Electronic Journal of Statistics
Vol. 4 (2010) 1–14
ISSN: 1935-7524
DOI: 10.1214/09-EJS447

The bias and skewness of M-estimators in regression

Christopher Withers

Applied Mathematics Group
Industrial Research Limited
Lower Hutt, New Zealand

and

Saralees Nadarajah

School of Mathematics
University of Manchester
Manchester, United Kingdom
e-mail: mbbsssn2@manchester.ac.uk

Abstract: We consider M estimation of a regression model with a nuisance parameter and a vector of other parameters. The unknown distribution of the residuals is not assumed to be normal or symmetric. Simple and easily estimated formulas are given for the dominant terms of the bias and skewness of the parameter estimates. For the linear model these are proportional to the skewness of the 'independent' variables. For a nonlinear model, its linear component plays the role of these independent variables, and a second term must be added proportional to the covariance of its linear and quadratic components. For the least squares estimate with normal errors this term was derived by Box [1]. We also consider the effect of a large number of parameters, and the case of random independent variables.

AMS 2000 subject classifications: Primary 62G08, 62G20.
Keywords and phrases: Bias reduction, M-estimates, non-linear, regression, robust, skewness.

Received July 2009.

Contents

1. Introduction

The asymptotic theory of M estimates for regression models has been the subject of many papers. We refer the readers to the excellent book by Maronna *et al.* [7] for a comprehensive review.

In this note, we give formulas for the dominant terms of the bias and skewness of M-estimates in linear and nonlinear regression models. These formulas could have applications in many areas, including bias reduction, confidence regions and Edgeworth expansions. For the least squares estimate the formula for bias is just that given by Box [1] for the case of normal errors. The main results are in Section 2 with proofs deferred to Section 6. The exact regularity conditions are not given but some sufficient conditions are discussed. Section 3 gives some applications. Section 4 considers the effect of a large number of parameters on bias and skewness, and shows how to adapt the results when the 'independent' variables are random. Section 5 gives some simulation results for an L$_1$-estimate. Section 7 extends our results to the case, where the residuals may have different distributions.

2. Main results

First consider the linear model: we observe

$$Y_N = \alpha + x_N^{'}\beta + e_N,\ 1 \le N \le n, \tag{2.1}$$

where x_N and β are, respectively, known and unknown p-vectors, $\{e_N\}$ are random residuals with an unknown distribution F, with density f (if it exists) not necessarily symmetric, and α is a nuisance parameter centering the residuals around zero in some way. We estimate the unknown parameters $\varphi = (\alpha, \beta)$ by $(\widehat{\alpha}, \widehat{\beta})$ minimising

$$\lambda = \sum_{N=1}^{n} \rho\left(Y_N - \alpha - x_N^{'}\beta\right), \tag{2.2}$$

where ρ is some smooth function for which a minimum exists. By smooth we mean that its derivatives exist except at a finite number of points.

It turns out as is well known that for $(\widehat{\alpha}, \widehat{\beta})$ to be consistent, α must satisfy the centering condition:

$$\rho_1 = 0,\ G = \text{ distribution of } \alpha + e_N, \tag{2.3}$$

where $\rho_1 = \int \rho^{(1)}(\nu - \alpha) dG(\nu)$ and $\rho^{(r)}$ denotes the rth derivative of ρ. In general, set

$$\rho_{rs\cdots} = \int \rho^{(r)}(e)\rho^{(s)}(e)\cdots dF(e),$$

where the number of indices on the left hand side is the same as the number of terms in the integrand on the right hand side. For example, $\rho_2 = \int \rho^{(2)}(e) dF(e)$,

$\rho_{11} = \int \rho^{(1)}(e)\rho^{(1)}(e)dF(e)$ and $\rho_{123} = \int \rho^{(1)}(e)\ \rho^{(2)}(e)\ \rho^{(3)}(e)\ dF(e)$. The condition in (2.3) makes α identifiable.

We now show that the bias and skewness of $\widehat{\beta}$ is essentially proportional to the 'skewness' (third central moments) of the $\{x_N\}$. Set

$$\overline{x} = n^{-1} \sum_{N=1}^{n} x_N,\ m_{ij\cdots} = n^{-1} \sum_{N=1}^{n} \left(x_N - \overline{x}\right)_i \left(x_N - \overline{x}\right)_j \cdots, \tag{2.4}$$

where $(\cdot)_i$ is the ith component, $M = (m_{ij})p \times p$ and $(m^{ij}) = M^{-1}$. Suppose

$$\text{det(M) is bounded away from zero as } n \to \infty. \tag{2.5}$$

Theorem 2.1. *Suppose (2.1), (2.3), (2.5) hold and $(\widehat{\alpha}, \widehat{\beta})$ minimises*

$$\lambda = \sum_{N=1}^{n} \rho\left(Y_N - \alpha - x_N^{'}\beta\right).$$

Then for ρ, F and $\{x_N\}$ suitably regular

$$n^{1/2}\left(\widehat{\beta} - \beta\right) \overset{\mathcal{L}}{\to} N_p\left(0, c_1 M^{-1}\right) \tag{2.6}$$

as $n \to \infty$, where $c_1 = \rho_{11}\rho_2^{-2}$. Furthermore, $\widehat{\beta}$ has bias, covariance, third cumulants and skewness as:

$$E\left(\widehat{\beta} - \beta\right)_a = n^{-1}K_a + O\left(n^{-2}\right)$$

for $1 \le a \le p$, where

$$K_a = c_2 \sum_{i,j,k=1}^{p} m^{ai}m^{jk}m_{ijk},\ c_2 = -\rho_{12}\rho_2^{-2} + \rho_{11}\rho_3\rho_2^{-3}/2, \tag{2.7}$$

and

$$cov\left(\widehat{\beta}_a, \widehat{\beta}_b\right) = n^{-2}K_{ab} + O\left(n^{-2}\right),$$
$$E\left(\widehat{\beta} - \beta\right)_a\left(\widehat{\beta} - \beta\right)_b = n^{-2}K_{ab} + O\left(n^{-2}\right),$$

where $K_{ab} = c_1 m^{ab}$ for $1 \le a, b \le p$. The third cumulants are given by

$$\kappa\left(\widehat{\beta}_a, \widehat{\beta}_b, \widehat{\beta}_c\right) = n^{-2}K_{abc} + O\left(n^{-3}\right)$$

and

$$E\left(\widehat{\beta} - \beta\right)_a\left(\widehat{\beta} - \beta\right)_b\left(\widehat{\beta} - \beta\right)_c = n^{-2}K_{abc}^{'} + O\left(n^{-3}\right)$$

for $1 \leq a, b, c \leq p$*, where*

$$K_{abc} = c_3 \sum_{i,j,k=1}^{p} m^{ai} m^{bj} m^{ck} m_{ijk}, \tag{2.8}$$
$$c_3 = \rho_{111}\rho_2^{-3} - 6\rho_{11}\rho_{12}\rho_2^{-4} + 3\rho_{11}^2\rho_3\rho_2^{-5}$$

and

$$K_{abc}^{'} = K_{abc} + \sum_{abc}^{3} K_{ab}K_c,$$

where $\sum_{abc}^3 f_{abc} = f_{abc} + f_{bca} + f_{cab}$*, while a plain* $\sum$ *sums repeated pairs of suffixes over their range (that is* i*,* j*,* k *over* $1 \cdots p$ *in (2.7), (2.8)).*

Note that the M on the right hand side of (2.6) should be interpreted as the limit of the M defined by (2.4) as $n \to \infty$.

Note that page 169 of Huber [5] essentially gave conditions for (2.6) to hold. In particular, if $p = 1$ in Theorem 2.1 and $\mu_r(\widehat{\beta})$ is the rth central moment of $\widehat{\beta}$, then

$$m_{11} = n^{-1}\sum_{1}^{n}\left(x_N - \overline{x}\right)^2, \ m_{111} = n^{-1}\sum_{1}^{n}\left(x_N - \overline{x}\right)^3,$$
$$E\widehat{\beta} = \beta + n^{-1}K_1 + O\left(n^{-2}\right), \ K_1 = c_2 m_{11}^{-2} m_{111},$$
$$\mu_2\left(\widehat{\beta}\right) = n^{-1}K_{11} + O\left(n^{-2}\right), \ K_{11} = c_1 m_{11}^{-1},$$
$$E\left(\widehat{\beta} - \beta\right)^2 = n^{-1}K_{11} + O\left(n^{-2}\right), \ K_{11} = c_1 m_{11}^{-1},$$
$$\mu_3\left(\widehat{\beta}\right) = n^{-2}K_{111} + O\left(n^{-3}\right), \ K_{111} = c_3 m_{11}^{-3} m_{111},$$
$$E\left(\widehat{\beta} - \beta\right)^3 = n^{-2}K_{111}^{'} + O\left(n^{-3}\right), \ K_{111}^{'} = c_3^{'} m_{11}^{-3} m_{111},$$

where $c_3^{'} = c_3 + 3c_1c_2 = \rho_{111}\rho_2^{-3} - 9\rho_{11}\rho_{12}\rho_2^{-4} + 9\rho_{11}^2\rho_3\rho_2^{-5}/2$.

A sufficient but not necessary condition on $\{x_N\}$ is that they be bounded. This ensures that $m_{ij\cdots}$ is bounded. For $\rho_{rs\cdots}$ to be finite a sufficient but not necessary condition is that $\rho^{(r)}, \rho^{(s)} \cdots$ are bounded.

Corollary 2.1. *For any* ρ *there exists* F *for which* $\widehat{\beta}$ *has arbitrarily large (asymptotic) variance.*

This is because $\sup c_1 = \infty$. We cannot have $\inf \rho^{(2)} > 0$ and $\rho^{(1)}$ bounded.

Note that models which assume (generally unrealistically) that $f = \dot{F}$ is symmetric (for example, normal) force the bias and skewness down to $\sim n^{-2}$ and n^{-3} (since $c_2 = c_3 = 0$ assuming ρ is symmetric).

These formulas for bias and skewness have an immediate application to experimental design: if $\{x_N\}$ are chosen to have skewness zero (or $\sim n^{-1}$) then the bias and skewness (third moments) of $\widehat{\beta}$ are reduced from $\sim n^{-1}$ and n^{-2} to $\sim n^{-2}$ and n^{-3}, and so the nominal level of the one-sided confidence interval for β_a based on approximate normality has error reduced from $\sim n^{-1/2}$ to $\sim n^{-1}$.

Example 2.1. *For the L_1-estimate, $\rho(e) =\mid e \mid$. So, $\rho_1 = E$ sign (e_1), $\alpha =$ median of G, $\rho_2 = 2f(0)$, $\rho_{11} = 1$, $\rho_3 = -2\dot{f}(0)$ and $\rho_{111} = \rho_{12} = 0$. (Here, we use $\rho^{(2)}(e) = 2\delta(e)$ for δ the Dirac function.) So, $c_1 = f(0)^{-2}/4$ (as is well known for $p = 0$, i.e. for the variance of the sample median), $c_2 = -\dot{f}(0)f(0)^{-3}/8$, $c_3 = -3\dot{f}(0)f(0)^{-5}/16$ and $c_3^{'} = -9\dot{f}(0)f(0)^{-5}/32$.*

We now consider the general *non-linear* model

$$Y_N = \alpha + f_N(\beta) + e_N, \ 1 \leq N \leq n. \tag{2.9}$$

That is we replace the regression functions $\{x_N^{'}\beta\}$ by smooth functions $\{f_N(\beta)\}$. We shall see that the role of x_N in Theorem 2.1 is now replaced by

$$x_N = \partial f_N(\beta)/\partial\beta,$$

but there is an additional term in the bias and skewness proportional to the covariances between the linear and quadratic components of the model, i.e.

$$m_{i,jk} = n^{-1} \sum_{N=1}^{n} \left(x_N - \overline{x}\right)_i \left(x_N - \overline{x}\right)_{jk},$$

where $(x_N)_{ij\cdots} = \partial_i\partial_j \cdots f_N(\beta)$ and $\partial_i = \partial/\partial\beta_i$.

Define $m_{ij\cdots}$ and m^{ij} as before. So, $\{m_{ijk}\}$ now consists of the third moments of the linear components of the model, $\{x_N = \partial f_N(\beta)/\partial\beta\}$.

Theorem 2.2. *Theorem 2.1 remains valid with $\{x_N^{'}\beta\}$ replaced by suitably smooth $\{f_N(\beta)\}$ in (2.1) and (2.2) and K_a, K_{abc} replaced by*

$$K_a = \sum_{i,j,k} m^{ai}m^{jk}\left(c_2 m_{ijk} - c_1 m_{i,jk}/2\right),$$

and

$$K_{abc} = \sum_{i,j,k} m^{ai}m^{bj}m^{ck}\left(c_3 m_{ijk} - c_1^2 \sum_{ijk}^{3} m_{i,jk}\right).$$

Example 2.2. *For the least squares estimate, $\rho(e) = e^2/2$. So, $Ee_1 = 0$, $c_1 = \mu_2(e_1)$, $c_2 = 0$, $c_3 = c_3^{'} = \mu_3(e_1)$. In particular, $\widehat{\beta}_a$ has bias $n^{-1}K_a + O(n^{-2})$, where $K_a = -\mu_2(e_1)\sum_{i,j,k=1}^{p} m^{ai}\ m^{jk}\ m_{i,jk}/2$. For F normal this result is essentially equation (2.24) in Box [1]; see also Clarke [2] for a less explicit formula. For $p = 1$ the formula for the skewness of $\widehat{\beta}$ yields $K_{111}^{'} = m_{11}^{-3}(\mu_3(e_1)m_{111} - 6\mu_2(e_1)^2 m_{11})$. Jennrich [6] proved asymptotic normality for this case.*

Example 2.3. *For Huber's estimate,*

$$\rho^{(1)}(e) = \begin{cases} e, & \mid e \mid \leq 1, \\ sign(e), & \mid e \mid > 1. \end{cases}$$

So,

$$\begin{aligned} \rho_1 &= \int_{-1}^{1} e dF(e) + \int_{1} dF - \int^{-1} dF, \;\; \rho_2 = \int_{-1}^{1} dF, \\ \rho_{11} &= \int_{-1}^{1} e^2 dF(e) + \int_{1} dF + \int^{-1} dF, \;\; \rho_3 = f(-1) - f(1), \\ \rho_{12} &= \int_{-1}^{1} e dF(e), \;\; \rho_{111} = \int_{-1}^{1} e^3 dF(e) + \int_{1} dF - \int^{-1} dF. \end{aligned}$$

For ρ_3 use $\rho^{(3)}(e) = \delta(e+1) - \delta(e-1)$, where δ is the Dirac function.

3. Applications

Note M, c_i, K_a, K_{abc} and $K_{abc}^{'}$ are easily estimated. Let $\widehat{F}$ be the empirical distribution of the estimated residuals $\{e_N(\widehat{\varphi}), 1 \leq N \leq n\}$, where $\varphi = (\alpha, \beta)$ and $e_N(\varphi) = Y_N - \alpha - f_N(\beta)$. Let $\widehat{M}$, $\widehat{c}_i$, $\widehat{K}_a$, ... denote their values at $(\widehat{\beta}, \widehat{F})$. Then for ρ suitably regular

$$E\widehat{c}_i = c_i + O\left(n^{-1}\right), \tag{3.1}$$

so

$$\widehat{\beta}_a - n^{-1}\widehat{K}_a \text{ estimates } \beta_a \text{ with bias } \sim n^{-2} \tag{3.2}$$

and variance $\sim n^{-1}$, and the confidence region

$$\left\{\beta : \left(\beta - \widehat{\beta}\right)^{'} \widehat{M} \left(\beta - \widehat{\beta}\right) \leq z\widehat{c}_1/n\right\} \text{ has level } P\left(\chi_p^2 \leq z\right) + O\left(n^{-1}\right). \tag{3.3}$$

Unlike the jackknife or bootstrap versions of $\widehat{\beta}_a$ which require $\sim n^2$ or more calculations to reduce the bias to $\sim n^{-2}$ (and retain variance $\sim n^{-1}$) the estimate in (3.2) only requires $\sim n$ calculations.

Estimates for which $\rho^{(1)}$ or $\rho^{(2)}$ discontinuous, such as the L$_1$-estimate or Huber's, fail (3.1). Let $\widetilde{f}$ and $\widetilde{\dot{f}}$ be kernel estimates of f and $\dot{f}$ with kernels of order m so that in the usual notation, bias $\sim h^m$ and variance $\sim h^{-1}n^{-1}$. Let $\widetilde{c}_1$ and $\widetilde{K}_a$ be the corresponding estimate of c_1 and K_a for these examples. Suppose $h = n^{-e}$, where $m^{-1} \leq e \leq 2$. Then $\widehat{\beta}_a - n^{-1}\widetilde{K}_a$ estimates β_a with bias $\sim n^{-2}$ and variance $\sim n^{-1}$, as for (3.2).

For Huber's estimate, (3.3) holds since (3.1) holds for $i = 1$. For the L$_1$-estimate (3.3) holds with $\widehat{c}_1, O(n^{-1})$ replaced by $\widetilde{c}_1, O(n^{\delta-1})$ for kernel estimates with $h = n^{-\delta}$, $\delta = (2m+1)^{-1}$.

Our expressions for bias and skewness enable us to calculate the first term of the Edgeworth expansion for the distribution of $Y_n = n^{1/2}(\widehat{\beta} - \beta)$ and its Studentised version. In particular, if $Z_n = n^{1/2}(\widehat{\beta}_a - \beta_a)K_{aa}^{-1/2}$ or $n^{1/2}(\widehat{\beta}_a - \beta_a)\widehat{K}_{aa}^{-1/2}$,

$$\Delta_a = K_a + K_{aa}^{-1} K_{aaa} \left(z^2 - 1\right)/6 \text{ and } \delta_n = n^{-1/2} \Delta_a K_{aa}^{-1/2}, \tag{3.4}$$

and Φ, ϕ are the distribution and density of a unit normal random variable, then $P(Z_n \leq z) = \Phi(z) - \delta_n \phi(z) + O(n^{-1})$. So, a one-sided confidence interval for β_a of level $\Phi(z) + O(n^{-1})$ is

$$\beta_a \geq \widehat{\beta}_a - n^{-1/2} \widehat{K}_{aa}^{-1/2} z - n^{-1} \widehat{\Delta}_a. \tag{3.5}$$

If we drop the last term in (3.5) we must replace $O(n^{-1})$ by $O(n^{-1/2})$; c.f. Withers [10, 11]. If $p = 1$ (3.4) can be written

$$\Delta_1 = m_{11}^{-2} \left\{ m_{111} \left(\rho_{111} \rho_{11}^{-1} \rho_2^{-1} \left(z^2 - 1\right)/6 + c_2 z^2 \right) - m_{1,11} c_1 z^2/2 \right\}.$$

See Withers [10] for more details on this type of applications. Note K_a and K_{abc} may also be used to obtain 'small sample asymptotics' for the density and tails of $\widehat{\beta}$ by the method of Easton and Ronchetti [3].

4. Some miscellaneous results

Here, we consider two separate topics: the effect of p large, and the effect of random independent variables. Suppose both n and p large. Let $\Delta =$max$|\ m^{ij}\ |$.

Now $p\Delta$ is bounded away from zero since $M \sim 1$ by assumption and $\sum_j m_{ij}\ m^{jk} = \delta_{ik}$. Typically $\Delta \sim p^{-1}$ as $p \to \infty$. From our formulas, $K_a \sim \Delta^2 p^2$ and $K_{abc}, K_{abc}^{'} \sim \Delta^3 p^3$. So, if $\Delta \sim p^{-1}$, $\widehat{\beta}$ has bias $\sim p/n$ and skewness (third cumulants) $\sim n^{-2}$, c.f. Portnoy [8] who shows for the linear case that $\widehat{\beta} - \beta = O_p(p/n)$.

Set s.e. = standard error = (variance)$^{1/2}$, r.b. = relative bias = bias/s.e. and r.s. = relative skewness = skewness/s.e.^{3}. Then for $\widehat{\beta}$, r.b. $\sim (p^6 \Delta^3/n)^{1/2} \sim (p^3/n)^{1/2}$ if $\Delta \sim p^{-1}$ and r.s. $\sim (p^6 \Delta^3/n)^{1/2} \sim (p/n)^{1/2}$ if $\Delta \sim p^{-1}$. This follows from the expression in Theorem 2.2 for K_a, K_{ab}, K_{abc}. So, r.b. can become infinite if $p^3/n \to \infty$. Huber [5] (see page 172) showed this for the linear case under the stronger condition $p^{3/2}/n \to \infty$.

We now ask, what if $\{x_N\}$ in the linear model or $\{f_N\}$ in the nonlinear model are random?

Theorem 4.1. *Suppose (2.9) holds with $f_N(\beta) = f(\beta, U_N)$, where $\{U_N\}$ is a random sample independent of $\{e_N\}$. Then Theorem 2.2 holds with the sample values m_{ij}, m_{ijk} and $m_{i,jk}$ replaced by their population values $\kappa((x_N)_i,\ (x_N)_j)$, $\kappa((x_N)_i,\ (x_N)_j,\ (x_N)_k)$ and $\kappa((x_N)_i,\ (x_N)_{jk})$.*

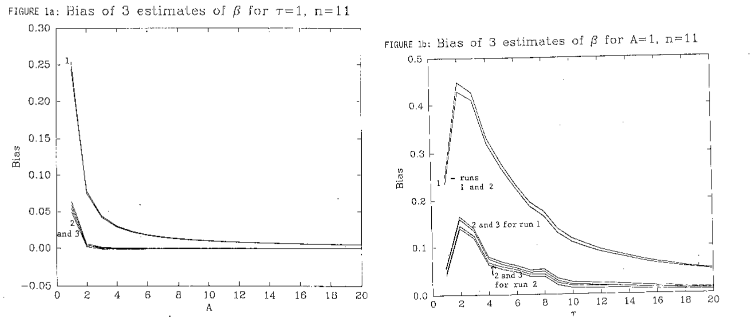

FIGURE 1a: Bias of 3 estimates of β for $\tau=1$, n=11

FIGURE 1b: Bias of 3 estimates of β for A=1, n=11

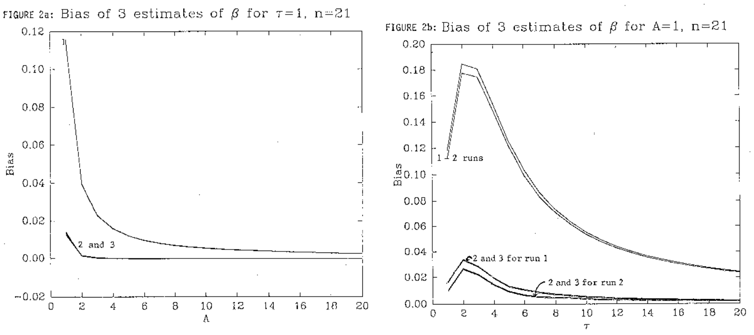

FIGURE 2a: Bias of 3 estimates of β for $\tau=1$, n=21

FIGURE 2b: Bias of 3 estimates of β for A=1, n=21

5. A numerical example

We noted that our formula for bias in the nonlinear model extends one of Box [1] for the least squares estimate. He gives a numerical example for F normal and $n = 3$ showing excellent agreement between exact bias (as estimated by 52,000 simulations) and our formula for K_a. Here, we do a similar comparison for the linear model with $p = 1$, $x_N = \widetilde{x}_{N-m-1}$, $\widetilde{x}_i = i/n + \tau(i/n)^2$ for $\mid i \mid\leq m$, where $n = 2m + 1$ and τ is assumed known. By changing τ this allows for an arbitrary value of μ_{3x}. We take $\rho(e) =\mid e \mid$ and $F(e) = G(1 + e)$, where $G(\nu) = 1 - \nu^{-A}/2$ for $\nu \geq 2^{-1/A}$ and $A > 0$. This is chosen as an example of a one-sided heavy-tailed distribution. (As noted in Example 2.1, α transforms G so that F has median zero.) So, $\widehat{\beta}$ has bias $n^{-1}K_1 + O(n^{-2})$, where $K_1 = c_2\mu_{2x}^{-2}\mu_{3x}$, $\mu_{2x} = \mu_X + \tau^2\mu_{2X}$, $\mu_{3x} = 3\tau\mu_{2X} + \tau^3\mu_{3X}$, and μ_X, μ_{rX} is the mean and rth central moment of $\{X_i = (i/n)^2, \mid i \mid\leq m\}$. Since $\{X_i\}$ has rth noncentral moment $\int_{-1/2}^{1/2} t^{2r}dt + O(n^{-1})$, $(\mu_X, \mu_{2X}, \mu_{3X}) = (1/12, 1/180, 1/3780) + O(n^{-1})$ and so $K_1 = K_1^* + O(n^{-1})$, where $K_1^* = 15(1 + \tau^2/15)^{-2}(\tau + \tau^3/63)c_2/4$, which is zero at $\tau = 0$ or ∞. Figures 1 and 2 plot the estimated biases of $\widehat{\beta}$, $\widehat{\beta} - n^{-1}K_1$ and $\widehat{\beta} - n^{-1}K_1^*$ (labeled 1, 2 and 3) against A and τ.

Estimates were obtained from two separate runs of 10^5 simulations each for $\beta = 1$, $\alpha = 0$ and $n = 11, 21$. Calculations were done using NAG routine E02 GAF. This took nearly 48 hours of CPU time on a VAX 780. The large number of simulations was required to obtain good accuracy, as indicated by the small variation between runs. This may be due to the non-uniqueness of the L_1-estimate, or more fundamentally the fact that $\rho(e) =\mid e \mid$ has a discontinuous derivative. The bias estimates $n^{-1}K_1$ and $n^{-1}K_1^*$ are seen to be excellent, and almost indistinguishable at the number of simulations.

6. Extensions to non-identical residual distributions

Here, we extend Theorems 2.1 and 2.2 to the case, where instead of being i.i.d. F,

$$\{e_N\} \text{ are independent with distributions } \{F_N\}. \tag{6.1}$$

We assume that each F_N is centered so that $E\rho^{(1)}(e_N) = 0$. Set $\rho_{Nrs\ldots} = E\rho^{(r)}(e_N)\rho^{(s)}(e_N)\ldots$ and define the linear operator $\varrho_{rs\ldots}$ by

$$\begin{aligned}\varrho_{rs\ldots}g_{i_1i_2\cdots} &= n^{-1}\sum_{N=1}^{n}\rho_{Nrs\ldots},g_{N.i_1}g_{M.i}\cdots,\\ \rho_{\mathbf{rs}\ldots}g_{i_1i_2\cdots,j_1j_2\cdots} &= n^{-1}\sum_{N=1}^{n}\rho_{Nrs\ldots}g_{N.i_1}g_{N.i_2}\cdots g_{N.j_1j_2\ldots}.\end{aligned}$$

Set $a_{ij} = \rho_2 g_{ij}$ and $(a^{ij}) = (a_{ij})^{-1}$.

Theorem 6.1. *Under the condition of Theorem 2.2 with the condition that $\{e_N\}$ are i.i.d. weakened to (6.1),*

$$n^{1/2}\left(\widehat{\beta}-\beta\right) \xrightarrow{\mathcal{L}} N_p(0,V)$$

as $n\to\infty$, where

$$V_{ij} = a^{ik}\left(\varrho_{11}g_{kl}\right)a^{lj}. \tag{6.2}$$

The other results of Theorem 2.2 hold with K_a, K_{abc} replaced by

$$K_a = a^{ai}\left\{-a^{jk}\varrho_{12}+V_{jk}\varrho_3\right\}g_{ijk}+a^{ai}\left\{\left(a^{jk}\varrho_{11}-V_{jk}\varrho_2\right)g_{j,ki}-V_{jk}\varrho_2 g_{i,jk}/2\right\},$$

and

$$\begin{aligned}K_{abc} &= a^{ai}\left\{a^{bi}a^{ck}\left(\varrho_{111}-V_{bj}\varrho_{12}\right)-3V_{bj}V_{ck}\varrho_3/2\right\}g_{ijk}\\ &\quad+\sum_{abc}^{3}\left\{a^{aj}V_{ba}a^{ci}\varrho_{11}g_{i,jk}-a^{ai}V_{bj}V_{ck}\varrho_2\sum_{ijk}^{3}g_{i,jk}/2\right\}.\end{aligned} \tag{6.3}$$

7. Proofs

Proof of Theorem 2.2. Set $\varphi' = (\alpha',\beta')$ and $g_N = g_N(\varphi) = \alpha + f_N(\beta)$ and let subscripts following a dot denote partial derivatives: $h_{\cdot i_1\cdots i_r} = \partial_1\cdots\partial_r h(\varphi)$ for

$\partial_i = \partial/\partial\varphi_i$, $1 \leq i \leq m = p+1$. The idea of the proof is to obtain a stochastic expansion for $\delta = \widehat{\varphi} - \varphi$.

Any smooth function $h(\varphi)$ has a Taylor series

$$h\left(\widehat{\varphi}\right) \approx \sum_{r=0}^{\infty} \sum_{i_1,\ldots,i_r} h_{i_1\cdots i_r} \delta_{i_1} \cdots \delta_{i_r}/r!.$$

So, $\widehat{\varphi}$ satisfies for $1 \leq i_0 \leq m$

$$0 = n^{-1}\partial\lambda\left(\widehat{\varphi}\right)/\partial\widehat{\varphi}_{i_0} \approx \sum_{r=0}^{\infty} \sum_{i_0,\ldots,i_r} R_{i_0\cdots i_r}\delta_{i_1}\cdots\delta_{i_r}, \tag{7.1}$$

where $R_{i_0\cdots i_r} = n^{-1}\sum_{N=1}^{n} R_{N.i_0\cdots i_r}/r!$ and $R_N(\varphi) = \rho(e_N(\varphi))$.

For X_n a finite subset of $\{R_{i_0\cdots i_r}\}$, $X_n = O_p(1)$ and $n^{1/2}(X_n - EX_n)$ is asymptotically normal and so $O_p(1)$. See Hajek and Sidak [4] for conditions. By the chain rule

$$R_{N.i} = -\rho^{(1)}\left(e_N\right) g_{N.i},\ R_{N.ij} = \rho^{(2)}\left(e_N\right) g_{N.i} g_{N.j} - \rho^{(1)}\left(e_N\right) g_{N.ij}$$

and

$$R_{N.ijk} = -\rho^{(3)}\left(e_N\right) g_{N.i} g_{N.j} g_{N.k} + \rho^{(2)}\left(e_N\right) \sum_{ijk}^{3} g_{N.ij} g_{N.k} - \rho^{(1)}\left(e_N\right) g_{N.ijk}.$$

Set $a_{ij} = ER_{ij} = \rho_2 g_{ij}$, where $g_{ij\cdots} = n^{-1}\sum_{N=1}^{n} g_{N.i} g_{N.j}\cdots$ and $g_{kl\cdots,ij\cdots} = n^{-1}\sum_{N=1}^{n} g_{N.k}\ g_{N.l}\ \cdots\ g_{N.ij\cdots}$. Then $ER_i = -\rho_1 g_i = 0$. So, R_i and $U_{ij} = R_{ij} - a_{ij}$ are $O_p(n^{-1/2})$.

Multiplying (7.1) by a^{hi_0}, where $a = (a_{ij})$ is $m \times m$ and $(a^{ij}) = a^{-1}$, gives

$$\delta_h \approx -\sum_{r=0}^{\infty} \sum_{i_1,\ldots,i_r} S_{hi_1\cdots i_r}\delta_{i_1}\cdots\delta_{i_r},\ 1 \leq h \leq m, \tag{7.2}$$

where $S_{hi_1\cdots i_r} = \sum_{i_0} a^{hi_0} R_{i_0\cdots i_r}$ for $r \neq 1$ and $\sum_{i_0} a^{hi_0} U_{i_0 i_1}$ for $r = 1$. So, $\delta = O_p(n^{-1/2})$. Iterating (7.2) gives

$$\delta_h = f_{hq}\left(\widehat{\theta}_q\right) + O_p\left(n^{-q/2}\right)$$

for $q \geq 2$, where

$$\widehat{\theta}_q = \{S_{hi_1\cdots i_r}, 0 \leq r < q\} = n^{-1}\sum_{N=1}^{n} h_{Nq}\left(e_N\right) \tag{7.3}$$

say, and

$$\begin{aligned} f_{h2}\left(\widehat{\theta}_2\right) &= -S_h, \\ f_{h3}\left(\widehat{\theta}_3\right) &= -S_h + \sum_i S_{hi} S_i - \sum_{i,j}\left(ES_{hij}\right) S_i S_j, \end{aligned} \tag{7.4}$$

and so on. From (7.3) with $q = 2$ we obtain

$$n^{1/2}\left(\widehat{\varphi} - \varphi\right) \xrightarrow{\mathcal{L}} N_m(0, V)$$

as $n \to \infty$, where $V = c_1 g^{-1}$ and $g = (g_{ij})$. Since $\widehat{\theta}$ is a weighted mean of i.i.d. random variables, by Withers [12]

$$E\widehat{\varphi} - \varphi \approx \sum_{n=1}^{\infty} n^{-r} C_r,$$

where $C_r = O(1)$, C_1 has hth component $C_{1h} = \sum_{i,j} K^{ij} f_{h3.ij}(\theta)/2$, $\theta = E\widehat{\theta}$ and $K^{i_1 \cdots i_r} = n^{r-1}\kappa^{i_1 \cdots i_r}(\widehat{\theta}) = n^{-1}\sum_{N=1}^{n} \kappa^{i_1 \cdots i_r}(h_{N3}(e_N))$.

By (7.4) for x, y elements of $\widehat{\theta}$, setting $\partial_x = \partial/\partial_x$ and $\sum_{xy}^{2} h_{xy} = h_{xy} + h_{yx}$,

$$\partial_x f_{h3}\left(\widehat{\theta}\right)\Big|_{\theta} = -I\left(x = S_h\right)$$

and

$$\begin{aligned} \partial_x \partial_y f_{h3}\left(\widehat{\theta}\right)\Big|_{\theta} \quad = \quad & \sum_{xy}^{2} \left\{ \sum_i I\left(x = S_i, y = S_{hi}\right) \right. \\ & \left. - \sum_{i,j} I\left(x = S_i, y = S_j\right) ES_{hij} \right\}. \end{aligned} \tag{7.5}$$

So, $C_{1h} = \sum_i n\kappa(S_i, S_{hi}) - \sum_{j,k} n\kappa(S_j, S_k) ES_{hjk}$.

Now

$$\begin{aligned} & n\kappa\left(S_i, S_j\right) = V_{ij}, \\ & n\kappa\left(S_l, S_{hk}\right) = \sum_{i,j} a^{il} a^{hj} \left\{-\rho_{12} g_{ijk} + \rho_{11} g_{i,jk}\right\}, \\ & ES_{hjk} = \sum_i a^{hi} \left\{ -\rho_3 g_{ijk} + \rho_2 \sum_{i,j,k}^{3} g_{i,jk} \right\} /2, \end{aligned}$$

where $g_{i,jk} = n^{-1}\sum_{N=1}^{n} g_{N.i} g_{N.jk}$. So,

$$C_{1h} = \sum_{i,j,k} g^{hi} g^{jk} \left\{c_2 g_{ijk} - c_1 g_{i,jk}/2\right\}.$$

By the top line on page 67 of Withers [9],

$$\kappa\left(\widehat{\varphi}_a, \widehat{\varphi}_b, \widehat{\varphi}_c\right) = n^{-2}\overline{K}_{abc} + O\left(n^{-3}\right),$$

where

$$\begin{aligned} \overline{K}_{abc} \quad = \quad & -n^2 \kappa\left(S_a, S_b, S_c\right) + n^2 \sum_{abc}^{3} \left[\sum_{ab}^{2} \sum_k \kappa\left(S_{ak}, S_b\right) \kappa\left(S_k, S_c\right) \right. \\ & \left. - \sum_{ab}^{2} \sum_{j,k} \kappa\left(S_j, S_b\right) \kappa\left(S_k, S_c\right) ES_{ajk} \right] \end{aligned}$$

by (7.5). So,

$$\overline{K}_{abc} = \sum_{i,j,k} g^{ai} g^{bj} g^{ck} \left\{ c_3 g_{ijk} - c_1^2 \sum_{ijk}^{3} g_{i,jk} \right\}.$$

Also $E(\widehat{\varphi}-\varphi)_a(\widehat{\varphi}-\varphi)_b(\widehat{\varphi}-\varphi)_c = n^{-2}\overline{K}'_{abc} + O(n^{-3})$, where $\overline{K}'_{abc} = \overline{K}_{abc} + \sum_{abc}^{3} V_{ab} C_{1c}$. Apply the Cramer-Wold device to the expression on page 580 of Withers [11].

Finally, replace (φ, f_N) by (φ_0, f_{oN}), where $f_{oN} = f_N - \overline{f}$, $\varphi_0 = (\alpha_n, \beta)$ and $\alpha_n = \alpha + \overline{f}(\beta)$; this is valid since $\alpha + f_N(\beta) = \alpha_n + f_{oN}(\beta)$. The result of the theorem holds since $\left\{ g = \begin{pmatrix} 1 & 0' \\ 0 & M \end{pmatrix} \right\}$ for this parameterisation. □

Note 7.1. *Note that the method of proof gives expansion for cumulants of $\widehat{\varphi}$ as power series in n^{-1}, and allows us to relax the i.i.d. conditions on $\{e_N\}$. Further details on the proofs are in Withers and Nadarajah [13] which treats the case of multivariate $\{Y_N\}$.*

Note 7.2. *The proof of Theorem 4.1 is the same except that $g_{ij\cdots}$ and $g_{i,jk}$ are replaced by their expectations.*

Proof of Theorem 6.1. This follows the proof above. Note $R_{i_0\cdots i_r}$ and $S_{hi_1\cdots i_r}$ are defined as above. Note $S_l = a^{li} R_i$ and $S_{hk} = a^{hj} R_{jk}$, so

$$\begin{aligned}
n\kappa\left(S_l, S_{hk}\right) &= a^{li} a^{hj} n\kappa\left(R_i, R_{jk}\right) \\
&= a^{li} a^{hj} n^{-1} \sum_{N=1}^{n} \kappa\Big(-\rho^{(1)}\left(e_N\right) g_{N.i}, \rho^{(2)}\left(e_N\right) g_{N.j} g_{N.k} \\
&\quad -\rho^{(1)}\left(e_N\right) g_{N.jk}\Big) \\
&= a^{li} a^{hj} n^{-1} \sum_{N=1}^{n} \left(-g_{N.i} g_{N.j} g_{N.k} \rho_{N12} + g_{N.i} g_{N.jk} \rho_{N11} \right) \\
&= a^{li} a^{hj} \left(-\varrho_{12} g_{ijk} + \varrho_{11} g_{i,jk} \right).
\end{aligned}$$

Similarly, $ER_{ijk} = -\varrho_3 g_{ijk} + \varrho_2 \sum_{ijk}^{3} g_{i,jk}$, so

$$C_{1h} = a^{jk} a^{hi} \left(-\varrho_{12} g_{ijk} + \varrho_{11} g_{j,ik} \right) - V_{jk} a^{hi} \left(\varrho_3 g_{ijk} - \sum_{ijk}^{3} \varrho_2 g_{i,jk} \right) /2.$$

Also $a_{ij} = \varrho_2 g_{ij}$ and

$$V_{ij} = n\ cov\left(S_i, S_j\right) = a^{ik} a^{jl} n\ cov\left(R_k, R_l\right) = a^{ik} a^{jl} \varrho_{11} g_{kl}.$$

Now write the last term for C_{1h} as

$$\left(a^{hi} \varrho_2 g_{i,jk} V_{jk} + a^{hi} \varrho_2 g_{j,ki} V_{jk} + \varrho_2 g_{k,ij} a^{hi} V_{jk} \right) /2.$$

So, the second plus fourth terms simplify to

$$a^{hi}\left\{\left(a^{jk}\varrho_{11}-V_{jk}\varrho_2\right)g_{j,ki}-V_{jk}\varrho_2 g_{i,jk}/2\right\},$$

while the first plus third terms add to $a^{hi}\{a^{jk}(-\varrho_{12}+V_{jk}a^{jk}\varrho_3)g_{ijk}$ since $V_{jk} = a^{jl}(\varrho_{11}g_{lm})a^{mk}$. This proves (6.2).

Put $f_{h.h}=\partial f_h/\partial S_h=-1$, $f_{h.i,hi}=\partial^2 f_h/\partial S_i\partial S_{hi}=1$ and $f_{h.i,j}=\partial^2 f_h/\partial S_i\partial S_j = -ES_{hij}$. So,

$$\begin{aligned} f_{a.ik}f_{b.j}f_{c.e}\kappa^{ij}\kappa^{kl} &= f_{a.ik}\; n\kappa\left(\widehat{\theta}_i,S_b\right)n\kappa\left(\widehat{\theta}_k,S_c\right) \\ &= n\kappa\left(S_i,S_b\right)n\kappa\left(S_{ai},S_c\right)-\left(ES_{ajk}\right)n\kappa\left(S_j,S_b\right)n\kappa\left(S_k,S_c\right) \\ &= V_{ib}n\kappa\left(S_{ai},S_c\right)-\left(ES_{ajk}\right)V_{jb}V_{kc}, \end{aligned}$$

and

$$n^2\kappa\left(S_a,S_b,S_c\right)=a^{ai}a^{bj}a^{ck}n^2\kappa\left(R_i,R_j,R_k\right)=-a^{ai}a^{bj}a^{ck}\varrho_{111}g_{ijk}.$$

So,

$$\begin{aligned} \overline{K}^{abc} &= a^{ai}a^{bj}a^{ck}\varrho_{111}g_{ijk}+\sum_{abc}^{3}\Bigg\{V_{kb}a^{ci}a^{ji}\left(-\varrho_{12}g_{ijk}+\varrho_{11}g_{i,jk}\right) \\ &\qquad -V_{jb}V_{kc}a^{ai}\left(-\varrho_3 g_{ijk}+\varrho_2\sum_{ijk}^{3}g_{i,jk}\right)/2\Bigg\}. \end{aligned}$$

Of these five terms, the first plus second simplifies to $a^{ai}a^{ck}(a^{bj})\varrho_{111}-3V_{bj}\varrho_{12})g_{ijk}$ and the fourth simplifies to $3a^{ai}V_{bj}V_{ck}\varrho_3 g_{ijk}/2$, so the first plus second plus fourth gives $a^{ai}\{a^{bj}a^{ck}\varrho_{111}-3V_{bj}a^{ck}\varrho_{12}+3V_{bj}V_{ck}\}g_{ijk}$. The third plus fifth terms give the last two of the three terms in (6.3). □

Acknowledgments

The authors would like to thank the Editor and the Associate Editor for careful reading and for their comments which greatly improved the paper.

References

[1] Box, M. J. (1971). Bias in non-linear estimation (with discussion). *Journal of the Royal Statistical Society* B **33** 171–201. MR0315827

[2] Clarke, G. P. Y. (1980). Moments of the least squares estimators in a nonlinear regression model. *Journal of the Royal Statistical Society* B **42** 227–237. MR0583361

[3] Easton, G. S. and Ronchetti, E. (1986). General saddelpoint approximations with applications to L statistics. *Journal of the American Statistical Association* **81** 420–430. MR0845879

[4] Hajek, J. and Sidak, Z. (1967). *Theory of Rank Tests*. Academic Press, New York. MR0229351

[5] Huber, P. J. (1981). *Robust statisitics*. Wiley, New York. MR0606374

[6] Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares indent estimators. *Annals of Mathematical Statistics* **40** 633–643. MR0238419

[7] Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester. MR2238141

[8] Portnoy, S. (1984). Asymptotic behaviour of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Annals of Statistics* **12** 1298–1309. MR0760690

[9] Withers, C. S. (1982a). The distribution and quantiles of a function of parameter estimates. *Annals of the Institute of Statistical Mathematics* A **34** 55–68. MR0650324

[10] Withers, C. S. (1982b). Second order inference for asymptotically normal random variables. *Sankhyā* **44** 19–27. MR0692891

[11] Withers, C. S. (1983). Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparametric confidence intervals. *Annals of Statistics* **11** 577–587. MR0696069

[12] Withers, C. S. (1987). Bias reduction by Taylor series. *Communications in Statistics—Theory and Methods* **16** 2369–2384. MR0915469

[13] Withers, C. S. and Nadarajah, S. (2009). Asymptotic properties of M-estimates. *Technical Report*, Applied Mathematics Group, Industrial Research Ltd., Lower Hutt, New Zealand.