The Calculus of M-Estimation

Leonard A. STEFANSKI and Dennis D. BOOS *

Abstract

Since the groundbreaking papers by Huber in the 1960's, M-estimation methods (estimating equations) have been increasingly important for asymptotic analysis and approximate inference. Now with the prevalence of programs like Maple and Mathematica, calculation of asymptotic variances in complex problems is just a matter of routine manipulation. The intent of this article is to illustrate the depth and generality of the M-estimation approach and thereby facilitate its use.

KEY WORDS: Asymptotic variance; Central Limit Theorem; Estimating equations; Maple; M-estimator.

Institute of Statistics Mimeo Series No. 2528

March 2001

*Leonard A. Stefanski and Dennis D. Boos are Professors, Department of Statistics, North Carolina State University, Raleigh, NC 27695–8203. Email addresses are: stefanski@stat.ncsu.edu,boos@stat.ncsu.edu.

1. INTRODUCTION

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

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

Here we are assuming that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ 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 $\boldsymbol{Y}_i$ represents the ith datum. In some applications it is advantageous to emphasize the dependence of ψ on particular components of $\boldsymbol{Y}_i$. For example, in a regression problem $\boldsymbol{Y}_i = (\boldsymbol{x}_i,\ Y_i)$ and (1) would typically be written

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

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

Huber (1964, 1967) introduced M-estimators and their asymptotic properties, and they were 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). Obviously, many others have made important contributions. For example, Godambe (1960) introduced the concept of an *optimum estimating function* in an M-estimator context, and that paper 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 want students to feel comfortable finding and using the asymptotic approximations that flow from the method. The key advantage is that a very large class of asymptotically normal statistics including delta method transformations can be put in the general M-estimator framework. This unifies large sample approximation methods, simplifies analysis, and makes computations routine.

An important practical consideration is the availability of a symbolic manipulation program

like Maple; otherwise, the matrix calculations can be overwhelming. Nevertheless, the general theory is straightforward.

We claim that many estimators not thought of as M-estimators can be written in the form of M-estimators. Consider as a simple example the mean deviation from the sample mean

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \overline{Y}|.$$

Is this an M-estimator? There is certainly no single equation of the form

$$\sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, \theta) = 0$$

that yields $\widehat{\theta}_1$. Moreover, there is no family of densities $f(y;\boldsymbol{\theta})$ such that $\widehat{\theta}_1$ is a component of the maximum likelihood estimator of $\boldsymbol{\theta}$. But if we let $\psi_1(y,\theta_1,\theta_2) = |y-\theta_2|-\theta_1$ and $\psi_2(y,\theta_1,\theta_2) = y-\theta_2$, then

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

yields $\widehat{\theta}_2 = \overline{Y}$ and $\widehat{\theta}_1 = (1/n)\sum_{i=1}^{n} |Y_i - \overline{Y}|$. We like to use the term "partial M-estimator" for an estimator that is not naturally an M-estimator until additional ψ functions are added. The key idea is simple: any estimator that would be an M-estimator if certain parameters were known, is a partial M-estimator because we can "stack" ψ 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, the sample median. Moreover, we can also add ψ functions to give delta method asymptotic results for transformations like $\widehat{\theta}_3 = \log(\widehat{\theta}_1)$. In this latter context, there are connections to Benichou and Gail (1989).

Certainly the combination of standard influence curve and "delta theorem" methodology can handle a larger class of problems than this enhanced M-estimation approach. However, we believe that the combination of a single approach along with a symbolic manipulator like Maple will make

this M-estimation approach much more likely to be used in complex problems.

A description of the basic approach is given 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. Extensions for regression are given in Section 5. A discussion of some testing problems is given in Section 6, and Section 7 summarizes the key features of the M-estimator method.

2. The Basic Approach

M-estimators solve (1), where the vector function $\boldsymbol{\psi}$ must be a known function that does not depend on i or n. For regression situations, the argument of $\boldsymbol{\psi}$ will be expanded to depend on regressors $\boldsymbol{x}_i$, but the basic $\boldsymbol{\psi}$ will still not depend on i. For the moment we will 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 clearly the population mean $\theta_0 = \int y dF(y)$ is the unique solution of $\int (y - \theta) dF(y) = 0$.

If there is one unique $\boldsymbol{\theta}_0$ satisfying (3), then in general there exists a sequence of M-estimators $\widehat{\boldsymbol{\theta}}$ such that the weak law of large numbers leads to $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ as $n \to \infty$. Furthermore, if $\boldsymbol{\psi}$ is suitably smooth, then Taylor expansion of $\boldsymbol{G}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \boldsymbol{\psi}(Y_i, \boldsymbol{\theta})$ gives

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

where $\boldsymbol{G}_n'(\boldsymbol{\theta}_0) = \left[\partial \boldsymbol{G}_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$. For n sufficiently large, we expect $\boldsymbol{G}_n'(\boldsymbol{\theta}_0)$ to be nonsingular so that we can rearrange and get:

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

Under suitable regularity conditions as $n \to \infty$,

$$-\boldsymbol{G}_n'(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\partial}{\partial \boldsymbol{\theta}^T}\boldsymbol{\psi}(Y_i,\boldsymbol{\theta}_0)\right] \xrightarrow{p} \mathrm{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}^T}\boldsymbol{\psi}(Y_1,\boldsymbol{\theta}_0)\right] = \boldsymbol{A}(\boldsymbol{\theta}_0). \tag{5}$$

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

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

If $\boldsymbol{A}(\boldsymbol{\theta}_0)$ exists, the Weak Law of Large Numbers gives (5). If $\boldsymbol{B}(\boldsymbol{\theta}_0)$ exists, then (6) follows from the Central Limit Theorem. The hard part to prove is (7). Huber(1967) was the first to give general results for (7), but there have been many others since then. We shall be content to observe that (7) holds in most regular situations where there are sufficent smoothness conditions on $\boldsymbol{\psi}$, and $\boldsymbol{\theta}$ has fixed dimension p as $n \to \infty$.

Putting (1) and (4)–(7) together with Slustky's Theorem, we have that

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

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

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

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

where $\boldsymbol{c}_n/\sqrt{n} \xrightarrow{p} \mathbf{0}$ as $n \to \infty$. Following the above arguments and noting that $\boldsymbol{c}_n/\sqrt{n}$ is absorbed in $\sqrt{n}\boldsymbol{R}_n^*$ of (4), we can see that as long as (9) and (4)–(7) hold, then (8) will also hold. This extension allows us to cover a much wider class of statistics including empirical quantiles, 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 truly come from the assumed parametric family $f(y;\boldsymbol{\theta})$, then $\boldsymbol{A}(\boldsymbol{\theta}_0) = \boldsymbol{B}(\boldsymbol{\theta}_0) =$

$I(\boldsymbol{\theta}_0)$, the information matrix. In this case the sandwich matrix $V(\boldsymbol{\theta}_0)$ reduces to the usual $I(\boldsymbol{\theta}_0)^{-1}$. One of the key contributions of M-estimation theory has been to point out what happens 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 $\boldsymbol{A}(\boldsymbol{\theta}_0) \neq \boldsymbol{B}(\boldsymbol{\theta}_0)$, and valid inference should be carried out using the correct limiting covariance matrix $V(\boldsymbol{\theta}_0) = \boldsymbol{A}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_0)\{\boldsymbol{A}(\boldsymbol{\theta}_0)^{-1}\}^T$, not $I(\boldsymbol{\theta}_0)^{-1}$.

Using the left-hand-side of (5), we define the empirical estimator of $\boldsymbol{A}(\boldsymbol{\theta}_0)$ by

$$\boldsymbol{A}_n(\boldsymbol{Y}, \widehat{\boldsymbol{\theta}}) = \boldsymbol{G}_n'(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\partial}{\partial \boldsymbol{\theta}^T}\boldsymbol{\psi}(Y_i, \widehat{\boldsymbol{\theta}})\right].$$

Note that for maximum likelihood estimation, $n\boldsymbol{A}_n(\widehat{\boldsymbol{\theta}})$ is the observed information matrix $\boldsymbol{I}_{\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}})$. Similarly, the empirical estimator of $\boldsymbol{B}(\boldsymbol{\theta}_0)$ is

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

Putting these together yields the empirical sandwich estimator

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

$\boldsymbol{V}_n(\boldsymbol{Y}, \widehat{\boldsymbol{\theta}})$ will generally be consistent for $V(\boldsymbol{\theta}_0)$ under mild regularity conditions (see Iverson and Randles (1989) for a general theorem on this convegence).

$\boldsymbol{V}_n(\boldsymbol{Y}, \widehat{\boldsymbol{\theta}})$ requires no analytic work beyond specifying the $\boldsymbol{\psi}$ function. In some problems, it is simpler to work directly with the limiting form $V(\boldsymbol{\theta}_0) = \boldsymbol{A}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_0)\{\boldsymbol{A}(\boldsymbol{\theta}_0)^{-1}\}^T$ and just plug in estimators for $\boldsymbol{\theta}_0$ and any other unknown quantities in $\boldsymbol{V}(\boldsymbol{\theta}_0)$. The notation $\boldsymbol{V}(\boldsymbol{\theta}_0)$ suggests that $\boldsymbol{\theta}_0$ is the only unknown quantity in $\boldsymbol{V}(\boldsymbol{\theta}_0)$, but in reality $\boldsymbol{V}(\boldsymbol{\theta}_0)$ often involves higher moments or other characteristics of the true distribution function F of Y_i. In fact there is a range of possibilities for estimating $\boldsymbol{V}(\boldsymbol{\theta}_0)$ depending on what model assumptions are used. For simplicity, we will just use the notation $\boldsymbol{V}_n(\boldsymbol{Y}, \widehat{\boldsymbol{\theta}})$ for the purely empirical estimator and $\boldsymbol{V}(\widehat{\boldsymbol{\theta}})$ for any of the expected value plus model assumption versions.

For maximum likelihood estimation with a correctly specified family, the three competing

estimators for $I(\boldsymbol{\theta})^{-1}$ are $V_n(\boldsymbol{Y}, \widehat{\boldsymbol{\theta}})$, $\left[I_{\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}})/n\right]^{-1} = \boldsymbol{A}_n(\boldsymbol{Y}, \widehat{\boldsymbol{\theta}})^{-1}$, and $I(\widehat{\boldsymbol{\theta}})^{-1} = \boldsymbol{V}(\widehat{\boldsymbol{\theta}})$. In this case the standard estimators $\left[I_{\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}})/n\right]^{-1}$ and $I(\widehat{\boldsymbol{\theta}})^{-1}$ are generally more efficient than $\boldsymbol{V}_n(\boldsymbol{Y}, \widehat{\boldsymbol{\theta}})$ for estimating $I(\boldsymbol{\theta})^{-1}$. (Clearly nothing can have smaller asymptotic variance for estimating $I(\boldsymbol{\theta})^{-1}$ than $I(\widehat{\boldsymbol{\theta}}_{\text{MLE}})^{-1}$.)

Now we illustrate these ideas with examples.

Example 1 Let $\widehat{\boldsymbol{\theta}} = (\overline{Y}, s_n^2)^T$, the sample mean and variance. Here

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

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 $\widehat{\theta}_2 = s_n^2 = n^{-1}\sum(Y_i - \overline{Y})^2$, when considered 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 our definition of a partial M-estimator.

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

$$\boldsymbol{A}(\boldsymbol{\theta}_0) = \mathrm{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}^T}\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)\right] = \mathrm{E}\begin{pmatrix} 1 & 0 \\ 2(Y_1 - \theta_{10}) & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{aligned} \boldsymbol{B}(\boldsymbol{\theta}_0) &= \mathrm{E}\left[\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)^T\right] \\ &= \mathrm{E}\begin{pmatrix} (Y_1 - \theta_{10})^2 & (Y_1 - \theta_{10})\left[(Y_1 - \theta_{10})^2 - \theta_{20}\right] \\ (Y_1 - \theta_{10})\left[(Y_1 - \theta_{10})^2 - \theta_{20}\right] & \left[(Y_1 - \theta_{10})^2 - \theta_{20}\right]^2 \end{pmatrix} \\ &= \begin{pmatrix} \theta_{20} & \mu_3 \\ \mu_3 & \mu_4 - \theta_{20}^2 \end{pmatrix} = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}, \end{aligned}$$

where μ_k is our notation for the kth central moment of Y_1 and the more familiar notation $\sigma^2 = \theta_{20}$ has been substituted at the end. In this case, since $\boldsymbol{A}(\boldsymbol{\theta}_0)$ is the identity matrix, $V(\boldsymbol{\theta}_0) = \boldsymbol{B}(\boldsymbol{\theta}_0)$.

To estimate $\boldsymbol{B}(\boldsymbol{\theta}_0)$, we may use

$$\begin{aligned}\boldsymbol{B}_n(\boldsymbol{Y},\widehat{\boldsymbol{\theta}}) &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} (Y_i-\overline{y})^2 & (Y_i-\overline{Y})\left[(Y_i-\overline{Y})^2-s_n^2\right] \\ (Y_i-\overline{Y})\left[(Y_i-\overline{Y})^2-s_n^2\right] & \left[(Y_i-\overline{Y})^2-s_n^2\right]^2 \end{pmatrix} \\ &= \begin{pmatrix} s_n^2 & m_3 \\ m_3 & m_4-s_n^4 \end{pmatrix},\end{aligned}$$

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

Note that $\widehat{\boldsymbol{\theta}}$ is a maximum likelihood estimator for the normal model density $f(y;\boldsymbol{\theta}) = (2\pi\theta_2)^{-1}\exp(-(y-\theta_1)^2/2\theta_2)$, but $\psi_1 = Y_i-\theta_1$ and $\psi_2 = (Y_i-\theta_1)^2-\theta_2$ are not the score functions that come from this normal density. The partial derivative of this normal log density yields $\psi_1 = (Y_i-\theta_1)/\theta_2$ 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 can lead to the same estimator. Of course different $\boldsymbol{\psi}$ functions associated with the same estimator yield different $\boldsymbol{A}$ and $\boldsymbol{B}$ but the same $\boldsymbol{V}$. For example, using these latter two ψ functions, the resulting $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices are

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

If we further assume that the data truly are normally distributed, then $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$ resulting in $\boldsymbol{A}(\boldsymbol{\theta}_0) = \boldsymbol{B}(\boldsymbol{\theta}_0) = \boldsymbol{I}(\boldsymbol{\theta}_0) = \mathrm{Diag}(1/\sigma^2, 1/2\sigma^4)$. Here the expected value model-based covariance estimator would be $\boldsymbol{V}(\widehat{\boldsymbol{\theta}}) = \mathrm{Diag}(1/s_n^2, 1/2s_n^4)$.

Note that the likelihood score ψ functions, $\boldsymbol{\psi}_{\mathrm{MLE}}$, are related to the original ψ functions by $\boldsymbol{\psi}_{\mathrm{MLE}} = \boldsymbol{C}\psi$, where $\boldsymbol{C} = \mathrm{diag}(1/\theta_{20}, 1/2\theta_{20}^2)$. A little algebra shows that all ψ of the form $\boldsymbol{C}\psi$, where $\boldsymbol{C}$ is nonsingular (but possibly depending on $\boldsymbol{\theta}_0$ and $Y_1,\ldots,Y_n$), lead to an equivalence class having the same estimator and asymptotic variance matrix $V(\boldsymbol{\theta}_0)$.

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 $\mathrm{E}X_1 = \mu_X$, 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 ψ function for $\widehat{\theta} = \overline{Y}/\overline{X}$ is $\psi(Y_i, X_i, \theta) = Y_i - \theta X_i$ leading to $\boldsymbol{A}(\boldsymbol{\theta}_0) = \mu_X$, $\boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{E}(Y_1 - \theta_0 X_1)^2$, $\boldsymbol{V}(\boldsymbol{\theta}_0) = \mathrm{E}(Y_1 - \theta_0 X_1)^2/\mu_X^2$, $\boldsymbol{A}_n(\boldsymbol{Y}, \widehat{\theta}) = \overline{X}$, and

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

and

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

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

Now consider the following $\boldsymbol{\psi}$ of dimension 3 that yields $\widehat{\theta}_3 = \overline{Y}/\overline{X}$ as the third component of $\widehat{\boldsymbol{\theta}}$:

$$\boldsymbol{\psi}(Y_i, X_i, \boldsymbol{\theta}) = \begin{pmatrix} Y_i - \theta_1 \\ X_i - \theta_2 \\ \theta_1 - \theta_3\theta_2 \end{pmatrix}$$

This is a quite interesting $\boldsymbol{\psi}$ function because the third component does not have any data involved in it. Nevertheless, this $\boldsymbol{\psi}$ satisfies all the requirements of a ψ function and illustrates how to build the "delta" method into the M-estimator framework. The A and B matrices are

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

This example illustrates the fact that $\boldsymbol{B}(\boldsymbol{\theta}_0)$ can be singular (although $\boldsymbol{A}(\boldsymbol{\theta}_0)$ generally cannot). In fact whenever a $\boldsymbol{\psi}$ funtion has components that involve no data, then the resulting B matrix will be singular. In Maple we computed $V(\boldsymbol{\theta}_0) = \boldsymbol{A}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{B}(\boldsymbol{\theta}_0)\{\boldsymbol{A}(\boldsymbol{\theta}_0)^{-1}\}^T$, and obtained for the (3,3) element

$$v_{33} = \frac{1}{\theta_{20}^2}\left[\sigma_Y^2 - 2\theta_{30}\sigma_{YX} + \theta_{30}^2\sigma_X^2\right].$$

This latter expression for the asymptotic variance of $\sqrt{n}\widehat{\theta}_3$ can be shown to be the same as

$\mathrm{E}(Y_1 - \theta_{30}X_1)^2/\mu_X^2$ obtained earlier upon noting that $\theta_{20} = \mu_X$.

Sample Maple Program

```
with(linalg):                                   Brings in the linear algebra package
vA:=[1,0,0,0,1,0,-1,theta[3],theta[2]];         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]);
```

$$\frac{{\sigma_y}^2 - 2\,\theta_3\,\sigma_{xy} + {\theta_3}^2\,{\sigma_x}^2}{{\theta_2}^2}$$

The above display is what appears on the Maple window for the last command.

Example 3 Further illustration of the "delta method." In the context of Example 1, suppose we are interested in $s_n = \sqrt{s_n^2}$ and $\log(s_n^2)$. We could of course just redefine θ_2 in Example 1 to be θ_2^2 and $\exp(\theta_2)$, respectively. Instead, we prefer to 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 seems conceptually simpler and it also gives the joint asymptotic distribution of all quantities. Now we have

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

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

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

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 *Posterior Mode.* Consider the standard Bayesian model in an iid framework where the posterior density is proportional to

$$\pi(\boldsymbol{\theta}) \prod_{i=1}^{n} f(Y_i|\boldsymbol{\theta}),$$

and π is the prior density. Posterior *mode* estimators satisfy (9) with $\boldsymbol{\psi}(y, \boldsymbol{\theta}) = \partial \log f(y|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ the same as for maximum likelihood and $\boldsymbol{c}_n = -\pi'(\widehat{\boldsymbol{\theta}})/\pi(\widehat{\boldsymbol{\theta}})$. Thus, as long as $\boldsymbol{c}_n/\sqrt{n} \xrightarrow{p} \mathbf{0}$, the Bayesian mode estimator will have the same asymptotic covariance matrix as maximum likelihood estimators.

Example 5 *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 some features of the M-estimation approach. Suppose that triples (Y_i, W_i, T_i) are observed such that

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \sigma_\varepsilon \varepsilon_{1,i} \\ W_i &= X_i + \sigma_U \varepsilon_{2,i} \\ T_i &= \gamma + \delta X_i + \sigma_\tau \varepsilon_{3,i} \end{aligned}$$

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,

$$\widehat{\beta}_{\mathrm{IV}} = \frac{\widehat{\beta}_{Y|T}}{\widehat{\beta}_{W|T}},$$

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 the Introduction, 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

$$\psi(Y,\ W,\ T,\ \theta) = \begin{pmatrix} \theta_1 - T \\ (Y - \theta_2 W)(\theta_1 - T) \end{pmatrix},$$

with associated M-estimator,

$$\widehat{\theta}_1 = \overline{T}, \qquad \widehat{\theta}_2 = \widehat{\beta}_{\mathrm{IV}}.$$

The $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices calculated at the true parameters assuming the instrumental variable model are

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ \alpha & \sigma_{X,\,T} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{B} = \begin{pmatrix} \sigma_T^2 & \alpha\sigma_T^2 \\ \alpha\sigma_T^2 & \sigma_T^2(\alpha^2 + \sigma_\epsilon^2 + \beta^2\sigma_U^2) \end{pmatrix},$$

which yield the asymptotic variance matrix

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

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 when there is doubt about the magnitude of σ_U^2, their joint asymptotic distribution is of interest. The M-estimator approach easily accommodates such calculations. For this task consider the ψ function

$$\boldsymbol{\psi}(Y,\ W,\ T,\ \theta) = \begin{pmatrix} \theta_1 - T \\ \theta_2 - W \\ (Y - \theta_3 W)(\theta_2 - W) \\ (Y - \theta_4 W)(\theta_1 - T) \end{pmatrix}.$$

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

The M-estimator associated with this $\boldsymbol{\psi}$ function is

$$\widehat{\theta}_1 = \overline{T}, \quad \widehat{\theta}_2 = \overline{W}, \quad \widehat{\theta}_3 = \widehat{\beta}_{Y|W}, \quad \widehat{\theta}_4 = \widehat{\beta}_{\text{IV}}.$$

The $\boldsymbol{A}$ 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 $\boldsymbol{B}$ matrix is unwieldy. However, primary interest lies in the lower 2×2 submatrix of the asymptotic variance matrix $\boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{A}^{-1}\right)^T$. 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 $(\widehat{\theta}_3,\ \widehat{\theta}_4)$,

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

The variance formula given above assumes normality of the errors $\varepsilon_{j,i}$ and the X_i in the model. Instrumental variable estimation works more generally and in the absence of distributional assumptions (beyond those of lack of correlation) estimated variances can be obtained using the sandwich formula. We illustrate the calculations with data from the Framingham Heart Study. For this illustration Y and W are systolic blood pressure and serum cholesterol respectively measured at the third exam, and T is serum cholesterol respectively measured at the second exam. The data include measurements on $n = 1615$ males aged 45 to 65.

The 4×1 $\boldsymbol{\psi}$ function was used to determine the estimates (standard errors in parentheses)

$$\begin{aligned} \widehat{\theta}_1 &= \overline{T} = 227.2(1.1), & \widehat{\theta}_2 &= \overline{W} = 228.4(1.0), \\ \widehat{\theta}_3 &= \widehat{\beta}_{Y|W} = 0.042(0.011), & \widehat{\theta}_4 &= \widehat{\beta}_{\mathrm{IV}} = 0.065(0.015). \end{aligned}$$

The empirical sandwich variance estimate (direct computer output) is

```
 1785.8453     1291.8722     -1.3658812     -3.8619519
 1291.8722     1718.3129     -1.1578449     -2.6815324
-1.3658812    -1.1578449     0.20737770     0.19878711
-3.8619519    -2.6815324     0.19878711     0.35584612
```

The estimated contrast $\widehat{\beta}_{\mathrm{IV}} - \widehat{\beta}_{Y|W} = 0.023$ has standard error 0.010, resulting in the test statistic $t = 2.29$. The test statistic statistic is consistent with the hypothesis that serum cholesterol is measured with non-negligible error.

3. CONNECTIONS TO THE INFLUENCE CURVE

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

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

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

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

Since the Influence Curve approach is more general, why use the M-estimator approach of this paper? Although the two methods have similarities, we have found that the M-estimator approach described in this paper is easier to routinely use as a "plug and crank" method. Especially in messy problems with a large number of parameters, it appears easier to stack ψ functions and compute $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices than it is to compute and then stack influence curves and then compute Σ. This may be largely a matter of taste, but we have seen resistance to using influence curves.

If one has already computed influence curves, then defining $\boldsymbol{\psi}(Y_i,\boldsymbol{\theta}) = \boldsymbol{IC}_{\widehat{\boldsymbol{\theta}}}(Y_i,\boldsymbol{\theta}_0) - (\theta - \theta_0)$ allows one to use the approach of this paper. In this case $\boldsymbol{A}(\boldsymbol{\theta}_0)$ is the identity matrix and $\boldsymbol{B}(\boldsymbol{\theta}_0) = \Sigma$. (A minor modification is that for the empirical variance estimators we need to define $\boldsymbol{\psi}(Y_i,\widehat{\boldsymbol{\theta}}) = \boldsymbol{IC}_{\widehat{\boldsymbol{\theta}}}(Y_i,\widehat{\boldsymbol{\theta}})$; that is, 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 6 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

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

It is not clear how to put this estimator directly in the M-estimator framework, but for distributions symmetic 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) gives

$$IC_{\widehat{\theta}_{HL}}(y;\theta_0) = \frac{F_0(y-\theta_0) - \frac{1}{2}}{\int f_0^2(y)dy},$$

where $f_0(y)$ is the density function of $F_0(y)$. The variance of this influence curve is

$$\frac{1}{12\left[\int f_0^2(y)dy\right]^2},$$

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 $\widehat{\theta}_{HL}$ and any set of M-estimators, we can stack $\psi(Y_i,\theta) = IC_{\widehat{\theta}_{HL}}(y;\theta_0) - (\theta - \theta_0)$ with the $\boldsymbol{\psi}$ functions of the M-estimators. The part of the $\boldsymbol{A}$ matrix associated with $\widehat{\theta}_{HL}$ will be all zeroes except for the diagonal element which will be a one. The diagonal element of the $\boldsymbol{B}$ matrix will be the asymptotic variance given above, but one will still need to compute correlations of $IC_{\widehat{\theta}_{HL}}(Y_1,\theta_0)$ with the other ψ functions to get the off-diagonal elements of the $\boldsymbol{B}$ matrix involving $\widehat{\theta}_{HL}$.

4. NONSMOOTH ψ FUNCTIONS

In some situations the $\boldsymbol{\psi}$ function may not be differentiable everywhere, thus causing a problem with the definition of the $\boldsymbol{A}$ matrix as the expected value of a derivative that does not exist. The modified definition of $\boldsymbol{A}$ is to just interchange the order of taking the derivative and then the expectation:

$$\boldsymbol{A}(\boldsymbol{\theta}_0) \equiv -\frac{\partial}{\partial \boldsymbol{\theta}^T}\left\{\mathrm{E}_F \boldsymbol{\psi}(Y_1,\boldsymbol{\theta})\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}. \tag{11}$$

It is important to note that the expectation is taken with respect to the true distribution of the data (denoted by E_F), but $\boldsymbol{\theta}$ within the $\boldsymbol{\psi}$ function is free to change in order to take the derivative. Of course after taking the derivative, we then substitute the true parameter value $\boldsymbol{\theta}_0$.

Example 7 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

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

This ψ function is continuous everywhere but not differentiable at $\pm k$. Thus we use definition (11) to calculate $A(\theta_0)$:

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

The notation ψ_k' inside the integral stands for the derivative of ψ_k where it exists, and for the two points where it doesn't exist ($y - \theta_0 = \pm k$), we delete $y = \theta_0 \pm k$ from the integral assuming that F is continuous at those points.

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

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

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

Example 8 The sample pth quantile $\widehat{\theta} = F_n^{-1}(p)$ satisfies $\sum \left[p - I(Y_i \le \widehat{\theta})\right] = c_n$, where $-c_n = n\left[F_n(\widehat{\theta}) - p\right] \le 1$. Thus the ψ function is $\psi(Y_i, \theta) = p - I(Y_i \le \theta)$ and we are using our extended definition (9).

This ψ function is discontinuous at θ_0, but we shall see that definition (11) of $A(\theta_0)$ continues

to give us the correct asymptotic results:

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

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

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

Also, we could easily stack any finite number of quantile ψ 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 ψ functions: we no longer can use $A_n(\boldsymbol{Y}, \widehat{\theta})$ to estimate $\boldsymbol{A}(\theta_0)$. In fact, the derivative of the pth quantile ψ function is zero everywhere except at the location of the jump discontinuity. There are several options for estimating $\boldsymbol{A}(\theta_0)$. One is to use a smoothing technique to estimate f (kernel density estimators, for example). Another is to approximate ψ by a smooth ψ function and use the $\boldsymbol{A}(\theta_0)$ from this smooth approximation.

Example 9 The positive mean deviation from the median is defined to be

$$\widehat{\theta}_1 = \frac{2}{n}\sum_{i=1}^{n}(Y_i - \widehat{\theta}_2)I(Y_i > \widehat{\theta}_2),$$

where $\widehat{\theta}_2$ is the sample median. Thus the $\boldsymbol{\psi}$ function is

$$\boldsymbol{\psi}(Y_i, \boldsymbol{\theta}) = \begin{pmatrix} 2(Y_i - \theta_2)I(Y_i > \theta_2) - \theta_1 \\ \frac{1}{2} - I(Y_i \le \theta_2) \end{pmatrix}.$$

Notice that the first component of $\boldsymbol{\psi}$ is continous everywhere but not differentiable at $\theta_2 = Y_i$. The second component has a jump discontinuity at $\theta_2 = Y_i$. To get $\boldsymbol{A}(\boldsymbol{\theta}_0)$, we first calculate the expected value of $\boldsymbol{\psi}(Y_1, \boldsymbol{\theta})$ (note that $\boldsymbol{\theta}$ is not $\boldsymbol{\theta}_0$):

$$\mathrm{E}_F\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}) = \begin{pmatrix} 2\int_{\theta_2}^{\infty}(y - \theta_2)dF(y) - \theta_1 \\ \frac{1}{2} - F(\theta_2) \end{pmatrix}.$$

To take derivatives of the first component, let us write $dF(y)$ as $f(y)dy$ and expand it out to get

$$2\int_{\theta_2}^{\infty} yf(y)\,dy - 2\theta_2\int_{\theta_2}^{\infty} f(y)\,dy - \theta_1 = 2\int_{\theta_2}^{\infty} yf(y)\,dy - 2\theta_2\left[1-F(\theta_2)\right] - \theta_1.$$

The derivative of this latter expression with respect to θ_1 is of course -1. The derivative with respect to θ_2 is $-2\theta_2 f(\theta_2) - 2\left[1-F(\theta_2)\right] + 2\theta_2 f(\theta_2) = -2\left[1-F(\theta_2)\right]$ (using the Fundamental Theorem of Calculus to get the first term). Setting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ means that $F(\theta_{20}) = 1/2$ because θ_{20} is the population median. Thus the derivative of the first component with respect to θ_2 and evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is just -1. The partial derivatives of the second component evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ are 0 and $-f(\theta_{20})$, respectively. Thus

$$\boldsymbol{A}(\boldsymbol{\theta}_0) = \begin{pmatrix} 1 & 1 \\ 0 & f(\theta_{20}) \end{pmatrix}.$$

Straightforward calculations for $\boldsymbol{B}(\boldsymbol{\theta}_0)$ yield

$$\boldsymbol{B}(\boldsymbol{\theta}_0) = \begin{pmatrix} b_{11} & \dfrac{\theta_{10}}{2} \\ \dfrac{\theta_{10}}{2} & \dfrac{1}{4} \end{pmatrix},$$

where $b_{11} = 4\int_{\theta_{20}}^{\infty}(y-\theta_{20})^2 f(y)\,dy - \theta_{10}^2$. Finally, $V(\boldsymbol{\theta}_0)$ is given by

$$V(\boldsymbol{\theta}_0) = \begin{pmatrix} b_{11} - \dfrac{\theta_{10}}{f(\theta_{20})} + \dfrac{1}{4f^2(\theta_{20})} & \dfrac{\theta_{10}}{2f(\theta_{20})} - \dfrac{1}{4f^2(\theta_{20})} \\ \dfrac{\theta_{10}}{2f(\theta_{20})} - \dfrac{1}{4f^2(\theta_{20})} & \dfrac{1}{4f^2(\theta_{20})} \end{pmatrix}.$$

5. REGRESSION M-ESTIMATORS

There are two situations of interest for M-estimator analysis of regression estimators. The first is where the independent variables are random variables and we can think in terms of iid $(\boldsymbol{X}, Y)$ pairs. This situation fits into our basic theory developed in Section 2 for iid sampling; see Example 5. The second situation is where the independent variables are fixed constants. This covers standard regression models as well as multi-sample problems like the one-way analysis of variance setup. For this second regression situation we need to introduce new notation to handle the non-iid character

of the problem.

A fairly simple setting to introduce notation is the nonlinear model

$$Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i \qquad i = 1, \ldots, n, \tag{12}$$

where g is a known differentiable function and $e_1, \ldots, e_n$ are independent with mean 0 and possibly unequal variances $\mathrm{Var}(e_i) = \sigma_i^2$, $i = 1, \ldots, n$, and the $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are known constant vectors. As usual we put the vectors together and define $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$. The least squares estimator satisfies

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

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

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left[\frac{1}{n} \sum_{i=1}^{n} -\boldsymbol{\psi}'(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}_0) \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1} \boldsymbol{\psi}(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}_0) + \sqrt{n} R_n^*, \tag{13}$$

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

$$\begin{aligned} A_n(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\beta}_0) &= \frac{1}{n} \sum_{i=1}^{n} \left[-\boldsymbol{\psi}'(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}_0) \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[g'(\boldsymbol{x}_i, \boldsymbol{\beta}_0) g'(\boldsymbol{x}_i, \boldsymbol{\beta}_0)^T - (Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta}_0)) g''(\boldsymbol{x}_i, \boldsymbol{\beta}_0) \right]. \end{aligned} \tag{14}$$

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

$$\begin{aligned} A_n(\boldsymbol{X}, \boldsymbol{\beta}_0) &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{E} \left[-\boldsymbol{\psi}'(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}_0) \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} g'(\boldsymbol{x}_i, \boldsymbol{\beta}_0) g'(\boldsymbol{x}_i, \boldsymbol{\beta}_0)^T. \end{aligned} \tag{15}$$

Notice that we have dropped out the $\boldsymbol{Y}$ from this quantity's name because the expectation elimi-

nates dependence on the Y_i. Also note that the second term for the least squares estimator drops out because of the modeling assumption (12). Finally, assuming that the limit exist, we define

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{\beta}_0) &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[-\boldsymbol{\psi}'(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}_0)\right] \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g'(\boldsymbol{x}_i, \boldsymbol{\beta}_0) g'(\boldsymbol{x}_i, \boldsymbol{\beta}_0)^T. \end{aligned} \tag{16}$$

In the linear regression case, note that $\boldsymbol{A}(\boldsymbol{\beta}_0) = \lim_{n\to\infty} \boldsymbol{X}^T\boldsymbol{X}/n$. This limit need not exist for the least squares estimator to be consistent and asymptotically normal, but it's existence is a typical assumption leading to those desired results. Definition (14) leads to the purely empirical estimator of $A_(\boldsymbol{\beta}_0)$:

$$\begin{aligned} A_n(\boldsymbol{X}, \boldsymbol{Y}, \widehat{\boldsymbol{\beta}}) &= \frac{1}{n}\sum_{i=1}^{n} \left[-\boldsymbol{\psi}'(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[g'(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) g'(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})^T - (Y_i - g(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}))g''(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})\right]. \end{aligned} \tag{17}$$

Since this is the negative of the Hessian in a final Newton iteration, this is sometimes preferred on computational grounds. But the estimated expected value estimator based on (15) is typically simpler:

$$\begin{aligned} A_n(\boldsymbol{X}, \widehat{\boldsymbol{\beta}}) &= \frac{1}{n}\sum_{i=1}^{n} \left\{\mathrm{E}\left[-\boldsymbol{\psi}'(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}_0)\right]\right\}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \\ &= \frac{1}{n}\sum_{i=1}^{n} g'(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) g'(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})^T. \end{aligned} \tag{18}$$

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

$$\begin{aligned} B_n(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\beta}_0) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\boldsymbol{\psi}(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}_0)\boldsymbol{\psi}(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}_0)^T \\ &= \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 g'(\boldsymbol{x}_i, \boldsymbol{\beta}_0) g'(\boldsymbol{x}_i, \boldsymbol{\beta}_0)^T. \end{aligned} \tag{19}$$

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

$$\begin{aligned} B_n(\boldsymbol{X}, \boldsymbol{Y}, \widehat{\boldsymbol{\beta}}) &= \frac{1}{n-p} \sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) \boldsymbol{\psi}(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})^T \\ &= \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - g(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}))^2 g'(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) g'(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})^T. \end{aligned} \tag{20}$$

Example 10 Huber (1973) discussed robust regression alternatives to least squares in the linear regression context. As a specific example, consider the linear model (12) with $g(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{x}^T \boldsymbol{\beta}$ and estimator of $\boldsymbol{\beta}$ satisfying

$$\sum_{i=1}^{n} \psi_k(Y_i - \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}) \boldsymbol{x}_i = 0, \tag{21}$$

where ψ_k is the "Huber" ψ function defined in Example 7. This is a slight abuse of notation since the official $\boldsymbol{\psi}(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}) = \psi_k(Y_i - \boldsymbol{x}^T \boldsymbol{\beta}) \boldsymbol{x}_i$; i.e., ψ is being used as both the original Huber function ψ_k and also as the generic estimating equation function. Since ψ_k is an odd function about zero, the defining equations $\mathrm{E}\psi_k(Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_0) \boldsymbol{x}_i = 0$ will be satisfied if the e_i have a symmetric distribution about zero. If the e_i are not symmetrically distributed and the X matrix contains a column of ones, then the intercept estimated by $\widehat{\boldsymbol{\beta}}$ will be different from that of least squares, but the other components of $\boldsymbol{\beta}_0$ will be the same.

Taking a derivative, we have

$$A_n(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\beta}_0) = \frac{1}{n} \sum_{i=1}^{n} \left[-\boldsymbol{\psi}'(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}_0) \right] = \frac{1}{n} \sum_{i=1}^{n} \psi_k'(e_i) \boldsymbol{x}_i \boldsymbol{x}_i^T$$

and $A_n(\boldsymbol{X}, \boldsymbol{\beta}_0) = n^{-1} \sum_{i=1}^{n} \mathrm{E}\psi_k'(e_i) \boldsymbol{x}_i \boldsymbol{x}_i^T$. Also, $B_n(\boldsymbol{X}, \boldsymbol{\beta}_0) = n^{-1} \sum_{i=1}^{n} \mathrm{E}\psi(e_i)^2 \boldsymbol{x}_i \boldsymbol{x}_i^T$. If we make the homogeneity assumption that the errors $e_1, \ldots, e_n$ all have the same distribution, then $A_n(X, \boldsymbol{\beta}_0) = \mathrm{E}\psi_k'(e_1) \boldsymbol{X}^T \boldsymbol{X}/n$, $B_n(\boldsymbol{X}, \boldsymbol{\beta}_0) = \mathrm{E}\psi_k(e_1)^2 \boldsymbol{X}^T \boldsymbol{X}/n$, and $V(\boldsymbol{X}, \boldsymbol{\beta}_0) = (\boldsymbol{X}^T \boldsymbol{X}/n)^{-1} \mathrm{E}\psi_k(e_1)^2 / [\mathrm{E}\psi_k'(e_1)]^2$.

Example 11 Generalized linear models have score equations

$$\sum_{i=1}^{n} \boldsymbol{D}_i(\boldsymbol{\beta}) \frac{(Y_i - \mu_i(\boldsymbol{\beta}))}{V_i(\boldsymbol{\beta}) \tau} = 0, \tag{22}$$

where $\mu_i(\boldsymbol{\beta}_0) = \mathrm{E}(Y_i) = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0)$, $\boldsymbol{D}_i(\boldsymbol{\beta}) = \partial \mu_i(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$, $V_i(\boldsymbol{\beta}_0) \tau_0 = \mathrm{Var}(Y_i)$, g is the link function,

and τ is an additional 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

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

Note that the second term involving derivatives of $\boldsymbol{D}_i/V_i$ drops out due to the assumption that $\mu_i(\boldsymbol{\beta}_0) = \mathrm{E}(Y_i)$. Now for certain misspecification of densities, the generalized linear model framework allows for estimation of τ and approximately correct inference as long as the the mean is modeled correctly and the mean-variance relationship is specified correctly. Details of this robustified inference may be found in McCullagh (1983) 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:

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

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

Liang and Zeger (1986) actually proposed a generalized set of estimating equations that accommodates independent clusters of correlated data. The form of the estimating equations and $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices are similar to the above except that the sums are over independent clusters. Dunlop (1994) gives 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 $\boldsymbol{A}$ matrix estimates from the independent case are still consistent, but more complicated methods must be used in estimating the $\boldsymbol{B}$ matrix; see Lumley and Heagerty (2000) and Kim and Boos (2001).

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

Game Number	1	2	3	4	5	6	7	8	9	10	11
FT's Made	4	5	5	5	2	7	6	9	4	1	13
FT's Attempted	5	11	14	12	7	10	14	15	12	4	27
Prop. Made	.80	.45	.36	.42	.29	.70	.43	.60	.33	.25	.48

Game Number	12	13	14	15	16	17	18	19	20	21	22	23
FT's Made	5	6	9	7	3	8	1	18	3	10	1	3
FT's Attempted	17	12	9	12	10	12	6	39	13	17	6	12
Prop. Made	.29	.50	1.0	.58	.30	.67	.17	.46	.23	.59	.17	.25

6. APPLICATION TO TEST STATISTICS

Recall that Wald test statistics for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ are quadratic forms like $\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T V_n(\widehat{\boldsymbol{\theta}})^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ 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 two applications .

Example 12 In the National Basketball Association (NBA) playoffs of 2000, Los Angeles Lakers star player Shaquille O'Neal played in 23 games. Table 1 gives his game-by-game free throw outcomes and Figure 1 displays the results.

It is often conjectured that players have streaks where they shoot better or worse. One way to think about that is to assume that the 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 might correspond to some games having high or low p_i values. Thus, a statistical formulation of the problem might be **"Can the above observed game-to-game variation in sample proportions be explained by binomial variability with a common p?"** (By the way, the apparent downward trend in sample proportions is not significant; the simple linear regression p-value=.24.)

For generality let k be the number of games. The score statistic for testing a common binomial

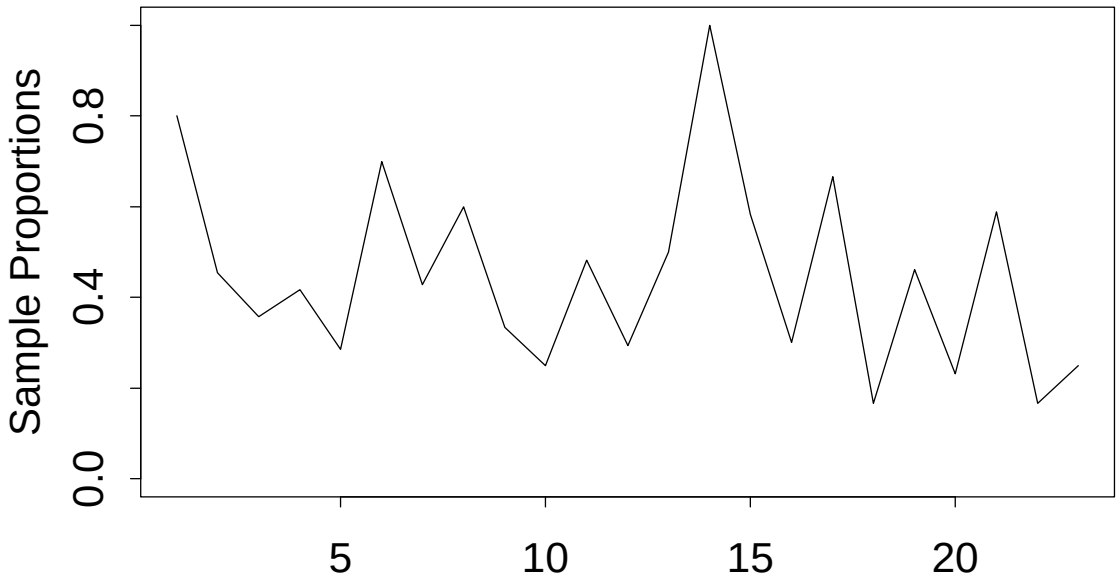

Figure 1: Shaq's Free Throw Percentages in the 2000 NBA Playoffs

proportion versus some differences

$$H_0 : p_1 = p_2 = \cdots = p_k = p \quad \text{vs.} \quad H_1 : p_i \neq p_j \quad \text{for at least one pair } i \neq j$$

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 fixed for this derivation (they aren't really; so this will be a conditional approach). T_S is also the simple chisquared 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 chisquared distribution with $k - 1 = 22$ degrees of freedom. But the chisquared approximation is based on each n_i going to infinity, and most of the n_i in our data set are quite small. Another approach then is to use 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 T_S/k as a parameter θ_1, and p

as θ_2, and form two ψ functions:

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

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

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

$\boldsymbol{B}_{12} = (1-2p), \boldsymbol{B}_{22} = \mu_n p(1-p)$. 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

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

Plugging in $\sum(1/n_i)$ for $\mathrm{E}(1/n_i)$ and $k^{-1}\sum n_i$ for μ_n and comparing T_S/k to a normal distribution with mean=1 and this estimated variance divided by k leads to 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 chisquared approximation seems better than the normal approximation. We might add that the results are very sensitive to game 14 where Shaq made 9 free throws out of 9. 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.

Example 13 Sen (1982) first derived generalized score tests based on the M-estimation formulation. Boos (1992) shows how the form of the test statistic arises from Taylor expansion. In this example we would like to point out how our M-estimation approach leads to the correct test statistic. In a sense, the $\boldsymbol{A}$ and $\boldsymbol{B}$ matrix formulation automatically does the Taylor expansion and computes the variance of the appropriate linear approximations. For simplicity we will present

results for the iid situation; regression extensions are similar.

Assume that $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$ and the null hypothesis is $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$, where $\boldsymbol{\theta}_1$ is of dimension $r \times 1$ and $\boldsymbol{\theta}_2$ is of dimension $(p - r) \times 1$. Assume that $\boldsymbol{\psi}^T = (\boldsymbol{\psi}_1^T, \boldsymbol{\psi}_2^T)$ is partioned similarly. Generalized score tests are based on $\sum \boldsymbol{\psi}_1(Y_i, \widetilde{\boldsymbol{\theta}})$, where $\widetilde{\boldsymbol{\theta}}^T = (\boldsymbol{\theta}_{10}^T, \widetilde{\boldsymbol{\theta}}_2^T)$ satisfies $\sum \boldsymbol{\psi}_2(Y_i, \widetilde{\boldsymbol{\theta}}) = \mathbf{0}$. The goal is to find the appropriate variance matrix to invert and put between $\sum \boldsymbol{\psi}_1(Y_i, \widetilde{\boldsymbol{\theta}})^T$ and $\sum \boldsymbol{\psi}_1(Y_i, \widetilde{\boldsymbol{\theta}})$.

To that end, let $\widehat{\boldsymbol{\theta}}_1^* = n^{-1} \sum_{i=1}^n \boldsymbol{\psi}_1(Y_i, \widetilde{\boldsymbol{\theta}})$ be an M-estimator that solves $\sum \left[\boldsymbol{\psi}_1(Y_i, \widetilde{\boldsymbol{\theta}}) - \widehat{\boldsymbol{\theta}}_1^* \right] = 0$. Then, thinking of $\boldsymbol{\theta}_1^*$ as a parameter that is the limit in probability of $\widehat{\boldsymbol{\theta}}_1^*$, the parameter for this new problem is $\boldsymbol{\theta}^*$ composed of $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2$; $\boldsymbol{\theta}_{10}$ is fixed and not a parameter in the new problem. The associated $\boldsymbol{\psi}$ functions are $\boldsymbol{\psi}_1^*(Y_i, \boldsymbol{\theta}^*) = \boldsymbol{\psi}_1(Y_i, \boldsymbol{\theta}) - \boldsymbol{\theta}_1^*$ and $\boldsymbol{\psi}_2^*(Y_i, \boldsymbol{\theta}^*) = \boldsymbol{\psi}_2(Y_i, \boldsymbol{\theta})$. Taking derivatives with respect to $\boldsymbol{\theta}^*$ and expectations, we find that

$$\boldsymbol{A}^* = \begin{pmatrix} \boldsymbol{I}_r & \boldsymbol{A}_{12} \\ \mathbf{0} & \boldsymbol{A}_{22} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{B}^* = \boldsymbol{B} = \begin{pmatrix} \mathrm{E}\boldsymbol{\psi}_1\boldsymbol{\psi}_1^T & \mathrm{E}\boldsymbol{\psi}_1\boldsymbol{\psi}_2^T \\ \mathrm{E}\boldsymbol{\psi}_2\boldsymbol{\psi}_1^T & \mathrm{E}\boldsymbol{\psi}_2\boldsymbol{\psi}_2^T \end{pmatrix},$$

where $\boldsymbol{I}_r$ is the $r \times r$ identity matrix and $\boldsymbol{A}$ and $\boldsymbol{B}$ without *'s refer to their form in the original problem. Finally, inverting and multiplying leads to the asymptotic variance of $\sum \boldsymbol{\psi}_1(Y_i, \widetilde{\boldsymbol{\theta}})/\sqrt{n} = \sqrt{n}\widehat{\boldsymbol{\theta}}_1^*$ given by

$$\boldsymbol{V}_{11} = \boldsymbol{B}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{B}_{21} - \boldsymbol{B}_{12}\{\boldsymbol{A}_{22}^{-1}\}^T\boldsymbol{A}_{12}^T + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{B}_{22}\{\boldsymbol{A}_{22}^{-1}\}^T\boldsymbol{A}_{12},$$

and the generalized score statistic is

$$T_{GS} = \sum \boldsymbol{\psi}_1(Y_i, \widetilde{\boldsymbol{\theta}})^T \boldsymbol{V}_{11}^{-1}(\widetilde{\boldsymbol{\theta}}) \sum \boldsymbol{\psi}_1(Y_i, \widetilde{\boldsymbol{\theta}}).$$

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 summarized makes standard error estimation and asymptotic analysis routine regardless of the complexity or dimension of the problem. In summary we would like to bring together the key features of M-estimators:

1. An M-estimator $\widehat{\boldsymbol{\theta}}$ satisfies (1): $\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{Y}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}$, 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. $\boldsymbol{A}(\boldsymbol{\theta}_0) = \mathrm{E}\left[-\partial\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}^T\right]$ is the Fisher information matrix in regular parametric models when $\boldsymbol{\psi}$ is the log-likelihood score function. More generally $\boldsymbol{A}(\boldsymbol{\theta}_0)$ must have an inverse but need not be symmetric. See also the extension (11) for non-differentiable $\boldsymbol{\psi}$.

4. $\boldsymbol{B}(\boldsymbol{\theta}_0) = \mathrm{E}\left[\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)\boldsymbol{\psi}(Y_1, \boldsymbol{\theta}_0)^T\right]$ is also the Fisher information matrix in regular parametric models when $\boldsymbol{\psi}$ is the log-likelihood score function. $\boldsymbol{B}(\boldsymbol{\theta}_0)$ always has the properties of a covariance matrix but will be singular when one component of $\widehat{\boldsymbol{\theta}}$ is a non-random function of the other components of $\widehat{\boldsymbol{\theta}}$.

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

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

REFERENCES

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

Boos, D. D. (1992), "On Generalized Score Tests," *The American Statistician*, 46, 327-333.

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

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*, John Wiley & Sons: New York.

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 It's 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.

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

Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *Annals of Statistics*, 1, 799-821.

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.

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

Sen, P. K. (1982), "On M Tests in Linear Models," *Biometrika*, 69, 245-248.

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