Z}_j)\big)_a \partial \big(\mathbb{E}(\mathbf{Z}_j)\big)_b \, \boldsymbol{\nu}_j^{ab},$$

$$\mathbf{Z}_j = \{\mathbf{Y}_j, \mathbf{Y}_j \mathbf{Y}_j'\} \quad \text{for } \mathbf{Y}_j \sim F_j,$$

$$\boldsymbol{\nu}_j^{ab} = \mathbb{E}\big[\big(\mathbf{Z}_j - \mathbb{E}(\mathbf{Z}_j)\big)_a \big(\mathbf{Z}_j - \mathbb{E}(\mathbf{Z}_j)\big)_b\big],$$

$$\mathbf{H} = \mathbf{H}\big(\mathbb{E}(\mathbf{Z}_1), \ldots, \mathbb{E}(\mathbf{Z}_q)\big) = T(F) \quad \text{of (4.1)}.$$

The rest of the proof is similar to that of Theorem 2.1. For example,

$$[11]_j = \mathbf{H}^j_{\cdot a,b} \, \boldsymbol{\mu}^j_{ab} + 2\mathbf{H}^j_{\cdot a,bc} \, \boldsymbol{\mu}^j_{abc} + \mathbf{H}^j_{\cdot ab,cd} \, \boldsymbol{\mu}^j_{abcd},$$

where $\mathbf{H}^j_{\cdot a} = \partial \mathbf{H} / \partial (\boldsymbol{\mu}_j)_a$, $\mathbf{H}^j_{\cdot bc} = \partial \mathbf{H} / \partial (\mathbf{M}_j)_{bc}$, and so on, $\boldsymbol{\mu}_j = \mathbb{E}(\mathbf{Y}_j)$, $\mathbf{M}_j = \mathbb{E}(\mathbf{Y}_j \mathbf{Y}_j')$. ∎

Of special interest is the case when all q distributions are univariate: $p_j \equiv 1$. In this case, $\mathbf{a}_j = \mathbf{X}_j'$ is a column vector, F_j has mean $\mathbf{a}_j' \boldsymbol{\beta}$, $\boldsymbol{\beta}(\boldsymbol{\mu}, \mathbf{V}) = \mathbf{D}^{-1}\mathbf{N}$, where $\mathbf{N} = n_0 \sum_{j=1}^{q} \mu_j \mathbf{a}_j / (n_j V_j)$ and $\mathbf{D} = n_0 \sum_{j=1}^{q} \mathbf{a}_j \mathbf{a}_j' / (n_j V_j)$. Therefore, we need $q \geqslant r$ for $\mathbf{D}$ to have full rank. Also

$$\mathbf{c}_j = \lambda_j \mathbf{L}^j = n_0 \mathbf{D}^{-1} \mathbf{a}_j / \left(n_j V_j\right)$$

is a column vector, and

$$K^j = \lambda_j / V_j - \mathbf{a}_j' \, \mathbf{D}^{-1} \mathbf{a}_j \, \lambda_j^2 \, / V_j^2 = (1 - \mathbf{a}_j' \mathbf{c}_j) \lambda_j / V_j$$

is a scalar. The leading bias term is

$$n_0^{-1} \mathbf{C} = - \sum_{j=1}^{q} \mathbf{c}_j (1 - \mathbf{a}_j' \mathbf{c}_j) \mu^j_{111} / (n_j V_j).$$

Note that $\sum_{j=1}^{q} \mathbf{c}_j \mathbf{a}_j' = \mathbf{I}_r$.

If also β is univariate ($r = 1$) then $D = \beta^{-2} D_0$, where $D_0 = \sum_{j=1}^{q} \lambda_j \mu_j^2 / V_j$, and the relative bias is (approximately)

$$n_0^{-1} C / \beta = -n_0^{-1} D_0^{-1} \sum_{j=1}^{q} \lambda_j^2 \, (1 - q_j) \, \nu_j,$$

where $q_j = (\lambda_j \mu_j^2/V_j)/D_0$, so $\sum_{j=1}^q q_j = 1$, and $\nu_j = \mu_j\ \mu_{111}^j/V_j^2$. These expressions are not changed by rescaling $\{F_j\}$. Consequently, for fixed $\{p_j, \nu_j\}$ the relative bias decreases from ∞ at $D_0 = 0$ to zero at $D_0 = \infty$. If also $F_j \equiv \mathcal{L}\,\sigma_j Y_0$ for some random variable Y_0 and $n_j \equiv n$, then this gives the asymptotic variance

$$(nq)^{-1}\beta^2 \mu_2(Y_0)\big(\mathbb{E}(Y_0)\big)^{-2}$$

and the relative bias (approximately)

$$-n^{-1}(1-q^{-1})\big(\mathbb{E}(Y_0)\big)\mu_3(Y_0)\mu_2(Y_0)^{-2}$$

which doubles as q increases from 2 to ∞: bias $O(n^{-2})$ is only achieved if one discards all samples save one. Of course, the cost of this procedure is to increase the covariance by a factor q. This illustrates that the bias need not decrease as the total sample size increases.

Suppose $p_j \equiv 1$, $r = 1$, $F_j = \mathcal{L}\sigma_j G_j$ for $G_j \sim \mathrm{Gamma}(\gamma_j)$ and σ_i some scale factor. Then, regardless of the choice of $\{Y_j\}$ and $\{\sigma_j\}$,

$$D_0 = \sum_{j=1}^{2} \lambda_j \gamma_j, \quad q_j = \lambda_j \gamma_j / D_0, \quad \nu_j = 2,$$

so D_0 approaches zero as all γ_j approach zero, and D_0 approaches ∞ if any γ_j approaches ∞.

If $n_j \equiv n$ and $\gamma_j \equiv \gamma$, then the relative bias is asymptotically

$$n_0^{-1} C/\beta = -2n^{-1}\gamma^{-1}(1-q^{-1})$$

which doubles from $-n^{-1}\gamma^{-1}$ for two samples to $-2n^{-1}\gamma^{-1}$ for $q = \infty$.

5. A SIMULATION STUDY

Here, we perform a simulation study to compare the usual maximum likelihood estimate, $\widehat{\boldsymbol{\beta}}$, with the bias reduced version, $\widehat{\boldsymbol{\beta}} - n^{-1}\mathbf{C}(\widehat{F})$, given by Theorem 1.1. For simplicity, we take into consideration the model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $i = 1, 2, \ldots, n$, where β_0 is the intercept parameter, β_1 is the slope parameter, and ϵ_i are independent errors not necessarily normally distributed. In our simulations, we take $\beta_0 = 0$, $\beta_1 = 1$, $x_i = i$, $i = 1, 2, \ldots, n$, $n = 10, 20, \ldots, 1000$ and consider the following distributions for $\{\epsilon_i, i = 1, 2, \ldots, n\}$:

1. skew normal distribution (Azzalini [2]) with $\lambda = 1$;
2. skew normal distribution (Azzalini [2]) with $\lambda = 5$;
3. standard Gumbel distribution;
4. skew Cauchy distribution (Behboodian et al. [3]) with $\lambda = 1$.

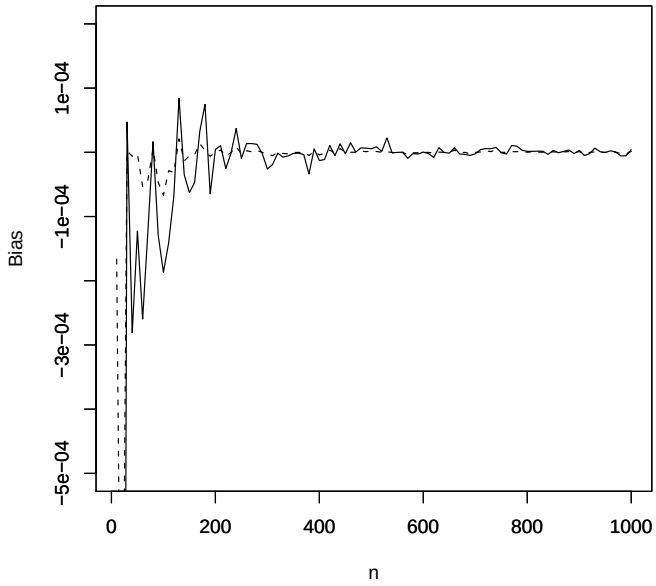

FIGURE 1. The biases for the maximum likelihood (solid curve) and bias reduced (broken curve) estimates of β_1 when the errors follow the skew normal distribution with $\lambda = 1$

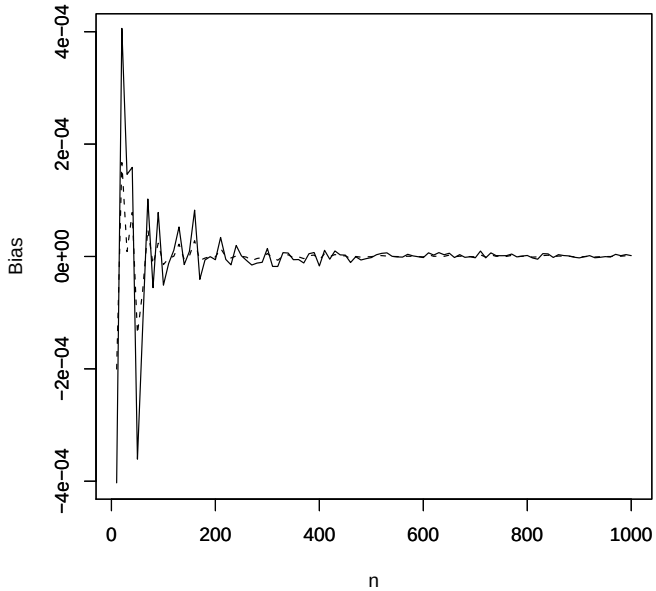

FIGURE 2. The biases for the maximum likelihood (solid curve) and bias reduced (broken curve) estimates of β_1 when the errors follow the skew normal distribution with $\lambda = 5$

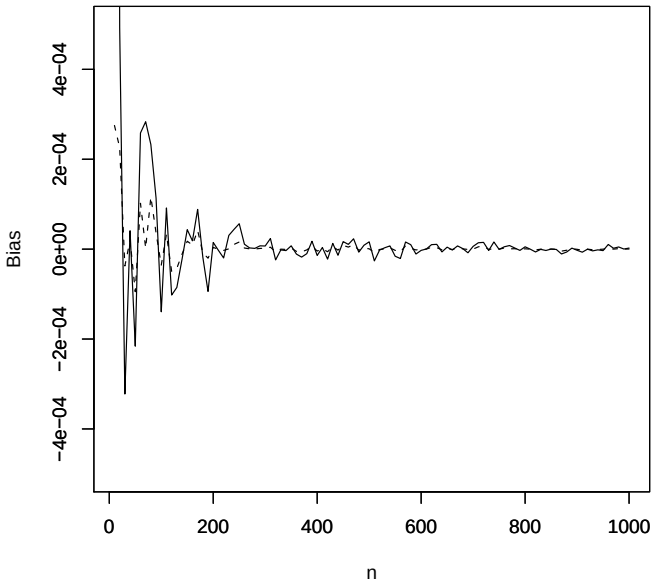

FIGURE 3. The biases for the maximum likelihood (solid curve) and bias reduced (broken curve) estimates of β_1 when the errors follow the standard Gumbel distribution

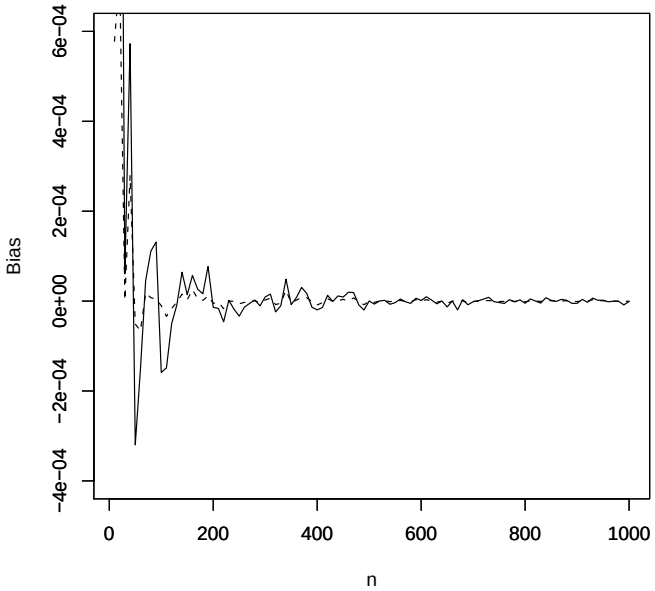

FIGURE 4. The biases for the maximum likelihood (solid curve) and bias reduced (broken curve) estimates of β_1 when the errors follow the skew Cauchy distribution with $\lambda = 1$

All of these four distributions are asymmetric. The skew normal distribution contains the standard normal as a particular case for $\lambda = 0$.

We assess the performance of the two estimates (the usual maximum likelihood estimate and its bias reduced version) by computing the bias of the slope parameter β_1. This was computed by fitting the simple linear regression model ten thousand times for every n. The biases of the slope parameter *versus* n are plotted in Figures 1–4 for the four distributions.

It is evident that the bias reduced version has smaller biases for all values of n and for the four distributions considered. The biases are consistently smaller for the estimate given by Theorem 1.1. The bias reduction appears substantial for all four distributions for n up to two hundred. With respect to n, the biases generally decrease in magnitude for all of the four distributions.

Acknowledgments. The authors would like to thank the Editor, the Handling Editor and the referee for careful reading and for their comments which greatly improved the paper.

REFERENCES

[1] H. Abdi, *Least squares*, in: *Encyclopedia of Social Sciences Research Methods*, M. Lewis-Beck, A. Bryman and T. Futing (Eds.), Sage, Thousand Oaks, California, 2003.

[2] A. Azzalini, *A class of distributions include the normal ones*, Scand. J. Statist. 12 (1985), pp. 171–178.

[3] J. Behboodian, A. Jamalizadeh and N. Balakrishnan, *A new class of skew-Cauchy distributions*, Statist. Probab. Lett. 76 (2006), pp. 1488–1493.

[4] J.-M. Bonny, M. Zanca, J.-Y. Boire and A. Veyre, *T_2 maximum likelihood estimation from multiple spin-echo magnitude images*, Magnet. Reson. Med. 36 (1996), pp. 287–293.

[5] S. G. Candy, A. J. Constable, T. Lamb and R. Williams, *A Von Bertalanffy growth model for toothfish at Heard Island fitted to length-at-age data and compared to observed growth from mark-recapture studies*, CCAMLR Science 14 (2007), pp. 43–66.

[6] K. Emrich and W. Urfer, *Benefits and complications of maximum likelihood estimation in (composite) interval mapping methods using EM and ECM*, Euphytica 137 (2004), pp. 155–163.

[7] J. A. Hausman and D. A. Wise, *Stratification on endogenous variables and estimation: the Gary income maintenance experiment*, in: *Structural Analysis of Discrete Data with Econometric Application*, C. Manski and D. McFadden (Eds.), Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1981, pp. 364–391.

[8] J. Jiang, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York 2007.

[9] K. G. Mehrotra, P. M. Kulkarni, R. M. Tripathi and J. E. Michalek, *Maximum likelihood estimation for longitudinal data with truncated observations*, Stat. Med. 19 (2000), pp. 2975–2988.

[10] M. M. Olsen, J. Swevers and W. Verdonck, *Maximum likelihood identification of a dynamic robot model: implementation issues*, Int. J. Robot. Res. 21 (2002), pp. 89–96.

[11] C. S. Withers, *Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparameteric confidence intervals*, Ann. Statist. 11 (1983), pp. 577–587.

[12] C. S. Withers, *Bias reduction by Taylor series*, Commun. Stat. – Theory and Methods 16 (1987), pp. 2369–2384.
[13] C. S. Withers, *Nonparametric confidence intervals for functions of several distributions*, Ann. Inst. Statist. Math. 40 (1988), 727–746.

Applied Mathematics Group
Industrial Research Limited
Lower Hutt, New Zealand
E-mail: c.withers@cri.irl.nz

School of Mathematics
University of Manchester
Manchester M13 9PL
United Kingdom
E-mail: mbbsssn2@manchester.ac.uk

Received on 11.2.2011;
revised version on 7.11.2011