

The Equivalence of Generalized Least Squares and Maximum Likelihood Estimates in the Exponential Family

A. CHARNES, E. L. FROME and P. L. YU*

The method of iterative weighted least squares can be used to estimate the parameters in a nonlinear regression model. If the dependent variables are observations from a member of the regular exponential family, then under mild conditions it is shown that the IWLS estimates are identical to those obtained using the maximum likelihood principle. An application is provided to illustrate the results.

1. INTRODUCTION

Let Y_1, Y_2, \dots, Y_n be a random sample of size n drawn from a population with density $h[y_i; f(\mathbf{x}_i, \boldsymbol{\theta})]$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ ($i = 1, \dots, n$) are a set of known values of the independent variables, $\boldsymbol{\theta} = [\theta_1, \dots, \theta_m]'$ are unknown parameters, and $f(\mathbf{x}_i, \boldsymbol{\theta})$ is a known function of the independent variables and the parameters. The regression function, $f(\mathbf{x}_i, \boldsymbol{\theta})$, will in general be nonlinear (with respect to the parameters), and we assume that it is a differentiable function of $\boldsymbol{\theta}$. Given the observed values y_1, \dots, y_n , the problem is to obtain estimates of the parameters $\theta_1, \dots, \theta_m$. One approach to this problem is to use the least squares principle, i.e.,

$$\text{Min}_{\boldsymbol{\theta}} \sum_{i=1}^n V(Y_i)^{-1} [y_i - f(\mathbf{x}_i, \boldsymbol{\theta})]^2, \quad (1.1)$$

where $V(Y_i)$ is the variance of Y_i . When the $V(Y_i)$'s are independent of $\boldsymbol{\theta}$, this requires the solution of the system of equations

$$\sum_{i=1}^n V(Y_i)^{-1} [y_i - f(\mathbf{x}_i, \boldsymbol{\theta})] (\partial f(\mathbf{x}_i, \boldsymbol{\theta}) / \partial \theta_j) = 0, \quad j = 1, \dots, m. \quad (1.2)$$

If $f(\mathbf{x}, \boldsymbol{\theta})$ is linear (or concave) in $\boldsymbol{\theta}$ and $V(Y_i)$, $i = 1, \dots, n$ are known, then the solution of (1.2) yields a weighted least squares estimate of $\boldsymbol{\theta}$. When the regression function is nonlinear in the parameters, or the variances depend on $\boldsymbol{\theta}$ (i.e., $V(Y|\mathbf{x})$ is a known functional form in $\boldsymbol{\theta}$), a solution of (1.2) can be obtained using an iterative

weighted least squares (IWLS) procedure. The usual IWLS approach is to

- i. obtain an initial estimate of $\boldsymbol{\theta}$,
- ii. replace $f(\mathbf{x}_i, \boldsymbol{\theta})$ with a first-order Taylor series approximation,
- iii. evaluate all expressions that involve $\boldsymbol{\theta}$ at the current estimate (this includes the variances if they depend on $\boldsymbol{\theta}$),
- iv. solve the resulting linear system of equations for a correction vector, say $\boldsymbol{\delta}$,
- v. set $\boldsymbol{\theta}^k \leftarrow \boldsymbol{\theta}^{k-1} + \boldsymbol{\delta}$, and repeat (ii)-(v) until $\{\boldsymbol{\theta}^k\}$ converges.

The resulting IWLS estimate will not necessarily be a solution of (1.1), but under conditions described in Section 2 will yield a maximum of the likelihood function.

The IWLS computational procedure is familiar to the statistician, since it reduces to a linear weighted least squares problem on each iteration. Various approaches to the numerical solution of linear system of equations required in iv have been presented by Lawson and Hanson [5]. We note that if a singular system is encountered, the procedure can be modified by using a generalized inverse (see [1]) to obtain a solution—i.e., sequential generalized nonlinear least squares (SGNLS). SGNLS estimation is a considerably more difficult problem both theoretically and computationally and will not be considered here.

The other approach to the estimation problem that we consider employs the maximum likelihood (ML) principle. The log of the likelihood function is

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \log h[y_i; f(\mathbf{x}_i, \boldsymbol{\theta})], \quad (1.3)$$

and the ML estimates are the solution of the system of equations

$$\partial L(\boldsymbol{\theta}) / \partial \theta_j = 0 \quad j = 1, \dots, m. \quad (1.4)$$

It is well-known that if the Y_i 's are normally distributed and the regression function is linear in the parameters, then the ML and LS estimators are identical. Bradley [2] has established that when the density of the Y_i 's is in the regular exponential family and the regression model is linear, then the ML and IWLS estimates

* A. Charnes is university system professor, and director of The Center for Cybernetic Studies, E.L. Frome is assistant professor of statistics-operations research, and P.L. Yu is associate professor of statistics-operations research, all at The University of Texas at Austin, Austin, Tex. 78712. This research was partly supported by the Office of Naval Research Contract N00014-67-A-0126-0009 with the Center for Cybernetic Studies, The University of Texas at Austin. The authors wish to thank the associate editor and referees for their constructive comments and suggestions.

satisfy (1.2). Nelder and Wedderburn [8] have shown that this result is also true for certain generalized linear models. The equivalence of LS estimates and ML estimates when the regression function is nonlinear in the parameters has been established for the normal distribution by Turner, Monroe, and Lucas [9], the binomial distribution by Moore and Zeigler [7], and for the Poisson distribution by Frome, Kutner, and Beauchamp [4].

Theorem 1 of Section 2 extends Bradley's result [2] to the nonlinear case, and Theorem 4 gives conditions which guarantee that a solution of (1.2) will provide a global maximum of the likelihood function. This additional result (Theorem 4)—which is necessary to establish the equivalence of the ML and IWLS methods of estimation—was not provided by Bradley.

2. RESULTS

Suppose that Y is a random variable with a density function of the regular exponential family; i.e.,

$$h(y; \beta) = \exp \{p(\beta)y - q(\beta) + g(y)\}, \quad (2.1)$$

where $E(Y) = \beta$, $p(\beta)$ and $q(\beta)$ are at least twice differentiable, and the range of Y does not depend on β . In some applications an additional "nuisance parameter" (e.g., the variance of a normal distribution) would appear in (2.1). We assume here that this nuisance parameter is either known or is constant for all values of the independent variable. Then, following the approach used by Bradley [2], differentiation (with respect to β) on both sides of $\int h(y; \beta) dy = 1$ yields

$$E(Y) = q'(\beta)/p'(\beta) = \beta \quad (2.2)$$

where $p'(\beta)$ and $q'(\beta)$ denote the derivatives with respect to β of $p(\beta)$ and $q(\beta)$. A second differentiation of the integral, along with evaluation of the derivative of (2.2), results in

$$V(Y) = p''(\beta)^{-1}. \quad (2.3)$$

The following theorem generalizes the results that were presented by Bradley [1].

Theorem 1: If Y_1, Y_2, \dots, Y_n is a random sample of size n from (2.1) with $E(Y_i) = f(\mathbf{x}_i, \theta)$, then a ML estimate of θ will satisfy (1.2), provided it is an interior point of Θ (the parameter space).

Proof: The log of the likelihood function—neglecting a constant that does not involve the θ 's—is

$$L = \sum_{i=1}^n p[f(\mathbf{x}_i, \theta)]y_i - \sum_{i=1}^n q[f(\mathbf{x}_i, \theta)]. \quad (2.4)$$

The ML estimates of θ are obtained by solving the likelihood equations (1.4), i.e.,

$$\begin{aligned} \partial L / \partial \theta_j &= \sum_i p'[f(\mathbf{x}_i, \theta)](\partial f(\mathbf{x}_i, \theta) / \partial \theta_j) y_i \\ &- \sum_i q'[f(\mathbf{x}_i, \theta)](\partial f(\mathbf{x}_i, \theta) / \partial \theta_j) = 0, \quad j = 1, \dots, m. \end{aligned}$$

By using (2.2) and (2.3), we obtain

$$\begin{aligned} \partial L / \partial \theta_j &= \sum_i p'[f(\mathbf{x}_i, \theta)]\{y_i - f(\mathbf{x}_i, \theta)\}(\partial f(\mathbf{x}_i, \theta) / \partial \theta_j) \\ &= \sum_i \{V(Y_i)^{-1}[y_i - f(\mathbf{x}_i, \theta)](\partial f(\mathbf{x}_i, \theta) / \partial \theta_j)\} = 0, \\ & \quad j = 1, \dots, m. \end{aligned} \quad (2.5)$$

Since (2.5) and (1.2) are the same, the proof is complete.

This result demonstrates that a solution of (1.2) will yield a critical point of the likelihood function. If the IWLS procedure converges to a stable solution (convergence is not guaranteed), will it be an optimal global solution to the maximization problem? The following theorem indicates conditions which guarantee that a solution of (1.2) will in fact be a ML estimate of θ .

Theorem 2: Let $L(\theta)$ be defined over a convex set Θ . If (i) $\hat{\theta} \in \Theta$ is a solution to (2.5), (ii) both $p[f(\mathbf{x}, \theta)]$ and $-q[f(\mathbf{x}, \theta)]$ are concave¹ in θ over Θ , and (iii) the y_i 's are nonnegative, then $\hat{\theta}$ is a global maximum of $L(\theta)$ over Θ . It is the unique global solution if at least one of the $p[f(\mathbf{x}_i, \theta)]y_i$, $-q[f(\mathbf{x}_i, \theta)]$, $i = 1, \dots, n$ is strictly concave over Θ .

Proof: L is concave if $p[f(\mathbf{x}, \theta)]$ and $-q[f(\mathbf{x}, \theta)]$ are, and the y_i 's are nonnegative. L is strictly concave if at least one of the $p[f(\mathbf{x}_i, \theta)]y_i$, $-q[f(\mathbf{x}_i, \theta)]$, $i = 1, \dots, n$ is. For a differentiable concave function, a point at which the gradient vanishes is a global maximum. Further, the maximum point of a strictly concave function is unique whenever it exists.

Remark: Sufficient conditions for $\hat{\theta}$ to be a global maximum of $L(\theta)$ over Θ are that (i) $L(\theta)$ be pseudoconcave (see [6]) over Θ , and (ii) that $\hat{\theta}$ satisfies (2.5). It is, however, difficult to verify pseudoconcavity without conditions similar to those in Theorem 2.

Theorem 3: Let $L(\theta)$ be defined over a set Θ which has a nonempty interior. Suppose that there is a $\theta^* \in \Theta$ and positive numbers M and ϵ such that (i) $L(\theta) \leq L(\theta^*) - \epsilon$ whenever $\|\theta\| > M$ and (ii) for every sequence $\{\theta^k\}$ of Θ which converges to a boundary point of Θ , $\lim_{k \rightarrow \infty} L(\theta^k) \leq L(\theta^*) - \epsilon$. Then there is a global maximum point of $L(\theta)$ on Θ . Each such maximum point satisfies the likelihood equations (2.5).

Proof: Define $\Theta^* = \{\theta \in \Theta \mid L(\theta) \geq L(\theta^*)\}$. Note that Θ^* is nonempty and is in the interior of Θ . Observe that Θ^* is closed—because Θ^* is interior to Θ and $L(\theta)$ is continuous over Θ —and bounded—this follows from (ii). Consequently, Θ^* is compact. Since $L(\theta)$ is continuous over Θ^* it has a maximum point, which is in turn a maximum point of $L(\theta)$ over Θ . Finally, it follows that since a maximum point must be an interior point of Θ , it must satisfy (2.5).

Theorems 2 and 3 can be combined as follows.

¹ If $f(\mathbf{x}, \theta)$ is concave in θ and $p(x)$ is concave and nondecreasing in x , then $p[f(\mathbf{x}, \theta)]$ is concave in θ . This fact may be useful in verifying the concavity of $p[f(\mathbf{x}, \theta)]$ and $-q[f(\mathbf{x}, \theta)]$.

Theorem 4: Under the assumptions of Theorems 2 and 3, $L(\theta)$ has a nonempty set of global maxima which coincides with the solution set of (2.5), i.e., the ML estimates and the IWLS estimates are identical.

3. APPLICATION

Suppose that Y_1, \dots, Y_n is a random sample from the Poisson distribution, and that the regression function is positive and linear in the parameters, i.e., $f(\mathbf{x}_i, \theta) = \sum_{j=1}^m x_{ij}\theta_j > 0$. In this situation the log of the likelihood function is

$$L(\theta) = \sum_{i=1}^n \log \left(\sum_{j=1}^m x_{ij}\theta_j \right) y_i - \sum_{i=1}^n \sum_{j=1}^m x_{ij}\theta_j, \quad (3.1)$$

and the likelihood equations (2.5) reduce to

$$\frac{\partial L}{\partial \theta_k} = \sum_{i=1}^n \left[y_i \left(\sum_{j=1}^m x_{ij}\theta_j \right)^{-1} - 1 \right] x_{ik} = 0, \quad k = 1, \dots, m. \quad (3.2)$$

If (3.2) has a solution, it will be a global maximum (from the first part of Theorem 2). If the matrix \mathbf{C} with entries

$$C_{rk} = - \frac{\partial^2 L}{\partial \theta_r \partial \theta_k} = \sum_{i=1}^n y_i x_{ir} x_{ik} \left[\sum_{j=1}^m x_{ij}\theta_j \right]^{-2}, \quad r, k = 1, \dots, m, \quad (3.3)$$

is positive definite in the region where $\sum_{j=1}^m x_{ij}\theta_j > 0$ for each i , then (3.1) is a strictly concave function of θ there, and the maximum point is unique. For any m -dimensional vector λ , we have

$$\lambda' \mathbf{C} \lambda = \sum_{i=1}^n y_i \left[\sum_{j=1}^m x_{ij}\theta_j \right]^{-2} \left[\sum_{j=1}^m x_{ij}\lambda_j \right]^2 \geq 0. \quad (3.4)$$

If we let \mathbf{X}^r denote the "reduced" \mathbf{X} matrix—obtained by considering only \mathbf{x}_i such that $y_i \neq 0$ —then \mathbf{C} is positive definite provided $\mathbf{X}^r \lambda = \mathbf{0}$ has only the trivial solution $\lambda = \mathbf{0}$. Assuming this, a solution of (3.2) will be the maximum likelihood estimator of θ . Further discussion (and numerical examples) of the application of regression analysis to Poisson distributed data have been presented elsewhere [3, 4, 8].

[Received October 1974. Revised April 1975.]

REFERENCES

- [1] Ben-Israel, A. and Greville, T.N.E., *Generalized Inverses: Theory and Applications*, New York: John Wiley & Sons, Inc., 1974.
- [2] Bradley, E.L., "The Equivalence of Maximum Likelihood and Weighted Least Squares Estimates in the Exponential Family," *Journal of the American Statistical Association*, 68 (March 1973), 199-200.
- [3] Erlander, S., Gustavsson, J. and Svensson, A., "On Asymptotic Simultaneous Confidence Regions for Regression Planes in a Poisson Model," *Review of the International Statistical Institute*, 40, No. 2 (1972), 111-22.
- [4] Frome, E.L., Kutner, M.H. and Beauchamp, J.J., "Regression Analysis of Poisson-Distributed Data," *Journal of the American Statistical Association*, 68 (December 1973), 935-40.
- [5] Lawson, C.L. and Hanson, R.J., *Solving Least Squares Problems*, Englewood Cliffs, N.J.: Prentice-Hall Inc., 1974.
- [6] Mangasarian, O.L., *Nonlinear Programming*, New York: McGraw-Hill Book Co., 1969.
- [7] Moore, R.H. and Zeigler, R.K., "The Use of Nonlinear Regression Methods for Analyzing Sensitivity and Quantal Response Data," *Biometrics*, 23 (September 1967), 563-7.
- [8] Nelder, J.A. and Wedderburn, R.W.M., "Generalized Linear Models," *Journal of the Royal Statistical Society, Ser. A*, 135, Part 3 (1972), 370-84.
- [9] Turner, M.E., Monroe, R.S. and Lucas, H.L., "Generalized Asymptotic Regression and Nonlinear Data Analysis," *Biometrics*, 17 (March 1961), 120-43.