The Calculus of M-Estimation

Author(s): Leonard A. Stefanski and Dennis D. Boos

Source: *The American Statistician*, Feb., 2002, Vol. 56, No. 1 (Feb., 2002), pp. 29-38

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/3087324

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/3087324?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Taylor & Francis, Ltd. and *American Statistical Association* are collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician*

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms

The Calculus of M-Estimation

Leonard A. STEFANSKI and Dennis D. BOOS

Since the seminal papers by Huber in the 1960s, M-estimation methods (also known as estimating equation methods) have been increasingly important for asymptotic analysis and approximate inference. This article illustrates the breadth and generality of the M-estimation approach, thereby facilitating its use in practice and in the classroom as a unifying approach to the study of large-sample inference.

KEY WORDS: Asymptotic variance; Central limit theorem; Estimating equations; Large-sample inference; Maple; M-estimator.

1. INTRODUCTION

M-estimators are solutions of the vector equation $\sum_{i=1}^n \boldsymbol{\psi}(\mathbf{Y}_i, \boldsymbol{\theta}) = \mathbf{0}$; that is, the M-estimator $\widehat{\boldsymbol{\theta}}$ satisfies

$$\sum_{i=1}^n \boldsymbol{\psi}(\mathbf{Y}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{1}$$

Here we are assuming that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are independent but not necessarily identically distributed, $\boldsymbol{\theta}$ is a p-dimensional parameter, and $\boldsymbol{\psi}$ is a known $(p \times 1)$-function that does not depend on i or n. In this description $\mathbf{Y}_i$ represents the ith datum. In some applications it is advantageous to emphasize the dependence of $\boldsymbol{\psi}$ on particular components of $\mathbf{Y}_i$. For example, in a regression problem $\mathbf{Y}_i = (\mathbf{x}_i, Y_i)$ and (1) would typically be written

$$\sum_{i=1}^n \boldsymbol{\psi}(Y_i, \mathbf{x}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}, \tag{2}$$

where $\mathbf{x}_i$ is the ith regressor.

Huber (1964, 1967) introduced M-estimators and their asymptotic properties. They played an important part of the development of modern robust statistics. Liang and Zeger (1986) helped popularize M-estimators in the biostatistics literature under the name *generalized estimating equations* (GEE). Others have made important contributions. For example, Godambe (1960) introduced the concept of an *optimum estimating function* in an M-estimator context, and his article could be called a forerunner of the M-estimator approach.

However, our goal is not to document the development of M-estimators or to give a bibliography of contributions to the literature. Rather we want to show that the M-estimator approach is simple, powerful, and more widely applicable than many readers imagine. We have found the methods useful in our research, and we have been teaching them in graduate courses since the 1980s. The methods we describe are not new, and similar accounts of the fundamental techniques appear elsewhere; see, for example, Carroll, Ruppert, and Stefanski (1995, Appendix A.3). Our intent with this article is to provide a coherent exposition of the M-estimator approach, and to do so in a widely accessible forum and with a variety of interesting examples.

One key advantage of the approach is that a very large class of asymptotically normal statistics can be put in the general M-estimator framework. This unifies large-sample approximation methods, simplifies analysis, and makes computations routine if sometimes tedious. Fortunately, the tedious derivative and matrix calculations often can be performed symbolically with programs such as Maple and Mathematica.

Many estimators not typically thought of as M-estimators can be written in the form of M-estimators. For example, consider the mean deviation from the sample mean, $\widehat{\theta}_1 = n^{-1}\sum_{i=1}^n |Y_i - \overline{Y}|$. Is this an M-estimator? There is no equation $\boldsymbol{\psi}$ such that the estimating equation $\sum_{i=1}^n \boldsymbol{\psi}(Y_i, \theta) = 0$ yields $\widehat{\theta}_1$. But if we define $\psi_1(y, \theta_1, \theta_2) = |y - \theta_2| - \theta_1$ and $\psi_2(y, \theta_1, \theta_2) = y - \theta_2$, then

$$\sum_{i=1}^n \boldsymbol{\psi}(Y_i, \widehat{\theta}_1, \widehat{\theta}_2) = \begin{pmatrix} \sum_{i=1}^n \left(|Y_i - \widehat{\theta}_2| - \widehat{\theta}_1\right) \\ \sum_{i=1}^n \left(Y_i - \widehat{\theta}_2\right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

where $\widehat{\theta}_2 = \overline{Y}$ and $\widehat{\theta}_1 = n^{-1}\sum_{i=1}^n |Y_i - \overline{Y}|$. We use *partial M-estimator* to denote an estimator that is not separately an M-estimator, but is a component of an M-estimator for suitably defined additional $\boldsymbol{\psi}$ functions. Any estimator that would be an M-estimator if certain parameters were known, is a partial M-estimator because we can stack $\boldsymbol{\psi}$ functions for each of the unknown parameters. This aspect of M-estimators is related to the general approach of Randles (1982) for replacing unknown parameters by estimators.

From the above example it should be obvious that we can replace $\widehat{\theta}_2 = \overline{Y}$ by any other estimator defined by an estimating equation; for example, a robust location estimator. Moreover, we can also add $\boldsymbol{\psi}$ functions to handle delta-method asymptotics for transformations of parameters, for example, $\widehat{\theta}_3 = \log(\widehat{\theta}_1)$; see Example 2 in this article and also Benichou and Gail (1989).

The application of standard influence curve analysis and delta-method techniques (Serfling 1980, chaps. 3, 6) handles a larger class of problems than the enhanced M-estimation methods described herein. However, enhanced M-estimator methods, implemented with the aid of symbolic mathematics software (for deriving analytic expressions) and standard numerical routines for derivatives and matrix algebra (for obtaining numerical estimates) provide a unified approach that is simple in implementation, easily taught, and applicable to a broad class of complex problems.

Leonard A. Stefanski and Dennis D. Boos are Professors, Department of Statistics, North Carolina State University, Raleigh, NC 27695–8203 (E-mail addresses: stefansk@stat.ncsu.edu; boos@stat.ncsu.edu). The authors thank the reviewers for suggestions that led to substantial improvements in the manuscript.

© 2002 American Statistical Association

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms

The basic approach is described in Section 2 along with a few examples. Connections to the influence curve are given in Section 3 and then extensions for nonsmooth ψ functions are given in Section 4. Regression estimators are discussed in Section 5 and a testing problem is considered in Section 6. Section 7 concludes with a summary of the key features of the M-estimator method.

2. THE BASIC APPROACH

M-estimators satisfy (1), where the vector function $\boldsymbol{\psi}$ is a known function that does not depend on i or n. For the moment we confine ourselves to the iid case where $Y_1, \ldots, Y_n$ are iid (possibly vector-valued) with distribution function F. The true parameter value $\boldsymbol{\theta}_0$ is defined by

$$\mathrm{E}_F \boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0) = \int \boldsymbol{\psi}(y, \boldsymbol{\theta}_0) dF(y) = \mathbf{0}. \tag{3}$$

For example, if $\psi(Y_i, \theta) = Y_i - \theta$, then the population mean $\theta_0 = \int y dF(y)$ is the unique solution of $\int (y - \theta) dF(y) = 0$.

If (3) determines $\boldsymbol{\theta}_0$ uniquely, then in general there exists a sequence of M-estimators $\widehat{\boldsymbol{\theta}}$ such that $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ as $n \to \infty$ (Huber 1967; Serfling 1980, chap. 7). Furthermore, if $\boldsymbol{\psi}$ is suitably smooth, then Taylor expansion of $\mathbf{G}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \boldsymbol{\psi}(Y_i, \boldsymbol{\theta})$ gives

$$\mathbf{0} = \mathbf{G}_n(\widehat{\boldsymbol{\theta}}) = \mathbf{G}_n(\boldsymbol{\theta}_0) + \dot{\mathbf{G}}_n(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathbf{R}_n,$$

where $\dot{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \left[\partial \mathbf{G}_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$. For n sufficiently large, we expect $\dot{\mathbf{G}}_n(\boldsymbol{\theta}_0)$ to be nonsingular so that upon rearrangement,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left[-\dot{\mathbf{G}}_n(\boldsymbol{\theta}_0)\right]^{-1} \sqrt{n}\mathbf{G}_n(\boldsymbol{\theta}_0) + \sqrt{n}\mathbf{R}_n^*. \tag{4}$$

Define $\dot{\boldsymbol{\psi}}(y, \boldsymbol{\theta}) = \partial \boldsymbol{\psi}(y, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T$. Under suitable regularity conditions as $n \to \infty$,

$$-\dot{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^n \left[-\dot{\boldsymbol{\psi}}(Y_i, \boldsymbol{\theta}_0)\right] \xrightarrow{p} \mathrm{E}\left[-\dot{\boldsymbol{\psi}}(Y_1, \boldsymbol{\theta}_0)\right] = \mathbf{A}(\boldsymbol{\theta}_0); \tag{5}$$

$$\sqrt{n}\mathbf{G}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{MVN}(0, \mathbf{B}(\boldsymbol{\theta}_0)), \quad \text{where} \quad \mathbf{B}(\boldsymbol{\theta}_0) = \mathrm{E}\left[\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)^T\right]; \tag{6}$$

$$\sqrt{n}\mathbf{R}_n^* \xrightarrow{p} \mathbf{0}. \tag{7}$$

If $\mathbf{A}(\boldsymbol{\theta}_0)$ exists, the weak law of large numbers gives (5). If $\mathbf{B}(\boldsymbol{\theta}_0)$ exists, then (6) follows from the central limit theorem. Proving (7) is difficult. Huber (1967) was the first to give general results for (7), but there have been many others since then (see, e.g., Serfling 1980, chaps. 5-8). We shall be content to note that (7) holds in most situations provided $\boldsymbol{\psi}$ is sufficiently smooth, and $\boldsymbol{\theta}$ has fixed dimension as $n \to \infty$.

Combining (1) and (4)–(7) and an appeal to Slutsky's Theorem show that

$$\widehat{\boldsymbol{\theta}} \quad \text{is} \quad \mathrm{AMN}\left(\boldsymbol{\theta}_0, \frac{\mathbf{V}(\boldsymbol{\theta}_0)}{n}\right) \quad \text{as} \quad n \to \infty, \tag{8}$$

where $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T$ (AMN means "asymptotically multivariate normal"). The matrix product defining the limiting covariance matrix is called the *sandwich matrix* of $\mathbf{A}(\boldsymbol{\theta}_0)$ and $\mathbf{B}(\boldsymbol{\theta}_0)$, because the "meat" $\mathbf{B}(\boldsymbol{\theta}_0)$ is placed between the "bread" $\mathbf{A}(\boldsymbol{\theta}_0)^{-1}$ and $\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T$.

Extension. Suppose that instead of (1), $\widehat{\boldsymbol{\theta}}$ satisfies

$$\sum_{i=1}^n \boldsymbol{\psi}(Y_i, \widehat{\boldsymbol{\theta}}) = \mathbf{c}_n, \tag{9}$$

where $\mathbf{c}_n / \sqrt{n} \xrightarrow{p} \mathbf{0}$ as $n \to \infty$. The arguments above can be repeated with the sole change that $\mathbf{c}_n / \sqrt{n}$ is absorbed into the remainder $\sqrt{n}\mathbf{R}_n^*$ in (4), thus showing that if (9) and (4)–(7) hold, then (8) follows. This extension allows us to cover a much wider class of statistics including empirical quantiles, certain estimators whose $\boldsymbol{\psi}$ function depends on n, and Bayesian estimators.

For maximum likelihood estimation

$$\boldsymbol{\psi}(y, \boldsymbol{\theta}) = \partial \log f(y; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$$

is often called the score function. If the data follow the assumed parametric family $f(y; \boldsymbol{\theta})$, then $\mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{B}(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0)$, the information matrix. In this case the sandwich matrix $\mathbf{V}(\boldsymbol{\theta}_0)$ reduces to the usual $I(\boldsymbol{\theta}_0)^{-1}$. One of the key contributions of M-estimation theory has been to facilitate analysis when the assumed parametric family is not correct. In such cases there is often a well-defined $\boldsymbol{\theta}_0$ satisfying (3) and $\widehat{\boldsymbol{\theta}}$ satisfying (8), but $\mathbf{A}(\boldsymbol{\theta}_0) \neq \mathbf{B}(\boldsymbol{\theta}_0)$, and inference should be carried out using the correct limiting covariance matrix $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T$, not $I(\boldsymbol{\theta}_0)^{-1}$.

Equations (5) and (6) motivate the empirical estimators of $\mathbf{A}(\boldsymbol{\theta}_0)$ and $\mathbf{B}(\boldsymbol{\theta}_0)$,

$$\mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}}) = -\dot{\mathbf{G}}_n(\widehat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^n \left[-\dot{\boldsymbol{\psi}}(Y_i, \widehat{\boldsymbol{\theta}})\right],$$

and

$$\mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\psi}(Y_i, \widehat{\boldsymbol{\theta}})\boldsymbol{\psi}(Y_i, \widehat{\boldsymbol{\theta}})^T.$$

Note that for maximum likelihood estimation, $n\mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})$ is the observed information matrix $\mathbf{I}_{\mathbf{Y}}(\widehat{\boldsymbol{\theta}})$. The sandwich matrix of these matrix estimators yields the empirical sandwich variance estimator

$$\mathbf{V}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}}) = \mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})^{-1}\mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})\{\mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})^{-1}\}^T, \tag{10}$$

which is consistent for $\mathbf{V}(\boldsymbol{\theta}_0)$ under mild regularity conditions (Iverson and Randles 1989).

Calculation of $\mathbf{V}_n(\mathbf{Y}, \widehat{\theta})$ requires no analytic work beyond specifying $\boldsymbol{\psi}$. In some problems, it is simpler to work with the limiting form $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T$, plugging in estimators for $\boldsymbol{\theta}_0$ and any other unknown quantities in $\mathbf{V}(\boldsymbol{\theta}_0)$. The notation $\mathbf{V}(\boldsymbol{\theta}_0)$ suggests that $\boldsymbol{\theta}_0$ is the only unknown quantity in $\mathbf{V}(\boldsymbol{\theta}_0)$, but in reality $\mathbf{V}(\boldsymbol{\theta}_0)$ often involves moments or other characteristics of the distribution function of $\mathbf{Y}_i$. In fact

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms

there is a range of possibilities for estimating $\mathbf{V}(\boldsymbol{\theta}_0)$ depending on model assumptions used. For simplicity, we use the notation $\mathbf{V}_n(\mathbf{Y}, \widehat{\theta})$ for the purely empirical estimator, and $\mathbf{V}(\widehat{\boldsymbol{\theta}})$ for any estimators that exploit model assumptions.

For maximum likelihood estimation with a correctly specified family, the three competing estimators for $I(\boldsymbol{\theta})^{-1}$ are $\mathbf{V}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})$, $\left[I_{\mathbf{Y}}(\widehat{\boldsymbol{\theta}})/n\right]^{-1} = \mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})^{-1}$, and $I(\widehat{\boldsymbol{\theta}})^{-1} = \mathbf{V}(\widehat{\boldsymbol{\theta}})$. In this case the standard estimators $\left[I_{\mathbf{Y}}(\widehat{\boldsymbol{\theta}})/n\right]^{-1}$ and $I(\widehat{\boldsymbol{\theta}})^{-1}$ are generally more efficient than $\mathbf{V}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})$ for estimating $I(\boldsymbol{\theta})^{-1}$. Clearly no estimator of $I(\boldsymbol{\theta})^{-1}$ has smaller asymptotic variance than $I(\widehat{\boldsymbol{\theta}})^{-1}$.

Now we illustrate these ideas with examples.

Example 1. Sample Moments. Let $\widehat{\boldsymbol{\theta}} = (\overline{Y}, s_n^2)^T$, the sample mean and variance, be the M-estimator defined by

$$\boldsymbol{\psi}(Y_i, \boldsymbol{\theta}) = \begin{pmatrix} Y_i - \theta_1 \\ (Y_i - \theta_1)^2 - \theta_2 \end{pmatrix}.$$

The first component, $\widehat{\theta}_1 = \overline{Y}$, satisfies $\sum(Y_i - \widehat{\theta}_1) = 0$, and is by itself an M-estimator. The second component $\widehat{\theta}_2 = s_n^2 = n^{-1}\sum(Y_i - \overline{Y})^2$, by itself, is not an M-estimator. However, when combined with $\widehat{\theta}_1$, the pair $(\widehat{\theta}_1, \widehat{\theta}_2)$ is a 2×1 M-estimator so that $\widehat{\theta}_2$ satisfies the definition of a partial M-estimator.

Now let us calculate $\mathbf{A}(\boldsymbol{\theta}_0)$ and $\mathbf{B}(\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0^T = (\theta_{10}, \theta_{20})$:

$$\begin{aligned} \mathbf{A}(\boldsymbol{\theta}_0) &= \mathrm{E}\left[-\dot{\boldsymbol{\psi}}(Y_1, \boldsymbol{\theta}_0)\right] \\ &= \mathrm{E}\begin{pmatrix} 1 & 0 \\ 2(Y_1 - \theta_{10}) & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned} \tag{11}$$

$$\mathbf{B}(\boldsymbol{\theta}_0) = \mathrm{E}\left[\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)^T\right] \tag{12}$$

with elements

$$\begin{aligned} \mathbf{B}(\boldsymbol{\theta}_0)_{11} &= \mathrm{E}(Y_1 - \theta_{10})^2 = \theta_{20} = \sigma^2 \\ \mathbf{B}(\boldsymbol{\theta}_0)_{12} &= \mathrm{E}(Y_1 - \theta_{10})[(Y_1 - \theta_{10})^2 - \theta_{20}] = \mu_3 \\ \mathbf{B}(\boldsymbol{\theta}_0)_{22} &= \mathrm{E}[(Y_1 - \theta_{10})^2 - \theta_{20}]^2 = \mu_4 - \sigma^4, \end{aligned}$$

where μ_k denotes the kth central moment of Y_1 and the more familiar notation $\sigma^2 = \theta_{20}$ is substituted at the end. In this case $\mathbf{A}(\boldsymbol{\theta}_0)$ is the identity matrix, and $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{B}(\boldsymbol{\theta}_0)$. To estimate $\mathbf{B}(\boldsymbol{\theta}_0)$, we may use $\mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})$with elements

$$\begin{aligned} \mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})_{11} &= \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^2 = s_n^2 \\ \mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})_{12} &= \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})[(Y_i - \overline{Y})^2 - s_n^2] = m_3 \\ \mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})_{22} &= \frac{1}{n}\sum_{i=1}^n [(Y_i - \overline{Y})^2 - s_n^2]^2 = m_4 - s_n^4, \end{aligned}$$

where the m_k are sample kth moments. Considering the form of $\mathbf{V}(\boldsymbol{\theta}_0)$ and plugging in empirical moment estimators leads to equality of the empirical estimator $\mathbf{V}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})$ and the expected value estimator $\mathbf{V}(\widehat{\boldsymbol{\theta}})$ in this case.

Note that $\widehat{\boldsymbol{\theta}}$ is a maximum likelihood estimator for the normal model, $N(\theta_1, \theta_2)$; but $\psi_1 = Y_i - \theta_1$ and $\psi_2 = (Y_i - \theta_1)^2 - \theta_2$ are not the score functions from the corresponding normal density. The latter are $\psi_1 = (Y_i - \theta_1)/\theta_2$ and $\psi_2 = (Y_i - \theta_1)^2/(2\theta_2^2) - 1/(2\theta_2)$. Thus, $\boldsymbol{\psi}$ functions are not unique—many different ones lead to the same estimator. However, different $\boldsymbol{\psi}$ functions associated with the same estimator yield different $\mathbf{A}$ and $\mathbf{B}$ but the same $\mathbf{V}$. For example, using these latter two $\boldsymbol{\psi}$ functions, the resulting $\mathbf{A}$ and $\mathbf{B}$ matrices are

$$\mathbf{A}(\boldsymbol{\theta}_0) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}, \quad \mathbf{B}(\boldsymbol{\theta}_0) = \begin{pmatrix} \frac{1}{\sigma^2} & \frac{\mu_3}{2\sigma^3} \\ \frac{\mu_3}{2\sigma^3} & \frac{\mu_4 - \sigma^4}{4\sigma^8} \end{pmatrix}.$$

Note that the sandwich matrix of these matrices is the same as the sandwich matrix of the matrices in (11) and (12). If we further assume that the data truly are normally distributed, then $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$ resulting in $\mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{B}(\boldsymbol{\theta}_0) = \mathbf{I}(\boldsymbol{\theta}_0) = \mathrm{diag}(1/\sigma^2, 1/2\sigma^4)$. Here the expected-value, model-based covariance estimator is $\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \mathrm{diag}(1/s_n^2, 1/2s_n^4)$.

Note that the likelihood score $\boldsymbol{\psi}$ functions, $\boldsymbol{\psi}_{\text{MLE}}$, are related to the original $\boldsymbol{\psi}$ functions by $\boldsymbol{\psi}_{\text{MLE}} = \mathbf{C}\boldsymbol{\psi}$, where $\mathbf{C} = \mathrm{diag}(1/\theta_2, 1/(2\theta_2^2))$. A little algebra shows that all $\boldsymbol{\psi}$ of the form $\mathbf{C}\boldsymbol{\psi}$, where $\mathbf{C}$ (may depend on $\boldsymbol{\theta}$) is nonsingular, lead to an equivalence class associated with the same estimator and asymptotic variance matrix $\mathbf{V}(\boldsymbol{\theta}_0)$.

Example 2. Ratio Estimator. Let $\widehat{\theta} = \overline{Y}/\overline{X}$, where $(Y_1, X_1), \ldots, (Y_n, X_n)$ is an iid sample of pairs with means $\mathrm{E}Y_1 = \mu_Y$ and $\mathrm{E}X_1 = \mu_X \neq 0$, variances $\mathrm{var}(Y_1) = \sigma_Y^2$ and $\mathrm{var}(X_1) = \sigma_X^2$, and covariance $\mathrm{cov}(Y_1, X_1) = \sigma_{YX}$. A $\boldsymbol{\psi}$ function for $\widehat{\theta} = \overline{Y}/\overline{X}$ is $\psi(Y_i, X_i, \theta) = Y_i - \theta X_i$ leading to $\mathbf{A}(\boldsymbol{\theta}_0) = \mu_X$, $\mathbf{B}(\boldsymbol{\theta}_0) = \mathrm{E}(Y_1 - \theta_0 X_1)^2$, $\mathbf{V}(\boldsymbol{\theta}_0) = \mathrm{E}(Y_1 - \theta_0 X_1)^2/\mu_X^2$, $\mathbf{A}_n(\mathbf{Y}, \widehat{\theta}) = \overline{X}$, and

$$\mathbf{B}_n(\mathbf{Y}, \widehat{\theta}) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \frac{\overline{Y}}{\overline{X}}X_i\right)^2,$$

and

$$\mathbf{V}_n(\mathbf{Y}, \widehat{\theta}) = \frac{1}{\overline{X}^2}\frac{1}{n}\sum_{i=1}^n \left(Y_i - \frac{\overline{Y}}{\overline{X}}X_i\right)^2.$$

This variance estimator is often encountered in finite population sampling contexts.

Now consider

$$\boldsymbol{\psi}(Y_i, X_i, \boldsymbol{\theta}) = (Y_i - \theta_1,\ X_i - \theta_2,\ \theta_1 - \theta_3\theta_2)^T$$

that yields $\widehat{\theta}_3 = \overline{Y}/\overline{X}$ as the third component of $\widehat{\boldsymbol{\theta}}$. This $\boldsymbol{\psi}$ function is interesting because the third component does not depend on the data. Nevertheless, this $\boldsymbol{\psi}$ satisfies all the requirements and illustrates how to implement the delta method via

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms

M-estimation. The $\mathbf{A}$ and $\mathbf{B}$ matrices are

$$\mathbf{A}(\boldsymbol{\theta}_0) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & \theta_{30} & \theta_{20} \end{pmatrix} \quad \mathbf{B}(\boldsymbol{\theta}_0) = \begin{pmatrix} \sigma_Y^2 & \sigma_{YX} & 0 \\ \sigma_{YX} & \sigma_X^2 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

This example illustrates the fact that $\mathbf{B}(\boldsymbol{\theta}_0)$ can be singular. Whenever a ψ function has components that do not depend on the data, the resulting B matrix is singular. However, $\mathbf{A}(\boldsymbol{\theta}_0)$ is generally nonsingular, a fact that is related to our assumptions that (1) and (3) have unique solutions.

Although this problem is amenable to hand calculation, it provides a good illustration of the advantages of using a symbolic mathematics program for doing some of the routine calculations associated with M-estimation. Using Maple we computed $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T$. The Maple code is shown in Figure 1.

The code shown in Figure 1 produces

$$\frac{{\sigma_y}^2 - 2\,\theta_{30}\,\sigma_{xy} + {\theta_{30}}^2\,{\sigma_x}^2}{{\theta_{20}}^2}.$$

This expression for the asymptotic variance of $\sqrt{n}\widehat{\theta}_3$ is identical to $\mathrm{E}(Y_1 - \theta_{30}X_1)^2/\mu_X^2$ after making the substitution $\theta_{20} = \mu_X$.

Example 3. Further illustration of the delta method via M-estimation. In the context of Example 1, suppose we are interested in $s_n = \sqrt{s_n^2}$ and $\log(s_n^2)$. We could just redefine θ_2 in Example 1 to be θ_2^2 and $\exp(\theta_2)$, respectively. Instead, we add $\psi_3(Y_i, \boldsymbol{\theta}) = \sqrt{\theta_2} - \theta_3$ and $\psi_4(Y_i, \boldsymbol{\theta}) = \log(\theta_2) - \theta_4$ because it is conceptually simpler and it also gives the joint asymptotic distribution of all estimators under study. Now we have

$$\mathbf{A}(\boldsymbol{\theta}_0) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\dfrac{1}{2\sqrt{\theta_{20}}} & 1 & 0 \\ 0 & -\dfrac{1}{\theta_{20}} & 0 & 1 \end{pmatrix}$$

$$\mathbf{B}(\boldsymbol{\theta}_0) = \begin{pmatrix} \dfrac{1}{\theta_{20}} & \dfrac{\mu_3}{2\theta_{20}^3} & 0 & 0 \\ \dfrac{\mu_3}{2\theta_{20}^3} & \dfrac{\mu_4 - \theta_{20}^2}{4\theta_{20}^4} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T$ is

$$\mathbf{V}(\boldsymbol{\theta}_0) = \begin{pmatrix} \theta_{20} & \mu_3 & \dfrac{\mu_3}{2\sqrt{\theta_{20}}} & \dfrac{\mu_3}{\theta_{20}} \\ \mu_3 & \mu_4 - \theta_{20}^2 & \dfrac{\mu_4 - \theta_{20}^2}{2\sqrt{\theta_{20}}} & \dfrac{\mu_4 - \theta_{20}^2}{\theta_{20}} \\ \dfrac{\mu_3}{2\sqrt{\theta_{20}}} & \dfrac{\mu_4 - \theta_{20}^2}{2\sqrt{\theta_{20}}} & \dfrac{\mu_4 - \theta_{20}^2}{4\theta_{20}} & \dfrac{\mu_4 - \theta_{20}^2}{2\theta_{20}^{3/2}} \\ \dfrac{\mu_3}{\theta_{20}} & \dfrac{\mu_4 - \theta_{20}^2}{\theta_{20}} & \dfrac{\mu_4 - \theta_{20}^2}{2\theta_{20}^{3/2}} & \dfrac{\mu_4 - \theta_{20}^2}{\theta_{20}^2} \end{pmatrix}.$$

Thus, the asymptotic variance of s_n is $(\mu_4 - \theta_{20}^2)/(4\theta_{20}) = (\mu_4 - \sigma^4)/4\sigma^2$, and the asymptotic variance of $\log(s_n^2)$ is $(\mu_4 - \theta_{20}^2)/\theta_{20}^2 = \mu_4/\sigma^4 - 1$.

Example 4. Instrumental Variable Estimation. Instrumental variable estimation is a method for estimating regression parameters when predictor variables are measured with error (Fuller 1967; Carroll et al. 1995). We use a simple instrumental variable model to illustrate features of the M-estimation approach. Suppose that triples (Y_i, W_i, T_i) are observed such that

$$Y_i = \alpha + \beta X_i + \sigma_\varepsilon \varepsilon_{1,i},$$

$$W_i = X_i + \sigma_U \varepsilon_{2,i},$$

and

$$T_i = \gamma + \delta X_i + \sigma_\tau \varepsilon_{3,i},$$

where $\varepsilon_{j,i}$ are mutually independent random errors with common mean 0 and variance 1. For simplicity also assume that $X_1, \ldots, X_n$ are iid, independent of the $\{\varepsilon_{j,i}\}$ and have finite variance. In the language of measurement error models, W_i is a measurement of X_i, and T_i is an instrumental variable for X_i (for estimating β), provided that $\delta \neq 0$ which we now assume. Note that $X_1, \ldots, X_n$ are latent variables and not observed. Let σ_S^2 and $\sigma_{S,T}$ denote variances and covariances of any random variables S and T.

The least squares estimator of slope obtained by regressing Y on W, $\widehat{\beta}_{Y|W}$, converges in probability to $\{\sigma_X^2/(\sigma_X^2 + \sigma_U^2)\}\,\beta$, and thus is not consistent for β when the measurement error variance $\sigma_U^2 > 0$. However, the instrumental variable estimator $\widehat{\beta}_{\mathrm{IV}} = \widehat{\beta}_{Y|T}/\widehat{\beta}_{W|T}$, where $\widehat{\beta}_{Y|T}$ and $\widehat{\beta}_{W|T}$ are the slopes from the least squares regressions of Y on T and W on T, respectively, is a consistent estimator of β under the stated assumptions regardless of σ_U^2.

The instrumental variable estimator $\widehat{\beta}_{\mathrm{IV}}$ is a partial M-estimator as defined in Section 1, and there are a number of ways to complete the ψ function in this case. Provided interest lies only in estimation of the β, a simple choice is $\boldsymbol{\psi}(Y,\ W,\ T,\ \theta) = (\theta_1 - T,\ (Y - \theta_2 W)(\theta_1 - T))^T$, with associated M-estimator, $\widehat{\theta}_1 = \overline{T}$, $\widehat{\theta}_2 = \widehat{\beta}_{\mathrm{IV}}$. The $\mathbf{A}$ and $\mathbf{B}$ matrices calculated at the true parameters assuming the instrumental variable model are

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ \alpha & \sigma_{X,T} \end{pmatrix}$$

and

$$\mathbf{B} = \begin{pmatrix} \sigma_T^2 & \alpha\sigma_T^2 \\ \alpha\sigma_T^2 & \sigma_T^2(\alpha^2 + \sigma_\epsilon^2 + \beta^2\sigma_U^2) \end{pmatrix},$$

which yield the asymptotic variance matrix

$$\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{A}^{-1}\right)^T = \begin{pmatrix} \sigma_T^2 & 0 \\ 0 & \sigma_T^2(\sigma_\epsilon^2 + \beta^2\sigma_U^2)/\sigma_{X,T}^2 \end{pmatrix}.$$

Under the stated assumptions the instrumental variable estimator and the naive estimator are both consistent for β when $\sigma_U^2 = 0$, yet have different asymptotic means when $\sigma_U^2 > 0$. Thus for certain applications their joint asymptotic distribution is of interest, for example, for inference about the difference

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms

```
with(linalg):                                   Bring in linear algebra package
vA:=[1,0,0,0,1,0,-1,theta[30],theta[20]];       Make a vector of the entries of A
A:=matrix(3,3,vA);                              Create A from vA
Ainv:=inverse(A);
vB:=[sigma[y]^2,sigma[xy],0,sigma[xy],sigma[x]^2,0,0,0,0];
B:=matrix(3,3,vB);
V:=multiply(Ainv,B,transpose(Ainv));
simplify(V[3,3]);
```

Figure 1. Maple code used to compute $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T$.

$\widehat{\beta}_{\mathrm{IV}} - \widehat{\beta}_{Y|W}$. The M-estimator approach easily accommodates such calculations. For this task consider the ψ function

$$\psi(Y,\, W,\, T,\, \theta) = (\theta_1 - T,\, \theta_2 - W,\, (Y - \theta_3 W)(\theta_2 - W),\, (Y - \theta_4 W)(\theta_1 - T))^T.$$

Note the change in the definitions of θ_2 and the ordering of the components of this ψ function. The configuration is primarily for convenience as it leads to a triangular $\mathbf{A}$ matrix. In general when the kth component of ψ depends only on $\theta_1, \ldots, \theta_k$, $k = 1, 2, \ldots$, the partial derivative matrix $\partial\psi/\partial\boldsymbol{\theta}^T$ is lower triangular and so too is the $\mathbf{A}$ matrix.

The M-estimator associated with this ψ function is $\widehat{\theta}_1 = \overline{T}$, $\widehat{\theta}_2 = \overline{W}$, $\widehat{\theta}_3 = \widehat{\beta}_{Y|W}$, $\widehat{\theta}_4 = \widehat{\beta}_{\mathrm{IV}}$. The $\mathbf{A}$ matrix calculated at the true parameters assuming the instrumental variable model is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \alpha + \beta\mu_X\sigma_U^2/\sigma_W^2 & \sigma_W^2 & 0 \\ \alpha & 0 & 0 & \sigma_{XT} \end{pmatrix}.$$

The expression for the $\mathbf{B}$ matrix is unwieldy. However, primary interest lies in the lower 2×2 submatrix of the asymptotic variance matrix $\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{A}^{-1}\right)^T$. We used Maple to calculate this submatrix and to substitute expressions for the various mixed moments of (Y, W, T) under the assumption of joint normality, resulting in the asymptotic covariance matrix for $(\widehat{\theta}_3,\, \widehat{\theta}_4)$,

$$\begin{pmatrix} (\sigma_\varepsilon^2\sigma_W^2 + \beta^2\sigma_U^2\sigma_X^2)/\sigma_W^4 & \left\{\sigma_\varepsilon^2\sigma_W^2 + \beta^2(\sigma_U^2\sigma_X^2 - \sigma_U^4)\right\}/\sigma_W^4 \\ \left\{\sigma_\varepsilon^2\sigma_W^2 + \beta^2(\sigma_U^2\sigma_X^2 - \sigma_U^4)\right\}/\sigma_W^4 & \sigma_T^2(\sigma_\varepsilon^2 + \beta^2\sigma_U^2)/\sigma_{X,T}^2 \end{pmatrix}.$$

The variance formula above assumes normality of the errors $\varepsilon_{j,i}$ and the X_i in the model. Instrumental variable estimation works more generally. In the absence of distributional assumptions (beyond lack of correlation) estimated variances can be obtained using empirical sandwich formulas.

3. CONNECTIONS TO THE INFLUENCE CURVE

The influence curve (Hampel 1974) $\mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(y; \boldsymbol{\theta}_0)$ of an estimator $\widehat{\boldsymbol{\theta}}$ based on an iid sample may be defined as satisfying

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(Y_i, \boldsymbol{\theta}_0) + \mathbf{R}_n,$$

where $\sqrt{n}\mathbf{R}_n \xrightarrow{p} 0$ as $n \to \infty$. If $\mathrm{E}[\mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(Y_1, \boldsymbol{\theta}_0)] = \mathbf{0}$ and $\mathrm{E}[\mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(Y_1, \boldsymbol{\theta}_0)\mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(Y_1, \boldsymbol{\theta}_0)^T] = \boldsymbol{\Sigma}$ exists, then by Slutsky's Theorem and the CLT, $\widehat{\boldsymbol{\theta}}$ is $\mathrm{AMN}(\mathbf{0}, \boldsymbol{\Sigma}/n)$. It is easy to verify that $\mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(y; \boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}(y; \boldsymbol{\theta}_0)$ for M-estimators. Thus

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathrm{E}\left[\mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(Y_1, \boldsymbol{\theta}_0)\mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(Y_1, \boldsymbol{\theta}_0)^T\right] \\ &= \mathrm{E}\left[\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)\{\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)\}^T\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T\right] \\ &= \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T = \mathbf{V}(\boldsymbol{\theta}_0). \end{aligned}$$

The influence curve approach is more general than the M-estimator approach; however, for many problems they are equivalent. Our experience teaching both methods indicates that students more readily learn the M-estimator approach and are therefore more likely to use it (and use it correctly) in their research and work. Especially in messy problems with a large number of parameters, it appears easier to stack ψ functions and compute $\mathbf{A}$ and $\mathbf{B}$ matrices than it is to compute and stack influence curves and then compute $\boldsymbol{\Sigma}$.

If the influence curve is known, then defining $\boldsymbol{\psi}(Y_i, \boldsymbol{\theta}) = \mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(Y_i, \boldsymbol{\theta}_0) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ allows one to use the approach of this article. In this case $\mathbf{A}(\boldsymbol{\theta}_0)$ is the identity matrix and $\mathbf{B}(\boldsymbol{\theta}_0) = \boldsymbol{\Sigma}$. (A minor modification is that for the empirical variance estimators we need to define $\boldsymbol{\psi}(Y_i, \widehat{\boldsymbol{\theta}}) = \mathbf{IC}_{\widehat{\boldsymbol{\theta}}}(Y_i, \widehat{\boldsymbol{\theta}})$; i.e., plugging in $\widehat{\boldsymbol{\theta}}$ for both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$.) More importantly, this fact allows one to combine M-estimators with estimators that may not be M-estimators but for which we have already computed influence curves. The next example illustrates this.

Example 5. Hodges and Lehmann (1963) suggested that estimators could be obtained by inverting rank tests, and the class of such estimators is called R-estimators. One of the most interesting R-estimators is called the Hodges–Lehmann location estimator,

$$\widehat{\theta}_{\mathrm{HL}} = \mathrm{median}\left\{\frac{X_i + X_j}{2}, 1 \le i \le j \le n\right\}.$$

It is not clear how to put this estimator directly in the M-estimator framework, but for distributions symmetric around θ_0, that is, having $F(y) = F_0(y - \theta_0)$ for a distribution F_0 symmetric about 0, Huber (1981, p. 64) gave $\mathrm{IC}_{\widehat{\theta}_{\mathrm{HL}}}(y; \theta_0) = \left\{\int f_0^2(y)dy\right\}^{-1}\{F_0(y - \theta_0) - 0.5\}$, where $f_0(y)$ is the density function of $F_0(y)$. The variance of this influence curve is $\left[12\left\{\int f_0^2(y)dy\right\}^2\right]^{-1}$, which is easily obtained after noting that $F_0(Y_1 - \theta_0)$ has a uniform distribution.

Now for obtaining the asymptotic joint distribution of $\widehat{\theta}_{\mathrm{HL}}$ and any set of M-estimators, we can stack $\psi(Y_i, \theta) = \mathrm{IC}_{\widehat{\theta}_{\mathrm{HL}}}(y; \theta_0) - (\theta - \theta_0)$ with the $\boldsymbol{\psi}$ functions of the M-estimators. The part of the $\mathbf{A}$ matrix associated with $\widehat{\theta}_{\mathrm{HL}}$ will be all zeroes except for

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms

the diagonal element which will be a one. The diagonal element of the $\mathbf{B}$ matrix will be the asymptotic variance given above, but one will still need to compute correlations of $\mathrm{IC}_{\widehat{\theta}_{\mathrm{HL}}}(Y_1, \theta_0)$ with the other $\boldsymbol{\psi}$ functions to get the off-diagonal elements of the $\mathbf{B}$ matrix involving $\widehat{\theta}_{\mathrm{HL}}$.

4. NONSMOOTH $\boldsymbol{\psi}$ FUNCTIONS

In some situations the $\boldsymbol{\psi}$ function may not be differentiable everywhere, thus invalidating the definition of the $\mathbf{A}$ matrix as the expected value of a derivative. The appropriately modified definition of $\mathbf{A}$ interchanges the order of differentiation and expectation,

$$\mathbf{A}(\boldsymbol{\theta}_0) \equiv -\frac{\partial}{\partial \boldsymbol{\theta}^T} \left\{ \mathrm{E}_F \boldsymbol{\psi}(Y_1, \boldsymbol{\theta}) \right\} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{13}$$

The expectation is with respect to the true distribution of the data (denoted E_F), but $\boldsymbol{\theta}$ within $\boldsymbol{\psi}$ varies freely with respect to differentiation, after which the true parameter value $\boldsymbol{\theta}_0$ replaces $\boldsymbol{\theta}$.

The next two examples illustrate (13) in two qualitatively different problems. In Example 6 the interchange of differentiation and expectation is justified (even though the $\boldsymbol{\psi}$ function is not everywhere differentiable) and (13) and (5) are equivalent. However, this is not the case in Example 7.

Example 6. Huber (1964) proposed estimating the center of symmetry of symmetric distributions using $\widehat{\theta}$ that satisfies $\sum \psi_k(Y_i - \widehat{\theta}) = 0$, where

$$\psi_k(x) = \begin{cases} x & \text{when} \quad |x| \le k, \\ \operatorname{sgn}(x)k & \text{when} \quad |x| > k. \end{cases}$$

This $\boldsymbol{\psi}$ function is continuous everywhere but not differentiable at $\pm k$. By definition (13), and assuming that F has density f,

$$\begin{aligned} A(\theta_0) &= -\frac{\partial}{\partial \theta} \left\{ \mathrm{E}_F \psi_k(Y_1 - \theta) \right\} \Big|_{\theta=\theta_0} \\ &= -\frac{\partial}{\partial \theta} \left\{ \int \psi_k(y - \theta) dF(y) \right\} \Big|_{\theta=\theta_0} \\ &= \int \left\{ -\frac{\partial}{\partial \theta} \psi_k(y - \theta) \right\} \Big|_{\theta=\theta_0} dF(y) \\ &= \int \psi_k'(y - \theta_0) dF(y) \end{aligned}$$

The notation ψ_k' inside the integral stands for the derivative of ψ_k where it exists. Verifying the second equality above is an instructive calculus exercise.

For $B(\theta_0)$ we have $B(\theta_0) = \mathrm{E}\psi_k^2(Y_1 - \theta_0) = \int \psi_k^2(y - \theta_0) dF(y)$, and thus

$$\mathbf{V}(\theta_0) = \frac{\int \psi_k^2(y - \theta_0) dF(y)}{\left[\int \psi_k'(y - \theta_0) dF(y) \right]^2}.$$

For estimating $A(\theta_0)$ and $B(\theta_0)$, our usual estimators are $A_n(\mathbf{Y}, \widehat{\theta}) = n^{-1} \sum_{i=1}^n \left[-\psi_k'(Y_i - \widehat{\theta}) \right]$ and $B_n(\mathbf{Y}, \widehat{\theta}) = n^{-1} \sum_{i=1}^n \psi_k^2(Y_i - \widehat{\theta})$ (or perhaps $(n-1)^{-1} \sum_{i=1}^n \psi_k^2(Y_i - \widehat{\theta})$). Here we can use $\psi_k'(Y_i - \widehat{\theta})$ because we expect to have data at $Y_i - \widehat{\theta} = \pm k$ with probability 0.

Example 7. Sample Quantiles. The sample pth quantile $\widehat{\theta} = F_n^{-1}(p)$ satisfies $\sum \left[p - I(Y_i \le \widehat{\theta}) \right] = c_n$, where $|c_n| = n|F_n(\widehat{\theta}) - p| \le 1$ and F_n is the empirical distribution function. Thus using the extended definition (9), the $\boldsymbol{\psi}$ function is $\psi(Y_i, \theta) = p - I(Y_i \le \theta)$. This $\boldsymbol{\psi}$ function is discontinuous at $\theta = Y_i$ and its derivative with respect to θ vanishes almost everywhere. However, definition (13) of $A(\theta_0)$ continues to give us the correct asymptotic results

$$\begin{aligned} A(\theta_0) &= -\frac{\partial}{\partial \theta} \left\{ \mathrm{E}_F \left[p - I(Y_1 \le \theta) \right] \right\} \Big|_{\theta=\theta_0} \\ &= -\frac{\partial}{\partial \theta} \left[p - F(\theta) \right] \Big|_{\theta=\theta_0} = f(\theta_0), \end{aligned}$$

$$B(\theta_0) = \mathrm{E}\left[p - I(Y_1 \le \theta_0) \right]^2 = p(1-p),$$

$$\mathbf{V}(\theta_0) = \frac{p(1-p)}{f^2(\theta_0)}.$$

Also, we could easily stack any finite number of quantile $\boldsymbol{\psi}$ functions together to get the joint asymptotic distribution of $(F_n^{-1}(p_1), \ldots, F_n^{-1}(p_k))$. There is a cost, however, for the jump discontinuities in these $\boldsymbol{\psi}$ functions: we no longer can use $A_n(\mathbf{Y}, \widehat{\theta})$ to estimate $\mathbf{A}(\theta_0)$. In fact, the derivative of the pth quantile $\boldsymbol{\psi}$ function is zero everywhere except at the location of the jump discontinuity. There are several options for estimating $\mathbf{A}(\theta_0)$. One is to use a smoothing technique to estimate f (e.g., kernel density estimators). Another is to approximate $\boldsymbol{\psi}$ by a smooth $\boldsymbol{\psi}$ function and use the $\mathbf{A}_n(\mathbf{Y}, \hat{\theta})$ from this smooth approximation.

5. REGRESSION M-ESTIMATORS

Regression M-estimators are a natural extension of location M-estimators. Although a number of different regression M-estimators have been proposed and studied, the fundamental ideas were established by Huber (1973; 1981, chap. 7).

There are two situations of interest for M-estimator analysis of regression estimators. In one the independent variables are random variables and the $(\mathbf{X}, Y)$ are modeled as iid pairs. This situation fits into the basic theory of Section 2 for iid sampling, for example, see Example 4. In the second situation the independent variables are fixed nonrandom constants. This covers standard regression models as well as ANOVA problems. For the second regression situation some new notation is required that reflects the non-iid character of the data.

A simple yet general setting to introduce notation is the nonlinear model

$$Y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + e_i \qquad i = 1, \ldots, n, \tag{14}$$

where g is a known differentiable function in $\boldsymbol{\beta}$, $e_1, \ldots, e_n$ are independent with mean 0 and possibly unequal variances $\operatorname{var}(e_i) = \sigma_i^2$, $i = 1, \ldots, n$, and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are known constant vectors. As usual we put the vectors together and define

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms

$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. The least squares estimator satisfies

$$\sum_{i=1}^{n} (Y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}))\dot{g}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}) = \mathbf{0},$$

where $\dot{g}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})$ denotes the partial derivative of g with respect to $\boldsymbol{\beta}$ evaluated at $\widehat{\boldsymbol{\beta}}$. Expanding this equation about the true value and rearranging, we get

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left[\frac{1}{n}\sum_{i=1}^{n} -\dot{\boldsymbol{\psi}}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\right]^{-1} \times \frac{1}{\sqrt{n}}\sum_{i=1} \boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) + \sqrt{n}R_n^*, \quad (15)$$

where $\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) = (Y_i - g(\mathbf{x}_i, \boldsymbol{\beta}_0))\dot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)$. We now give general definitions for a number of quantities followed by the result for the least squares estimator.

$$\begin{aligned} A_n(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}_0) &= \frac{1}{n}\sum_{i=1}^{n}\left[-\dot{\boldsymbol{\psi}}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\right] \quad (16)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\dot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)\dot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)^T - (Y_i - g(\mathbf{x}_i, \boldsymbol{\beta}_0))\ddot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)\right]. \quad (17)\end{aligned}$$

The notation principle is the same as before: all arguments of a quantity will be included in its name if those quantities are required for calculation. Now taking expectations with respect to the true model, define

$$\begin{aligned} A_n(\mathbf{X}, \boldsymbol{\beta}_0) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[-\dot{\boldsymbol{\psi}}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\dot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)\dot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)^T. \quad (18)\end{aligned}$$

Notice that we have dropped $\mathbf{Y}$ from this quantity's name because the expectation eliminates dependence on the Y_i. Also note that the second term for the least squares estimator drops out because of the model assumption (14). Finally, assuming that the limit exist, define

$$\begin{aligned} \mathbf{A}(\boldsymbol{\beta}_0) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[-\dot{\boldsymbol{\psi}}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\right]\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\dot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)\dot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)^T. \quad (19)\end{aligned}$$

In the linear regression case, note that $\mathbf{A}(\boldsymbol{\beta}_0) = \lim_{n\to\infty}\mathbf{X}^T\mathbf{X}/n$. This limit need not exist for the least squares estimator to be consistent and asymptotically normal, but its existence is a typical assumption leading to those desired results. Definition (17) leads to the purely empirical estimator of $A(\boldsymbol{\beta}_0)$:

$$\begin{aligned} &A_n(\mathbf{X}, \mathbf{Y}, \widehat{\boldsymbol{\beta}})\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[-\dot{\boldsymbol{\psi}}(Y_i, \mathbf{x}_i, \widehat{\boldsymbol{\beta}})\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\dot{g}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})\dot{g}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})^T - (Y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}))\ddot{g}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})\right]. \quad (20)\end{aligned}$$

This is the negative of the Hessian in a final Newton iteration, and thus is sometimes preferred on computational grounds. But the estimated expected value estimator based on (18) is typically simpler:

$$\begin{aligned} A_n(\mathbf{X}, \widehat{\boldsymbol{\beta}}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathrm{E}\left[-\dot{\boldsymbol{\psi}}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\right]\right\}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\\ &= \frac{1}{n}\sum_{i=1}^{n}\dot{g}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})\dot{g}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})^T. \quad (21)\end{aligned}$$

For the "B" matrices, we have in this expanded notation

$$\begin{aligned} B_n(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}_0) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)^T\\ &= \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\dot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)\dot{g}(\mathbf{x}_i, \boldsymbol{\beta}_0)^T, \quad (22)\end{aligned}$$

and $\mathbf{B}(\boldsymbol{\beta}_0)$ is just the limit of $B_n(\mathbf{X}, \boldsymbol{\beta}_0)$ as $n \to \infty$. A natural estimator of $\mathbf{B}(\boldsymbol{\beta}_0)$ is

$$\begin{aligned} &B_n(\mathbf{X}, \mathbf{Y}, \widehat{\boldsymbol{\beta}})\\ &= \frac{1}{n-p}\sum_{i=1}^{n}\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \widehat{\boldsymbol{\beta}})\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \widehat{\boldsymbol{\beta}})^T \quad (23)\\ &= \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}))^2\dot{g}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})\dot{g}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})^T. \quad (24)\end{aligned}$$

Example 8. Robust regression. Huber (1973) discussed robust regression alternatives to least squares in the linear regression context. As a specific example, consider the regression model in (14) with $g(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}^T\boldsymbol{\beta}$ and estimator of $\boldsymbol{\beta}$ satisfying

$$\sum_{i=1}^{n}\psi_k(Y_i - \mathbf{x}^T\widehat{\boldsymbol{\beta}})\mathbf{x}_i = 0, \quad (25)$$

where ψ_k is the function defined in Example 6. In this case $\boldsymbol{\psi}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}) = \psi_k(Y_i - \mathbf{x}^T\boldsymbol{\beta})\mathbf{x}_i$ (note the subtle but important distinction between $\boldsymbol{\psi}$ and ψ_k). Because ψ_k is an odd function about zero, the defining equations $\mathrm{E}\psi_k(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0)\mathbf{x}_i = 0$ will be satisfied if the e_i have a symmetric distribution about zero. If the e_i are not symmetrically distributed and the $\mathbf{X}$ matrix contains a column of ones, then the intercept estimated by $\widehat{\boldsymbol{\beta}}$ will be different from that estimated by least squares, but this is the only component of $\boldsymbol{\beta}_0$ affected by asymmetry.

Differentiating results in

$$\begin{aligned} \mathbf{A}_n(\mathbf{X}, \mathbf{Y}, \boldsymbol{\beta}_0) &= \frac{1}{n}\sum_{i=1}^{n}\left[-\dot{\boldsymbol{\psi}}(Y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\dot{\psi}_k(e_i)\mathbf{x}_i\mathbf{x}_i^T,\end{aligned}$$

and $\mathbf{A}_n(\mathbf{X}, \boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^{n}\mathrm{E}\dot{\psi}_k(e_i)\mathbf{x}_i\mathbf{x}_i^T$. Also, $\mathbf{B}_n(\mathbf{X}, \boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^{n}\mathrm{E}\psi(e_i)^2\mathbf{x}_i\mathbf{x}_i^T$. If the errors $e_1, \ldots, e_n$ are

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms

Table 1. Shaquille O'Neal Free Throws in 2000 NBA Playoffs

Game number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
FT's made	4	5	5	5	2	7	6	9	4	1	13	5	6	9	7	3	8	1	18	3	10	1	3
FT's attempted	5	11	14	12	7	10	14	15	12	4	27	17	12	9	12	10	12	6	39	13	17	6	12
Prop. made	.80	.45	.36	.42	.29	.70	.43	.60	.33	.25	.48	.29	.50	1.0	.58	.30	.67	.17	.46	.23	.59	.17	.25

identically distributed, then $\mathbf{A}_n(\mathbf{X}, \boldsymbol{\beta}_0) = \mathrm{E}\dot{\psi}_k(e_1)\mathbf{X}^T\mathbf{X}/n$, $\mathbf{B}_n(\mathbf{X}, \boldsymbol{\beta}_0) = \mathrm{E}\psi_k(e_1)^2\mathbf{X}^T\mathbf{X}/n$, and $\mathbf{V}(\mathbf{X}, \boldsymbol{\beta}_0) = (\mathbf{X}^T\mathbf{X}/n)^{-1}\mathrm{E}\psi_k(e_1)^2/[\mathrm{E}\dot{\psi}_k(e_1)]^2$.

Example 9. Generalized linear models have score equations

$$\sum_{i=1}^{n} \mathbf{D}_i(\boldsymbol{\beta}) \frac{(Y_i - \mu_i(\boldsymbol{\beta}))}{V_i(\boldsymbol{\beta})\tau} = 0, \tag{26}$$

where $\mu_i(\boldsymbol{\beta}_0) = \mathrm{E}(Y_i) = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta}_0)$, $\mathbf{D}_i(\boldsymbol{\beta}) = \partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, $V_i(\boldsymbol{\beta}_0)\tau_0 = \mathrm{var}(Y_i)$, g is the link function, and τ is a variance parameter. Taking expectations of the negative of the derivative with respect to $\boldsymbol{\beta}$ of the above sum evaluated at $\boldsymbol{\beta}_0$ yields the Fisher information matrix

$$\sum_{i=1}^{n} \frac{\mathbf{D}_i(\boldsymbol{\beta}_0)\mathbf{D}_i(\boldsymbol{\beta}_0)^T}{V_i(\boldsymbol{\beta}_0)\tau_0}.$$

Note that the second term involving derivatives of $\mathbf{D}_i/V_i$ drops out due to the assumption that $\mu_i(\boldsymbol{\beta}_0) = \mathrm{E}(Y_i)$. Now for certain misspecification of densities such as overdispersed binomial or Poisson models, the generalized linear model framework allows for estimation of τ and approximately correct inference as long as the mean is modeled correctly and the mean-variance relationship is specified correctly. Details of this robustified inference may be found in McCullagh and Nelder (1989, chap. 9) under the name *quasi-likelihood*. Note, though, that only one extra parameter τ is used to make up for possible misspecification. Instead, Liang and Zeger (1986) noticed that the M-estimator approach could be used here without τ and with only the mean correctly specified:

$$\mathbf{A}_n(\mathbf{X}, \widehat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{D}_i(\widehat{\boldsymbol{\beta}})\mathbf{D}_i(\widehat{\boldsymbol{\beta}})^T}{V_i(\widehat{\boldsymbol{\beta}})},$$

and

$$\mathbf{B}_n(\mathbf{X}, \mathbf{Y}, \widehat{\boldsymbol{\beta}}) = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(Y_i - \mu_i(\widehat{\boldsymbol{\beta}}))^2\mathbf{D}_i(\widehat{\boldsymbol{\beta}})\mathbf{D}_i(\widehat{\boldsymbol{\beta}})^T}{V_i^2(\widehat{\boldsymbol{\beta}})}.$$

Liang and Zeger (1986) proposed a generalized set of estimating equations that accommodates independent clusters of correlated data. The form of the estimating equations and $\mathbf{A}$ and $\mathbf{B}$ matrices are similar to the above except that the sums are over independent clusters. Dunlop (1994) gave a simple introduction to these generalized estimating equations (GEE). In time series and spatial analyses, there is often correlation among all the Y_i with no independent replication. In such cases the $\mathbf{A}$ matrix estimates from the independent case are still consistent, but more complicated methods must be used in estimating the $\mathbf{B}$ matrix; see Lumley and Heagerty (1999) and Kim and Boos (2001).

6. APPLICATION TO TEST STATISTICS

Recall that Wald test statistics for $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ are quadratic forms like $\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T \mathbf{V}_n(\widehat{\boldsymbol{\theta}})^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$. Thus M-estimation is directly useful for creating such statistics. Score statistics are created from the defining equations (1), but the variance estimates used to define them are not as simple to derive by the M-estimation method as Wald statistics. Here we illustrate how to find appropriate variances estimates for score statistics in one application.

Example 10. Testing Equality of Success Probabilities. In the National Basketball Association (NBA) playoffs of 2000, Los Angeles Lakers player Shaquille O'Neal played in 23 games. Table 1 gives his game-by-game free throw outcomes and Figure 2 displays the results. The apparent downward trend in sample proportions is not significant; the simple linear regression p value = .24.

It is often conjectured that players have streaks where they shoot better or worse for an entire game or a significant portion of a game. This can be modeled by assuming that the number of free throws made in the ith game, Y_i, is binomial (n_i, p_i) conditional on n_i, the number of free throws attempted in the ith game, and p_i, the probability of making a free throw in the ith game. Having streaks would be manifest by some games having high or low p_i values. The question of streaks can be addressed statistically by testing

$$\mathrm{H}_0 : p_1 = p_2 = \cdots = p_k = p \quad \text{vs.} \quad \mathrm{H}_1 : p_i \neq p_j \quad \text{for at least one pair } i \neq j$$

where k denotes the number of games. The score statistic for testing a common binomial proportion versus some differences is given by

$$T_S = \sum_{i=1}^{k}(Y_i - n_i\widetilde{p})^2/n_i\widetilde{p}(1-\widetilde{p}),$$

where $\widetilde{p} = \sum Y_i/\sum n_i$ is the estimate of the common value of p under the null hypothesis. The sample sizes $n_1, \ldots, n_k$ were assumed nonrandom for this derivation (they are not really; so this is a conditional approach). T_S is also the simple chi-squared goodness-of-fit statistic with the $2k$ cell expected values $n_1\widetilde{p}, n_1(1-\widetilde{p}), \ldots, n_k\widetilde{p}, n_k(1-\widetilde{p})$.

Using the above data, we find $T_S = 35.51$ and the p value is .034 based on a chi-squared distribution with $k-1 = 22$ degrees of freedom. But the chi-squared approximation is based on each n_i going to infinity, and most of the n_i in our dataset are quite small. An alternative approach uses the normal approximation based on $k \to \infty$. To find the asymptotic variance of T_S using the M-estimator approach, we need to treat the expected value of

This content downloaded from
:ffff:ffff:ffff:ffff:ffff:ffff on Thu, 01 Jan 1976 12:34:56 UTC
All use subject to https://about.jstor.org/terms

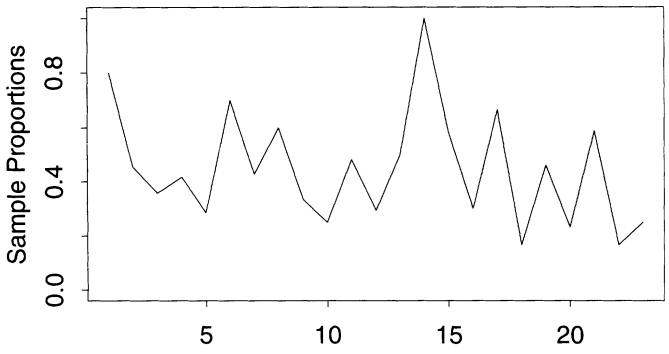

Figure 2. Shaq's Free Throw Percentages in the 2000 NBA Playoffs.

T_S/k as a parameter θ_1, and p as θ_2, and form two ψ functions:

$$\begin{aligned}\psi_1(Y_i, n_i, \theta_1, p) &= \frac{(Y_i - n_i p)^2}{n_i p(1-p)} - \theta_1 \\ \psi_2(Y_i, n_i, \theta_1, p) &= Y_i - n_i p.\end{aligned}$$

For calculating the $\mathbf{A}$ and $\mathbf{B}$ matrices we can treat the n_i as fixed constants as in regression or as random variables with some distribution. Taking the latter approach and noting that $\theta_1 = 1$ under H_0, we get $\mathbf{A}_{11} = 1, \mathbf{A}_{12} = (1-2p)/[p(1-p)], \mathbf{A}_{21} = 0, \mathbf{A}_{22} = \mathrm{E}(n_i) = \mu_n$,

$$\mathbf{B}_{11} = 2 + \frac{(1 - 6p + 6p^2)}{p(1-p)} \mathrm{E}\left(\frac{1}{n_i}\right),$$

$\mathbf{B}_{12} = (1-2p), \mathbf{B}_{22} = \mu_n p(1-p)$. We have used the assumption that conditionally under H_0 that $Y_i | n_i$ is binomial(n_i, p). The asymptotic variance of interest is then

$$\begin{aligned}V_{11} = \left[\mathbf{A}^{-1}\mathbf{B}\{\mathbf{A}^{-1}\}^T\right]_{11} &= \mathbf{B}_{11} - \frac{2\mathbf{A}_{12}\mathbf{B}_{12}}{\mathbf{A}_{22}} + \frac{\mathbf{A}_{12}^2\mathbf{B}_{22}}{\mathbf{A}_{22}^2} \\ &= 2 + \frac{(1 - 6p + 6p^2)}{p(1-p)} \mathrm{E}\left(\frac{1}{n_i}\right) \\ &\quad - \frac{(1-2p)^2}{\mu_n p(1-p)}.\end{aligned}$$

An estimate $\widehat{V}_{11}$ of V_{11} was calculated by plugging in $k^{-1}\sum(1/n_i)$ for $\mathrm{E}(1/n_i)$, $k^{-1}\sum n_i$ for μ_n and $\widetilde{p}$ for p in the expression for V_{11}. Comparing T_S/k to the $\mathrm{N}(1,\ k^{-1}\widehat{V}_{11})$ distribution results in a p value for the Shaq free throw data of .026. We also ran two parametric bootstraps with 10,000 resamples: conditional on $(n_1, \ldots, n_{23})$ yielding p value = .042 and also with the n_i drawn with replacement from $(n_1, \ldots, n_{23})$ yielding p value = .037. So the chi-squared approximation p value is closer to the bootstrap p values than is the p value from the normal approximation.

In summary, there is some evidence that Shaq is "streaky," that is, his foul-shooting probabilities are not constant across games. However, the effect does appear to be strong, and the results are very sensitive to game 14 where Shaq made nine free throws out of nine. Also the related score statistic derived by Tarone (1979) from the beta-binomial model is weighted differently and results in a p value of .25.

7. SUMMARY

M-estimators represent a very large class of statistics including, for example, maximum likelihood estimators and basic sample statistics like sample moments and sample quantiles as well as complex functions of these. The approach we have assimilated from the literature and described herein makes standard error estimation and asymptotic analysis routine regardless of the complexity or dimension of the problem. In summary we would like to concisely present the key features of M-estimators:

1. An M-estimator $\widehat{\boldsymbol{\theta}}$ satisfies (1): $\sum_{i=1}^n \boldsymbol{\psi}(\mathbf{Y}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}$, where $\boldsymbol{\psi}$ is a known function not depending on i or n. See also the extensions (2) and (9).

2. Many estimators that do not satisfy (1) or the extensions (2) and (9) are components of higher-dimensional M-estimators and thus are amenable to M-estimator techniques using the method of stacking. Such estimators are called *partial M-estimators*.

3. $\mathbf{A}(\boldsymbol{\theta}_0) = \mathrm{E}\left[-\partial\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}^T\right]$ is the Fisher information matrix in regular parametric models when $\boldsymbol{\psi}$ is the log-likelihood score function. More generally $\mathbf{A}(\boldsymbol{\theta}_0)$ must have an inverse but need not be symmetric. See also the extension (13) for non-differentiable $\boldsymbol{\psi}$.

4. $\mathbf{B}(\boldsymbol{\theta}_0) = \mathrm{E}\left[\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)^T\right]$ is also the Fisher information matrix in regular parametric models when $\boldsymbol{\psi}$ is the log-likelihood score function. $\mathbf{B}(\boldsymbol{\theta}_0)$ always has the properties of a covariance matrix but will be singular when one component of $\widehat{\boldsymbol{\theta}}$ is a nonrandom function of the other components of $\widehat{\boldsymbol{\theta}}$. In general, $\mathbf{A}(\boldsymbol{\theta}_0) \neq \mathbf{B}(\boldsymbol{\theta}_0)$; the noteworthy exception arises when a parametric model is assumed and the $\boldsymbol{\psi}$ function results in the maximum likelihood estimates.

5. Under suitable regularity conditions, $\widehat{\boldsymbol{\theta}}$ is AMN$(\boldsymbol{\theta}_0, \mathbf{V}(\boldsymbol{\theta}_0)/n)$ as $n \to \infty$, where $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\}^T$ is the sandwich matrix.

6. One generally applicable estimator of $\mathbf{V}(\boldsymbol{\theta}_0)$ for differentiable $\boldsymbol{\psi}$ functions is the empirical sandwich estimator $\mathbf{V}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}}) = \mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})^{-1}\mathbf{B}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})\{\mathbf{A}_n(\mathbf{Y}, \widehat{\boldsymbol{\theta}})^{-1}\}^T$.

[Received March 2001. Revised September 2001.]

REFERENCES

Benichou, J., and Gail, M. H. (1989), "A Delta Method for Implicitly Defined Random Variables," *The American Statistician*, 43, 41–44.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman & Hall.

Dunlop, D. D. (1994), "Regression for Longitudinal Data: A Bridge from Least Squares Regression," *The American Statistician*, 48, 299–303.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Godambe, V. P. (1960),"An Optimum Property of Regular Maximum Likelihood Estimation," *Annals of Mathematical Statistics*, 31, 1208–1211.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

Hodges, J. L. Jr., and Lehmann, E. L. (1963), "Estimates of Location Based on Rank Rests," *Annals of Mathematical Statistics*, 34, 598–611.

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73–101.

——— (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the 5th Berkeley Symposium*, 1, 221–233.

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms

——— (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *Annals of Statistics*, 1, 799–821.

——— (1981), *Robust Statistics*, New York: Wiley.

Iverson, H. K., and Randles, R. H. (1989), "The Effects on Convergence of Substituting parameter Estimates into U-Statistics and Other families of Statistics," *Probability Theory and Related Fields*, 81, 453–471.

Kim, H.-J., and Boos, D. D. (2001), "Variance Estimation in Spatial Regression Using a Nonparametric Semivariogram Based on Residuals," Institute of Statistics Mimeo Series #2524, North Carolina State University, Raleigh.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lumley, T., and Heagerty, P.J. (1999), "Weighted Empirical Adaptive Variance Estimators for Correlated Data Regression," *Journal of the Royal Statistical Society, Ser. B*, 61, 459–477.

McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, London: Chapman and Hall.

Randles, R. H. (1982), "On the Asymptotic Normality of Statistics With Estimated Parameters," *The Annals of Statistics*, 10 462–474.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Tarone, R. E. (1979), "Testing the Goodness of Fit of the Binomial Distribution," *Biometrika*, 66, 585–590.

This content downloaded from
139.18.6.251 on Fri, 20 Nov 2020 10:55:25 UTC
All use subject to https://about.jstor.org/terms