

## APPROXIMATE MAXIMUM LIKELIHOOD ESTIMATION IN LINEAR REGRESSION\*

MICHAEL A. MAGDALINOS

*Department of Statistics and Information Science,  
The Athens University of Economics and Business,  
76, Patission Street, Athens 104 34, Greece*

(Received March 3, 1989; revised October 25, 1991)

**Abstract.** The application of the ML method in linear regression requires a parametric form for the error density. When this is not available, the density may be parameterized by its cumulants ( $\kappa_i$ ) and the ML then applied. Results are obtained when the standardized cumulants ( $\gamma_i$ ) satisfy  $\gamma_i = \kappa_{i+2}/\kappa_2^{(i+2)/2} = O(v^i)$  as  $v \rightarrow 0$  for  $i > 0$ .

*Key words and phrases:* Regression, maximum likelihood, non-normal errors, Edgeworth approximation.

### 1. Introduction and Summary

Consider the linear regression model

$$(1.1) \quad y_j = x_j' \beta + e_j, \quad 1 \leq j \leq n$$

where  $\beta \in \mathbb{R}^m$  and  $e_j$  are unobserved errors.

ASSUMPTION 1. The vector  $x_j' = (x_{j1}, \dots, x_{j(m-1)}, 1)$  is non-stochastic, its elements are bounded for all  $j$ , and the matrix

$$(1.2) \quad Q = \sum_{j=1}^n x_j x_j' / n$$

converges to a positive definite matrix as  $n \rightarrow \infty$ .

ASSUMPTION 2. The errors  $e_j$  are iid with zero mean, variance  $\sigma^2$ , and finite cumulants ( $\kappa_i$ ) up to the eighth order.

---

\* Research financed in part by the Research Center of the Athens University of Economics and Business.

ASSUMPTION 3. For some  $v > 0$ , the standardized cumulants

$$(1.3) \quad \gamma_i = \kappa_{i+2}/\sigma^{i+2} = O(v^i) \quad \text{as } v \rightarrow 0, \quad i \geq 1.$$

Note that  $\gamma_1$  is the skewness and  $\gamma_2$  is the kurtosis coefficient of the error distribution. Moreover, the density function of the standardized errors

$$(1.4) \quad u_j = e_j/\sigma$$

admits a valid Edgeworth expansion. (Assumption 3 can be easily justified when the dependent variable  $y_j$  is an aggregate quantity (or its average), as in the example presented in Section 4. Then the error  $e_j$  in (1.1) can be written as a scalar multiple of the standardized sum of the errors committed by the  $N$  individuals over which the aggregation is taken. Under mild regularity conditions the density of this sum admits an Edgeworth expansion and (1.3) is satisfied for  $v = N^{-1/2}$  (Bhattacharya and Rao ((1976), p. 194)). For the general case a justification of Assumption 3 can be given through the theory of elementary errors (Edgeworth (1905)). The regression errors are often said to represent the impact of the sum of a large number of stochastic variables (elementary errors), each of which is itself insufficiently important to include in the regression (see, e.g. Cramer ((1946), p. 231), Johnston ((1972), p. 10)). Again, we can show that the density of this sum admits an Edgeworth expansion and that (1.3) is satisfied for  $v = N^{-1/2}$ , where  $N$  is the number of elementary errors.)

In Section 2 we prove the following results.

LEMMA 1.1. *If the Assumptions 1 to 3 are satisfied, then the log likelihood function of observations is*

$$L(\theta) = L^*(\theta) + O_p(v^3) \quad \text{as } v \rightarrow 0$$

where

$$(1.5) \quad L^*(\theta) = -(n/2) \ln(2\pi) + \sum_{j=1}^n l_j(\theta),$$

and

$$(1.6) \quad l_j(\theta) = -\frac{1}{2} \left[ \ln \sigma^2 + u_j^2 - \frac{1}{3} \gamma_1 (u_j^3 - 3u_j) \right. \\ \left. + \frac{1}{6} \gamma_1^2 (3u_j^2 - 2) + \frac{1}{4} (\delta^2 - 4) (u_j^4 - 6u_j^2 + 3) \right],$$

$$(1.7) \quad \delta^2 = \gamma_1^2 - \frac{1}{3} \gamma_2 + 4 < \frac{2}{3} (\gamma_1^2 + 7),$$

$$\begin{aligned} u_j &= (y_j - x'_j \beta) / \sigma, \\ \theta' &= (\beta', \sigma, \gamma_1, \delta). \end{aligned}$$

Usually, the maximum likelihood estimator is defined as the solution of the normal equations  $\partial L / \partial \theta = 0$ , and its asymptotic properties depend crucially on the fact that, under the usual regularity conditions, the expectation of the score vector  $\partial L / \partial \theta$  is zero. In our case, however, the function (1.5) is only an approximation of the true likelihood, so the expectation of the score vector is not zero.

We shall correct this deficiency by defining the approximate maximum likelihood (AML) estimator of  $\theta$  as the solution of the equations

$$(1.8) \quad s(\theta) = \frac{\partial L^*}{\partial \theta} - E \left[ \frac{\partial L^*}{\partial \theta} \right] = 0,$$

where  $s(\theta)$  is the approximate score vector (ASV).

**THEOREM 1.1.** *If the Assumptions 1 to 3 are satisfied, then the AML estimator of  $\theta$  is the solution of the equations*

$$(1.9) \quad s_i(\theta) = \sum_{j=1}^n s_{ij}(\theta) = 0 \quad (i = 1, \dots, 4)$$

where,

$$(1.10) \quad \begin{aligned} s_{1j} &= \frac{1}{2\sigma} x_j [(\gamma_1^2 + 2)u_j - \gamma_1(u_j^2 - 1) + (\delta^2 - 4)(u_j^3 - 3u_j - \gamma_1)], \\ s_{2j} &= \frac{1}{2\sigma} [2(u_j^2 - 1) - \gamma_1(u_j^3 - u_j) + \gamma_1^2 u_j^2 + (\delta^2 - 4)(u_j^4 - 3u_j^2 - \gamma_2)], \\ s_{3j} &= \frac{1}{6} [u_j^3 - 3u_j - \gamma_1(3u_j^2 - 2)], \\ s_{4j} &= \frac{1}{4} \delta (\gamma_2 - u_j^4 + 6u_j^2 - 3). \end{aligned}$$

We define the approximate information matrix (AIM) as

$$(1.11) \quad J(\theta) = E \left[ -\frac{1}{n} \frac{\partial s}{\partial \theta'} \right].$$

Also, we define the  $4 \times 4$  matrix  $D$  with elements

$$(1.12) \quad \begin{cases} d_{11} = (1 + \gamma_1^2) / 2\sigma^2, & d_{21} = \gamma_1(2\delta^2 - 9) / \sigma^2, \\ d_{31} = 0, & d_{41} = -\gamma_1 \delta / \sigma, \\ d_{12} = \gamma_1(3\delta^2 - 14) / 2\sigma^2, & d_{22} = [4 - \gamma_1^2 + 2(\delta^2 - 4)(2\gamma_2 + 3)] / 2\sigma^2, \\ d_{32} = -\gamma_1 / 2\sigma, & d_{42} = -\delta \gamma_2 / \sigma, & d_{13} = (\delta^2 - 4) / 2\sigma, \\ d_{23} = \gamma_1(6\delta^2 - 25) / 2\sigma, & d_{33} = 1/6, & d_{43} = -3\gamma_1 \delta / 2, & d_{14} = 0, \\ d_{24} = -3\delta(\delta^2 - 4) / \sigma, & d_{34} = 0, & d_{44} = 6\delta^2 / 4, \end{cases}$$

and we partition the  $m \times m$  matrices  $Q$  and  $Q^{-1}$  as

$$(1.13) \quad Q = \begin{bmatrix} Q_{11} & q'_1 \\ q_1 & 1 \end{bmatrix}, \quad Q^{-1} = \begin{bmatrix} G_{11} & g'_1 \\ g_1 & g \end{bmatrix}$$

where  $Q_{11}$  and  $G_{11}$  are  $(m-1) \times (m-1)$ ,  $q_1$  and  $g_1$  are  $(m-1) \times 1$ , and  $g$  is scalar. Notice that, if we write

$$x'_j = (z'_j, 1), \quad z. = \sum_{j=1}^n z_j/n$$

then the matrix  $G_{11}$  can be written in the familiar form

$$G_{11} = \left[ \sum_{j=1}^n (z_j - z.)(z_j - z.)' / n \right]^{-1}.$$

COROLLARY 1.1. *An analytic form for the AIM is*

$$(1.14) \quad J(\theta) = \begin{bmatrix} d_{11}Q_{11} & q_1d'_1 \\ d_2q'_1 & D \end{bmatrix}$$

where  $d'_1$  is the first row, and  $d_2$  is the first column of the matrix  $D$ , and

$$(1.15) \quad J^{-1}(\theta) = d_{11}^{-1} \begin{bmatrix} G_{11} & g_1i' \\ ig'_1 & d_{11}D^{-1} + (g-1)ii' \end{bmatrix}$$

where  $i = (1, 0, 0, 0)'$ .

The simple form of the inverse AIM suggests that an efficient algorithm for solving the equations (1.9) can be based on the method of scoring

$$(1.16) \quad \hat{\theta}_{r+1} = \hat{\theta}_r + J^{-1}(\hat{\theta}_r)s(\hat{\theta}_r)/n \quad (r = 1, 2, \dots).$$

The  $m \times m$  matrix  $Q$  needs to be inverted only once. Each iteration of the algorithm (1.16) requires only the inversion of the  $4 \times 4$  matrix  $D$ , so the algorithm is very fast even for large equations. A natural starting value for the algorithm is the LS estimator of  $\theta$ . Let  $\hat{\beta}$  and  $\hat{\sigma}^2$  be the usual LS estimators of  $\beta$  and  $\sigma^2$ . We define

$$(1.17) \quad \begin{cases} \tilde{u}_j = (y_j - x'_j\hat{\beta})/\hat{\sigma}, \\ \tilde{\gamma}_1 = \sum_{j=1}^n \tilde{u}_j^3/n, \quad \tilde{\gamma}_2 = \left( \sum_{j=1}^n \tilde{u}_j^4/n \right) - 3, \\ \tilde{\delta}^2 = \max\{0, \tilde{\gamma}_1^2 - \tilde{\gamma}_2/3 + 4\}, \quad \tilde{\theta} = (\tilde{\beta}', \tilde{\sigma}, \tilde{\gamma}_1, \tilde{\delta})'. \end{cases}$$

Also note that the one-step estimate  $\hat{\theta}_2$  will have the same asymptotic covariance matrix with the AML estimator as  $n \rightarrow \infty$ .

Following the standard maximum likelihood arguments, we can show that, as  $v \rightarrow 0$ , the AML estimator is consistent and asymptotically efficient to the order  $O(v^2)$ . This result, however, is of theoretical interest only and it will not be proved here. Having motivated the AML with Assumption 3, we now drop this assumption and consider its behaviour as  $n \rightarrow \infty$ . In that case the model is not necessarily identified, so we need an additional assumption.

**ASSUMPTION 4.** The matrix  $D$ , defined in (1.12), is non-singular in a neighborhood of the true values.

We define the  $4 \times 5$  matrix  $F$  with elements

$$(1.18) \quad \begin{cases} f_{11} = \gamma_1(5 - \delta^2)/2\sigma, & f_{12} = (\gamma_1 - 3\delta^2 + 14)/2\sigma, \\ f_{13} = -\gamma_1/2\sigma, & f_{14} = (\delta^2 - 4)/2\sigma, & f_{15} = 0, \\ f_{21} = (4\gamma_2 - \gamma_2\delta^2 - 2)/2\sigma, & f_{22} = \gamma_1/2\sigma, \\ f_{23} = (\gamma_1^2 - 3\delta^2 + 14)/2\sigma, & f_{24} = -\gamma_1/2\sigma, & f_{25} = (\delta^2 - 4)/2\sigma, \\ f_{31} = \gamma_1/3, & f_{32} = -1/2, & f_{33} = -\gamma_1/2, & f_{34} = 1/6, & f_{35} = 0, \\ f_{41} = \delta(\gamma_2 - 3)/4, & f_{42} = 0, \\ f_{43} = 3\delta/2, & f_{44} = 0, & f_{45} = -\delta/4, \end{cases}$$

and the  $5 \times 5$  matrix  $M$  with elements

$$(1.19) \quad \mu_{ij} = \mu_{i+j-2}, \quad \mu_k = E(u_1^k), \quad (k = 0, 1, \dots, 8).$$

Obviously,

$$(1.20) \quad \mu_0 = 1, \quad \mu_1 = 0, \quad \mu_2 = 1, \quad \mu_3 = \gamma_1, \quad \mu_4 = \gamma_2 + 3,$$

whereas the moments  $\mu_5$  to  $\mu_8$  depend on the higher order cumulants of  $u_1$ .

Also, we define the  $(m+7) \times 1$  vector

$$\theta^+ = (\theta', \mu_5, \mu_6, \mu_7, \mu_8)'$$

and the  $5 \times 5$  matrix

$$(1.21) \quad C = F M F'.$$

We write  $c_1$  for the first column and  $c_{11}$  for the leading element of  $C$ .

**LEMMA 1.2.** *If Assumptions 1 and 2 are satisfied, then as  $n \rightarrow \infty$*

$$(1.22) \quad s(\theta)/n^{1/2} \xrightarrow{d} N(0, \lim \Phi(\theta^+))$$

where,

$$(1.23) \quad \Phi(\theta^+) = \begin{bmatrix} c_{11}Q_{11} & q_1c_1' \\ c_1q_1' & C \end{bmatrix}.$$

THEOREM 1.2. *If Assumptions 1, 2 and 4 are satisfied, then as  $n \rightarrow \infty$*

$$(1.24) \quad n^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \lim V(\theta^+))$$

where,

$$(1.25) \quad V(\theta^+) = (c_{11}/d_{11}^2) \begin{bmatrix} G_{11} & g_1 i' \\ i g_1' & d_{11}^2 D^{-1} C D'^{-1} / c_{11} + (g-1) i i' \end{bmatrix}.$$

Moreover, it is easy to show that the estimates

$$\hat{\mu}_k = \frac{1}{n} \sum_{j=1}^n (y_j - x_j' \hat{\beta})^k / \hat{\sigma}^k, \quad (k = 5, \dots, 8)$$

are consistent, so that  $V(\hat{\theta}^+)$  is a consistent estimate of  $V(\theta^+)$ .

As usual, the AML has bias of order  $O(n^{-1})$  (see Section 3), so the contribution of the bias to the mean square error of the AML estimator is of order  $O(n^{-2})$ , i.e. asymptotically negligible compared with the contribution of the variance, which by Theorem 1.2 is of order  $O(n^{-1})$ . As the variance of the LS estimator is also of order  $O(n^{-1})$ , we may use the ratio of the two variances to define the relative asymptotic efficiency. In particular, Theorem 1.2 implies that, as  $n \rightarrow \infty$ , the asymptotic relative efficiency of the AML estimator with respect to the LS estimator is given by

$$(1.26) \quad e = d_{11}^2 \sigma^2 / c_{11}.$$

COROLLARY 1.2. *If Assumptions 1, 2 and 4 are satisfied, then*

$$(1.27) \quad e = (\gamma_1^2 + 2)^2 / [4 + 2\gamma_1^2 - (6\gamma_1^2 - 25\gamma_1^2\gamma_2 + 7\gamma_2^2)/3 - \delta_2(3\gamma_1^4 + 11\gamma_1^2\gamma_2 - 6\gamma_2^2)/3 - \delta_2(2\gamma_1\mu_5 - \delta_2\mu_6)],$$

where,

$$\delta_2 = \delta^2 - 4 = \gamma_1^2 - \gamma_2/3.$$

When the error distribution is normal,  $\gamma_1 = \gamma_2 = \delta_2 = 0$ , so that  $e = 1$ .

Small departures from the normality assumption imply that the AML estimator is likely to be more efficient than the LS estimator. In particular, expanding (1.27) under Assumption 3, we have

$$(1.28) \quad e = 1 + \gamma_1^2/2 + O(v^4) \quad \text{as} \quad v \rightarrow 0,$$

and, if  $\gamma_1 = 0$

$$(1.29) \quad e = 1/(1 - \gamma_2^2/6) + O(v^6) \quad \text{as} \quad v \rightarrow 0.$$

Moreover, it is easy to show that, when  $e > 1$ , the tests of hypotheses on the slope parameters  $(\beta_1, \dots, \beta_{m-1})$  are asymptotically more powerful when calculated from the AML estimator than when calculated from the LS estimator. In particular, the length of the AML confidence interval for the slope parameters is equal to the length of the corresponding LS confidence interval divided by  $\sqrt{e}$ .

2. Proofs

PROOF OF LEMMA 1.1. Assumption 3 implies that the density of standardized errors  $u_j = e_j/\sigma$  is

$$(2.1) \quad f(x) = (2\pi)^{-1/2} \exp(-x^2/2)[1 + p(x)] + O(v^3) \quad \text{as } v \rightarrow 0,$$

where

$$p(x) = \frac{1}{6}\gamma_1(x^3 - 3x) + \frac{1}{24}\gamma_2(x^4 - 6x^2 + 3) + \frac{1}{72}\gamma_1^2(x^6 - 15x^4 + 45x^2 - 15)$$

(see, e.g. Cramer ((1946), p. 229)). Taking logarithms in (2.1) we find

$$(2.2) \quad \begin{aligned} \ln f(x) &= \frac{1}{2} \ln(2\pi) - \frac{1}{2}x^2 + \ln[1 + p(x)] + O(v^3) \\ &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2}x^2 + p(x) - \frac{1}{2}[p(x)]^2 + O(v^3) \\ &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2}x^2 + \frac{1}{6}\gamma_1(x^3 - 3x) \\ &\quad + \frac{1}{24}[\gamma_2(x^4 - 6x^2 + 3) - \gamma_1^2(3x^4 - 12x^2 + 5)] + O(v^3). \end{aligned}$$

Following Kendal and Stuart ((1977), p. 95) we can show that

$$(2.3) \quad \gamma_1^2 - 2 < \gamma_2 < 3(\gamma_1^2 + 4) + O(v^4).$$

In order to incorporate<sup>1)</sup> these restrictions into the expansion (2.2) we may consider a  $\delta \in \mathbb{R}$ , such that

$$\gamma_2 = 3(\gamma_1^2 + 4 - \delta^2) + O(v^4),$$

which, substituted in (2.2) implies

$$(2.4) \quad f(x) = (2\pi)^{-1/2} \exp \left[ -\frac{1}{2}q(x) \right] + O(v^3)$$

where

$$(2.5) \quad q(x) = x^2 - \frac{1}{3}\gamma_1(x^3 - 3x) + \frac{1}{6}\gamma_1^2(3x^2 - 2) + \frac{1}{4}(\delta^2 - 4)(x^4 - 6x^2 + 3).$$

Then, the density of the disturbances  $e_j = \sigma u_j$  is

$$(2.6) \quad f_*(x) = (2\pi\sigma^2)^{-1/2} \exp \left[ -\frac{1}{2}q(x/\sigma) \right] + O(v^3),$$

---

<sup>1)</sup> We can show that a necessary condition for identification, i.e. for the validity of Assumption 4, is that the inequalities (2.3) hold exactly. This is the reason for using the parameter  $\delta$  as defined in (1.7) instead of the more natural parameter  $\gamma_2$ .

which implies (1.5), since the residual term is a polynomial with coefficients  $O(v^3)$  in the variables  $u_t = O_p(1)$  as  $v \rightarrow 0$ .

PROOF OF THEOREM 1.1. From the definition of  $u_j$ , we have that for  $k = 0, 1, \dots$

$$(2.7) \quad \begin{aligned} \partial u_j^k / \partial \beta &= -k x_j u_j^{k-1} / \sigma, & \partial u_j^k / \partial \sigma &= -k u_j^k / \sigma, \\ \partial (u_j^k / \sigma) / \partial \sigma &= -(k+1) u_j^k / \sigma^2 \end{aligned}$$

so that,

$$\begin{aligned} \partial l_j / \partial \beta &= x_j [2u_j - \gamma_1 (u_j^2 - 1) + \gamma_1^2 u_j + (\delta^2 - 4)(u_j^3 - 3u_j)] / 2\sigma, \\ \partial l_j / \partial \sigma &= [-2 + 2u_j^2 - \gamma_1 (u_j^3 - u_t) + \gamma_1^2 u_j^2 + (\delta^2 - 4)(u_j^4 - 3u_j^2)] / 2\sigma, \\ \partial l_j / \partial \gamma_1 &= [u_j^3 - 3u_j - \gamma_1 (3u_j^2 - 2)] / 6, \\ \partial l_j / \partial \delta &= -\delta (u_j^4 - 6u_j^2 + 3) / 4. \end{aligned}$$

Taking expectations and subtracting we find (1.10).

PROOF OF COROLLARY 1.1. Using (2.7) and

$$\partial \gamma_2 / \partial \gamma_1 = 6\gamma_1, \quad \partial \gamma_2 / \partial \delta = -6\delta,$$

we find

$$(2.8) \quad -\frac{1}{n} \frac{\partial s}{\partial \theta'} = \frac{1}{n} \sum_{j=1}^T \begin{bmatrix} x_j x'_j s_j^{11} & x_j s_j^{12} & x_j s_j^{13} & x_j s_j^{14} \\ x'_j s_j^{21} & s_j^{22} & s_j^{23} & s_j^{24} \\ x'_j s_j^{31} & s_j^{32} & s_j^{33} & s_j^{34} \\ x'_j s_j^{41} & s_j^{42} & s_j^{43} & s_j^{44} \end{bmatrix}$$

where

$$(2.9) \quad \begin{aligned} s_j^{11} &= [\gamma_1^2 + 2 - 2\gamma_1 u_j + 3(\delta^2 - 4)(u_j^2 - 1)] / 2\sigma^2, \\ s_j^{21} &= [4u_j - \gamma_1 (3u_j^2 - 1) + 2\gamma_1^2 u_j + (\delta^2 - 4)(4u_j^3 - 6u_j)] / 2\sigma^2, \\ s_j^{31} &= (u_j^2 - 1 - 2\gamma_1 u_j) / 2\sigma, & s^{41} &= -\delta (u_j^3 - 3u_j) / \sigma, \\ s_j^{12} &= [2(\gamma_1^2 + 2)u_j - \gamma_1 (3u_j^2 - 1) + (\delta^2 - 4)(4u_j^3 - 6u_j - \gamma_1)] / 2\sigma^2, \\ s_j^{22} &= [2(3u_j^2 - 1) - \gamma_1 (4u_j^3 - 2u_j) \\ &\quad + 3\gamma_1^2 u_j^2 + (\delta^2 - 4)(5u_j^4 - 9u_j^2 - \gamma_2)] / 2\sigma, \\ s_j^{32} &= (u_j^3 - u_j - 2\gamma_1 u_j^2) / 2\sigma, & s_j^{42} &= -\delta (u_j^4 - 3u_j^2) / \sigma, \\ s_j^{13} &= (u_j^2 - 2\gamma_1 u_j + \delta^2 - 5) / 2\sigma, \\ s_j^{23} &= [u_j^3 - 2\gamma_1 u_j^2 - u_j + 6\gamma_1 (\delta^2 - 4)] / 2\sigma, \\ s_j^{33} &= (3u_j^2 - 2) / 6, & s_j^{43} &= -3\delta \gamma_1 / 2, & s_j^{14} &= -\delta (u_j^3 - 3u_j - \gamma_1) / \sigma, \\ s_j^{24} &= -\delta [u_j^4 - 3u_j^2 - \gamma_2 + 3(\delta^2 - 4)] / \sigma, & s_j^{34} &= 0, \\ s_j^{44} &= (u_j^4 - 6u_j^2 + 3 - \gamma_2 + 6\delta^2) / 4. \end{aligned}$$



The elements of the matrix  $D$  are

$$d_{ik} = E(s_j^{ik}) \quad (i, k = 1, \dots, 4)$$

and they are given in (1.12). Taking expectations in (2.8) and using the partitioning (1.13) of the matrix  $Q$  we prove (1.14). Let  $I_n$  denote the  $n \times n$  identity matrix,  $O_n$  the  $n \times 1$  zero vector, and  $O_{nm}$  the  $n \times m$  zero matrix. From  $QQ^{-1} = I_m$ , using the partitioning (1.13), we find

$$(2.10) \quad \begin{aligned} Q_{11}G_{11} + q_1g'_1 &= I_{m-1}, & q'_1G_{11} + g'_1 &= O'_{m-1}, \\ Q_{11}g_1 + q_1g &= O_{m-1}, & q'_1g_1 + g &= 1. \end{aligned}$$

Multiplying (1.14) and (1.15) we find

$$JJ^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where, if  $a = (g - 1)/d_{11}$

$$\begin{aligned} A_{11} &= Q_{11}G_{11} + q_1d'_1ig'_1/d_{11} = Q_{11}G_{11} + q_1g'_1 = I_{m-1}, \\ A_{21} &= d_2q'_1G_{11}/d_{11} + Dig'_1/d_{11} = d_2(q'_1G_{11} + g'_1)/d_{11} = O_{4(m-1)}, \\ A_{12} &= Q_{11}g_1i' + q_1d'_1(D^{-1} + ii'a) = (Q_{11}g_1 + q_1g)i' = O_{(m-1)4}, \\ A_{22} &= d_2q'_1g_1i'/d_{11} + D(D^{-1} + ii'a) \\ &= I_4 + (q'_1g_1 + d_{11}a)d_2i'/d_{11} = I_4. \end{aligned}$$

PROOF OF LEMMA 1.2. We write  $x'_j = (z'_j, 1)$ ,  $z'_j = (x_{j1}, \dots, x_{j(m-1)})$  and we define the  $(m+3) \times 4$  matrices  $Z_j$  and the  $5 \times 1$  vectors  $u_j^*$  as

$$Z_j = \begin{bmatrix} z_j & O_{(m-1)3} \\ & I_4 \end{bmatrix}, \quad u_j^* = (1, u_j, u_j^2, u_j^3, u_j^4)'$$

From (1.9) and (1.10) we have that the score vector can be written as

$$s(\theta) = \sum_{j=1}^n s_j, \quad s_j = (s'_{1j}, s_{2j}, s_{3j}, s_{4j})' = Z_j F u_j^*.$$

The random vectors  $s_j$  are independently distributed with zero mean and variance

$$\begin{aligned} \Phi_j &= E(s_j s'_j) = Z_j F E(u_j^* u_j^{*'}) F' Z'_j = Z_j F M F' Z'_j = Z_j C Z'_j \\ &= \begin{bmatrix} c_{11} z_j z'_j & z_j c'_1 \\ c_1 z'_j & C \end{bmatrix}. \end{aligned}$$

Since the third moments of  $s_j$  exist, the assumptions of Lindeberg-Feller central limit theorem (see, e.g., Spanos ((1986), p. 177)) are satisfied, and  $s(\theta)/n^{1/2}$

converges in distribution to a normal vector with zero mean and variance  $\Phi(\theta) = \sum_{j=1}^n \Phi_j/n$ .

PROOF OF THEOREM 1.2. Assumptions 1 and 2 together with (2.8) and (2.9) imply that the variances of the elements of the matrix  $\partial s/\partial\theta'$  are bounded, and Kolmogorov's law of large numbers implies that

$$(2.11) \quad -\frac{1}{n} \frac{\partial s}{\partial\theta'} \xrightarrow{\text{a.s.}} J(\theta)$$

(see, e.g., Spanos ((1986), pp. 170–171)). Therefore, the AML estimator is a point of attraction of the iteration (1.16) with probability 1 as  $n \rightarrow \infty$  (see, e.g., Ortega and Reheinboldt ((1970), p. 311)). On the other hand, it is easy to show that, if the iterations (1.16) start with a root  $n$  consistent estimator, then it converges to a root  $n$  consistent estimator. Consequently, the AML estimator is root  $n$  consistent, i.e.

$$(2.12) \quad n^{1/2}(\hat{\theta} - \theta) = O_p(1) \quad \text{as} \quad n \rightarrow \infty.$$

Moreover, expanding the normal equations around the true values  $\theta$ , and using (2.12) we find

$$s(\theta) + \frac{\partial s}{\partial\theta'}(\hat{\theta} - \theta) + O_p(n^{-1})$$

so that

$$(2.13) \quad n^{1/2}(\hat{\theta} - \theta) = \left[ -\frac{1}{n} \frac{\partial s}{\partial\theta'} \right]^{-1} s(\theta)/n^{1/2} + O_p(n^{-1}).$$

Lemma 1.2 and (2.13) imply the convergence (1.24), where

$$V(\theta^+) = J^{-1}(\theta)\Phi(\theta^+)J'^{-1}(\theta).$$

Using (1.13), (1.23) and (2.10) we can show that

$$J^{-1}\Phi = \begin{bmatrix} (c_{11}/d_{11})I_{(m-1)} & O_{(m-1)4} \\ -(c_{11}/d_{11})iq'_1 + D^{-1}c_1q'_1 & D^{-1}C \end{bmatrix}$$

and post-multiplying by the transpose of (1.13) we prove (1.25).

PROOF OF COROLLARY 1.2. The first row of the matrix  $F$  is

$$(2.14) \quad f_1 = \frac{1}{2\sigma}(\gamma_1(1 - \delta_2), \gamma_2 - 2\gamma_1^2 + 2, -\gamma_1, \delta_2)$$

so that

$$e = (\gamma_1^2 + 2)^2/4\sigma^2 f_1 M f'_1.$$

Substituting the definition (1.19) of  $M$  and making use of (1.20) we prove (1.27).

### 3. Estimating bias and testing for normality

Although the AML estimator is root  $n$  consistent, it is not unbiased. In this section we calculate the bias to the order  $O(n^{-1})$ . To simplify the notation, we shall use the summation convention: In any product of symbols, summation over an index is understood if that index appears as superscript and as subscript.

The  $r$ -th component of the score vector  $s$ , and its derivatives with respect to the elements of  $\theta$  ( $s_{rs}$ ,  $s_{rst}$ , say) are expressible as a sum of  $n$  independent and identical random variables, so their joint distribution may be approximated by the normal approximation. It is convenient to make the dependence on  $n$  explicit by writing

$$\begin{aligned} s_r &= n^{1/2} z_r, & s_{rs} &= n\kappa_{rs} + n^{1/2} z_{rs}, \\ s_{rst} &= n\kappa_{rst} + n^{1/2} z_{rst}, \end{aligned}$$

where  $\kappa_{rs}$ ,  $\kappa_{rst}$  are the expectations of  $s_{rs}/n$ ,  $s_{rst}/n$ , respectively. From the normal approximation follows that  $z_r$ ,  $z_{rs}$ ,  $z_{rst}$  are of order  $O_p(1)$  as  $n \rightarrow \infty$ . The likelihood equations (1.9) may be expanded in a Taylor series in the components of the vector  $d = n^{1/2}(\hat{\theta} - \theta)$ . Using the summation convention

$$0 = s_r + s_{rs}d^s/n^{1/2} + s_{rst}d^s d^t/2n + \dots$$

Multiplying by  $n^{-1/2}$  and collecting terms of the same order, we find

$$(3.1) \quad 0 = z_r + \kappa_{rs}d^s + n^{-1/2}(z_{rs}d^s + \kappa_{rst}d^s d^t/2) + O_p(n^{-1}).$$

Note that  $-\kappa_{rs}$  is the  $(r, s)$ -th element of the matrix (1.14). We write  $\kappa^{rs}$  for the  $(r, s)$ -th element of its inverse (1.15). From (3.1) we have  $d^i = \kappa^{ij}z_j + O_p(n^{-1/2})$ , which substituted back in (3.1) implies

$$d^r = \kappa^{rs}z_s + n^{-1/2}(\kappa^{rs}\kappa^{tu}z_{st}z_u + \kappa^{rs}\kappa_{sij}\kappa^{iu}\kappa^{jv}z_u z_v/2) + O_p(n^{-1}).$$

Taking expectations, we prove the following theorem:

**THEOREM 3.1.** *If the Assumptions 1, 2 and 4 are satisfied, then the bias of the  $r$ -th element AML estimator is*

$$(3.2) \quad E(\hat{\theta}^r - \theta^r) = \frac{1}{n}(\kappa^{rs}\kappa^{tu}\epsilon_{stu} + \kappa^{ruv}\phi_{uv}) + O(n^{-3/2}) \quad \text{as } n \rightarrow \infty,$$

where  $\kappa^{rs}$ ,  $\phi_{rs}$  are the  $(r, s)$ -th elements of the matrices (1.15) and (1.23) respectively,

$$\begin{aligned} \kappa^{rvu} &= \kappa^{rs}\kappa_{sij}\kappa^{iu}\kappa^{jv}/2, \\ \kappa_{vij} &= E\left[\frac{1}{n}\frac{\partial^2 s_v}{\partial\theta^i\partial\theta^j}\right], & \epsilon_{vij} &= E\left[\frac{1}{n}\frac{\partial s_v}{\partial\theta^i}s_j\right], \end{aligned}$$

and  $s_v$  is the  $v$ -th element of the score vector (1.8).

Using the same method as in the proof of Corollary 1.1, we can calculate the unknown quantities  $\kappa_{vij}$  and  $\epsilon_{vij}$  ( $v, i, j = 1, \dots, m+3$ ), which can be expressed

in terms of  $128 = 2 \times 4^3$  parameters of the form (1.12). These parameters are too numerous to be presented here, but they can be consistently estimated. Then substituting in (1.26) we have a numerical estimate of the bias, which can be used to correct the bias of the AML estimator with an error of order  $O(n^{-3/2})$ . However, caution should be exercised in applying this correction, as the size corrections usually increase the second order variance of the estimator (see, e.g., Cox and Hinkley (1974), p. 310).

An interesting hypothesis is that the error distribution is normal. In our parameterization, the test for normality is equivalent to testing the hypothesis

$$H_0 : \gamma_1 = 0 \quad \text{and} \quad \delta = 2.$$

The corresponding Wald test statistic is calculated from the AML estimator as

$$(3.3) \quad W = n\hat{d}'\hat{V}_{22}^{-1}\hat{d}$$

where  $\hat{d}' = (\hat{\gamma}_1, \hat{\delta} - 2)$  and  $V_{22}$  is the lower  $2 \times 2$  submatrix of  $D^{-1}CD^{-1}$ . The Lagrange multiplier statistic for testing  $H_0$  is calculated from the LS estimator as

$$(3.4) \quad \text{LM} = s(\tilde{\theta})'\Phi(\tilde{\theta}^+)s(\tilde{\theta})/n = n(\tilde{\gamma}_1^2/6 + \tilde{\gamma}_2^2/24),$$

where  $\tilde{\gamma}_1, \tilde{\gamma}_2$  are defined in (1.17)<sup>2</sup>. Under the normality assumption the statistics  $W$  and LM are asymptotically distributed as chi-square with two degrees of freedom.

After a preliminary LS estimation, the test (3.4) can be used to help the decision of whether or not to proceed and use the AML method. However, caution should be exercised in the interpretation of the test, because in this case we face a decision rather than a testing problem. The fact that we are not able to reject normality does not mean that the error distribution is exactly normal. Even small departures from the normality assumption (which the tests might not be able to detect) reduce drastically the efficiency of the LS estimator (see Hampel *et al.* ((1986), p. 309) and the references cited therein). In many situations, we may improve considerably the efficiency by using the AML method, even though we might not be able to reject the normality assumption. It seems reasonable to proceed to the AML estimation in all cases where the preliminary estimate of the relative asymptotic efficiency of the AML method is greater than one ( $\tilde{e} > 1$ ).

#### 4. Comparison with other estimators

Following Stone (1975), Bickel (1982), Manski (1984), Schick (1986), Bickel and Ritov (1987) and Kreiss (1987), among others, propose the use of non-parametric scoring (NPS) estimators, which use the residuals from a preliminary

---

<sup>2</sup>) Bera and Jarque (1982) using a Pearson family as a priori specification of the error density, derived the same test. A less formal derivation can be based on the fact that, under the assumption of normality,  $n^{1/2}\tilde{\gamma}_1$  and  $n^{1/2}\tilde{\gamma}_2$  are asymptotically distributed as independent  $N(0, 6)$  and  $N(0, 24)$  variables, respectively.

consistent estimation to nonparametrically estimate the score function in the one-step scoring method. The NPS estimators are adaptive, i.e. they are asymptotically fully efficient, under the assumption that the bandwidth of the nonparametric shape estimation is converging to zero as the sample size tends to infinity. In practical terms, the shrinking bandwidth corresponds to an increasing generosity of nuisance parameterization, so that the property of adaptation is less appealing than the property of asymptotic efficiency in the traditional finite dimensional parametric estimation. Moreover, the computational requirements of the NPS estimation are heavy, as they typically increase at the same rate as the square of the sample size. In empirical work, the most undesirable feature of the NPS estimation is the sensitivity of its performance to the choice of the bandwidth. As shown in Hsieh and Manski (1987) the application of NPS estimation with preselected bandwidth can lead to a very inefficient estimate, and should be avoided. On the other hand, the automatic methods of bandwidth selection (cross-validation, bootstrapping, etc.) produce data-dependent bandwidths, whereas the standard theory relies quite heavily on the assumption of exogenous bandwidth.

Similar are the problems encountered with other infinite dimensional parameterizations of the error density. For example, Gallant and Nychka (1987) and Gallant and Tauchen (1989) put the density equal to a Hermite (or similar) series, and they have proven that the structural parameters and many aspects of the unknown density will be estimated consistently, provided that the length of the series increases with the sample size. Again, the choice of the length of the series will be the crucial factor determining the performance of the method in finite samples. With fixed length, the approximation is entirely ad hoc, and the resulting estimates do not seem to have any optimal properties.

The main idea that motivated the AML method is that it is not really necessary to estimate consistently the unknown density function: A simple approximation can be used to reproduce the basic shape of the error distribution, and to improve the efficiency of LS. The crucial factor here is that the asymptotic properties of the resulting estimator should be independent of the validity of the approximation. This is a natural generalization of the corresponding LS property (Spanos (1986), p. 450) and allows comparison of the two estimators.

We have applied the AML method on 42 different economic data sets, half of which were cross-sectional, and half were time-series data. The greatest relative efficiency found was  $e = 3.57$  (in Johnston's example ((1972), p. 147)) and in 23 cases  $e$  was greater than 1.5. Only in one case  $e$  was less than one ( $e = 0.987$ ), and in two cases the algorithm (1.16) did not converge. The difference between the LS and the AML estimates of the structural parameters is relatively small, but often there are considerable differences in the estimates of the error parameters. The various test statistics can also differ considerably. In particular, the LM test for normality was in all cases less than the corresponding Wald test.

A typical example of the application of the AML method will be presented in more detail. Maddala ((1977), p. 116) gives data on per-capita food consumption ( $q_t$ ), a price index of food ( $p_t$ ), and deflated per-capita disposable income ( $y_t$ ) for U.S.A. and for the years  $t = 1927, \dots, 1941, 1948, \dots, 1963$ . The LS and the AML

Table 1. Empirical application: food consumption data.

Period	1927-1941, 1948-1963		1927-1941		1948-1963	
	LS	AML	LS	AML	LS	AML
ESTIMATOR						
Price	-0.0449 (.038)	-0.2160 (.035)	-0.2160 (.046)	-0.2219 (.037)	-0.2262 (.131)	-0.3097 (.101)
Income	0.2810 (.016)	0.2867 (.014)	0.3781 (.030)	0.3785 (.025)	0.1495 (.041)	0.1414 (.032)
Constant	76.953 (2.65)	77.729 (2.46)	86.083 (3.56)	86.573 (2.81)	107.94 (16.8)	117.16 (13.0)
$\sigma$	0.9848 (.146)	0.9871 (.099)	0.7557 (.108)	0.7562 (.094)	0.5802 (.109)	0.5996 (.076)
$\gamma_1$	0.0297 (.430)	0.0317 (.307)	-0.1323 (.633)	-0.1081 (.356)	0.0855 (.612)	0.3902 (.389)
$\gamma_2$	-0.5956 (.870)	-0.7422	-1.0431 (1.26)	-1.0758	-0.0910 (1.22)	-0.9707
$\delta$	2.0492	2.0611 (.027)	2.0893	2.0905 (.031)	2.0094	2.1156 (.048)
F test	990.7	1157.9	159.9	243.3	128.2	251.2
Norm. test	0.463	5.09	0.724	12.37	0.025	10.79
$\bar{R}^2$	0.967	0.967	0.899	0.900	0.872	0.864
Rel. Eff.	1.094	1.161	1.526	1.526	1.671	1.792

Note: The number in parentheses are standard errors. The LS estimates of standard errors for the error parameters ( $\sigma$ ,  $\gamma_1$ ,  $\gamma_2$ ) are correct under the assumption of normality. The Norm(ality) test is the LM test (1.31) in the LS estimation and the Wald test (1.30) in the AML estimation. Rel(ative) Eff(iciency) is the parameter (1.27) with moments and cumulants estimated from the LS or the AML method.  $\bar{R}^2$  is the proportion of the variation in the dependent variable "explained" by the regression and it is corrected for degrees of freedom.

estimates of the linear regression

$$q_t = \beta_1 p_t + \beta_2 y_t + \beta_0 + e_t,$$

are given in Table 1.

The first two columns of Table 1 present the LS and AML estimates respectively for the whole period. The efficiency increase (16%) is rather small, and this instigated the assumption that the error generating process might not be the same before and after the war. To test this assumption, we estimated separately the pre-war and post-war periods. We found that both the structural and the error parameters differ significantly between the two periods. The differences in the estimated error densities are presented in Fig. 1, where the dashed line is the LS normal approximation of the error density for the whole period, the lines 1, 2, and 3 are the AML exponential approximations for the whole period, the pre-war, and post-war periods, respectively.

It is clear that the pre-war and post-war densities (lines 2 and 3, respectively) differ, and that they are significantly different from the normal density. When

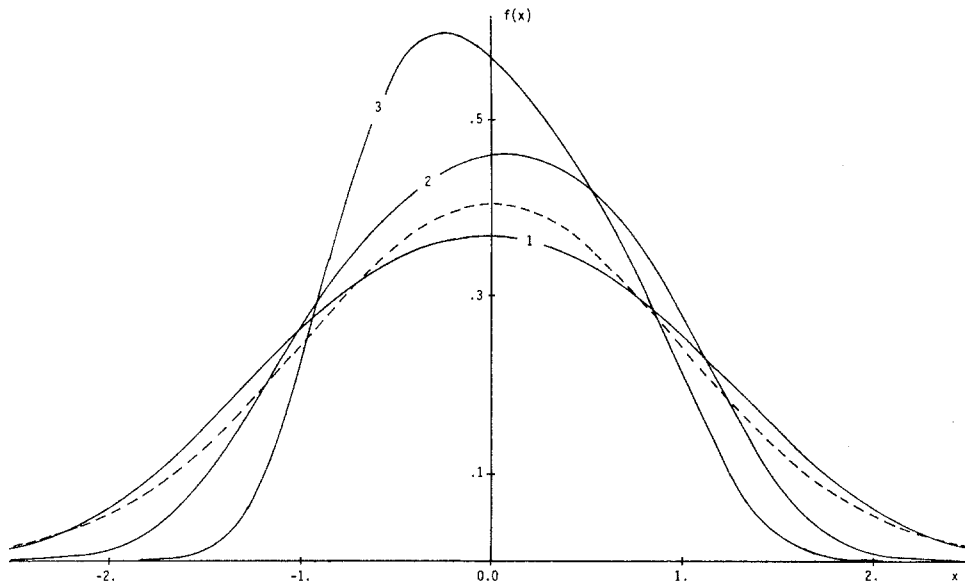


Fig. 1. Exponential approximation of the error densities: 1)  $\sigma^2 = .974275$ ,  $\gamma_1 = .031683$ ,  $\gamma_2 = -.742172$ . 2)  $\sigma^2 = .571918$ ,  $\gamma_1 = -.108149$ ,  $\gamma_2 = -1.075757$ . 3)  $\sigma^2 = .359550$ ,  $\gamma_1 = .390179$ ,  $\gamma_2 = -.970680$ . The dashed line stands for the  $N(0, .969849)$  density.

pooled together, however, they produce a density (line 1) that is not significantly different from the normal (dashed line). That explains the non-significance of the Wald test for normality for the whole period, as well as the small increase in efficiency for the whole sample.

#### Acknowledgements

The author is indebted to M. Drettakis, G. Phillips, and E. Kekalaki for helpful comments on an earlier version of this paper. The detailed comments and suggestions of two anonymous referees were greatly appreciated.

#### REFERENCES

- Bera, A. K. and Jarque, C. M. (1982). Model specification tests: A simultaneous approach, *J. Econometrics*, **20**, 59–82.
- Bhattacharya, R. N. and Rao, R. R. (1976). *Normal Approximations and Asymptotic Expansions*, Wiley, New York.
- Bickel, P. (1982). On adaptive estimation, *Ann. Statist.*, **10**, 647–671.
- Bickel, P. and Ritov, Y. (1987). Efficient estimation in the errors in variables model, *Ann. Statist.*, **15**, 513–540.
- Cox, D. R. and Hinkley, D. V. (1982). *Theoretical Statistics*, Chapman and Hall, London.
- Cramer, H. (1946). *Mathematical Methods of Statistics*, Princeton University Press, New Jersey.
- Edgeworth, F. Y. (1905). The law of the error, *Transactions of the Cambridge Philosophical Society*, **20**, 36–65, 113–141.

- Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation, *Econometrica*, **55**, 363–390.
- Gallant, A. R. and Tauchen, G. (1989). Semi-nonparametric estimation of conditionally constrained heterogenous processes, *Econometrica* (to appear).
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*, Wiley, New York.
- Hsieh, D. A. and Manski, C. F. (1987). Monte Carlo evidence on adaptive maximum likelihood estimation of a regression, *Ann. Statist.*, **15**, 541–551.
- Johnston, J. (1972). *Econometric Methods*, 2nd ed., McGraw-Hill Kogakusha, Tokyo.
- Kendal, M. and Stuart, A. (1977). *The Advanced Theory of Statistics*, Vol. 1, C. Griffin & Co., London.
- Kreiss, J. P. (1987). On adaptive estimation in stationary ARMA processes, *Ann. Statist.*, **15**, 112–133.
- Maddala, G. S. (1977). *Econometrics*, McGraw-Hill Kogakusha, Tokyo.
- Manski, C. F. (1984). Adaptive estimation of non-linear regression models, *Econometric Rev.*, **3**, 145–194.
- Ortega, J. M. and Reheinboldt, W. C. (1970). *Iterative Solution of Nonlinear Equation in Several Variables*, Academic Press, New York.
- Schick, A. (1986). On asymptotically efficient estimation in semiparametric models, *Ann. Statist.*, **14**, 1139–1151.
- Spanos, A. (1986). *Statistical Foundations of Econometric Modelling*, Cambridge University Press, Cambridge.
- Stone, C. J. (1975). Adaptive maximum likelihood estimation of a location parameter, *Ann. Statist.*, **3**, 267–284.