A statistic $\hat{\theta}$ is an M-estimator if it solves a system of equations of the form

$$\underset{p\times 1}{\underset{\sim}{0}} = \sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta}) \qquad (\theta \in \mathbb{R}^p)$$

$\{Y_1, \cdots, Y_n\}$ is the sample.

where φ is a known function that does not depend on the data.

Data: $Y_1, \cdots, Y_n$ independent (scalar or vector), not necessarily iid

Eg: the sample mean. $Y_1, \cdots, Y_n \overset{iid}{\sim} f_Y(y|\theta)$.

$\varphi(y, \theta) = y - \theta$

solve $0 = \sum_{i=1}^{n} \varphi(Y_i, \theta) = \sum_{i=1}^{n} (Y_i - \theta) = n(\bar{Y}_n - \theta) \Rightarrow \hat{\theta} = \bar{Y}_n$

so $\hat{\theta} = \bar{Y}_n$ is an M-estimator

Eg.
$$\varphi(y, \underset{\sim}{\theta}) = \begin{pmatrix} y - \theta_1 \\ \theta_2 - (y - \theta_1)^2 \end{pmatrix} \qquad \underset{\sim}{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

$$\underset{\sim}{0} = \sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta}) = \sum_{i=1}^{n} \begin{pmatrix} Y_i - \theta_1 \\ \theta_2 - (Y_i - \theta_1)^2 \end{pmatrix} = \begin{pmatrix} n(\bar{Y}_n - \theta_1) \\ n\theta_2 - \sum_{i=1}^{n}(Y_i - \theta_1)^2 \end{pmatrix}$$

$$\Rightarrow \quad \hat{\theta}_1 = \bar{Y}_n$$
$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\theta}_1)^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$$

However. $\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$ is not an M-estimator by itself.

Because there is no function φ such that

$0 = \sum_{i=1}^{n} \varphi(Y_i, \hat{\theta}_2)$ holds

In case like this. we say $\hat{\theta}_2$ is a partial M-estimator.

and to be precise, $\hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}$ is a full M-estimator.

Eg. Non-linear least square. $Y_i = g(X_i, \underset{\sim}{\theta}) + \varepsilon_i$

where $E(\varepsilon_i) = 0$ independent $Var(\varepsilon_i) = \sigma^2$

g known differentiable of $\underset{\sim}{\theta} : \mathbb{R}^p \to \mathbb{R}$

Estimator $\hat{\theta} = \operatorname{argmin} \sum_{i=1}^{n} \{Y_i - g(X_i, \underset{\sim}{\theta})\}^2$

as long as the minimizer is not on the boundary of the parameter space

we have

$$\underset{\sim}{0} = \sum_{i=1}^{n} \{Y_i - g(X_i, \hat{\underset{\sim}{\theta}})\} \dot{g}(X_i, \hat{\theta})$$

where $\dot{g}(X, \underset{\sim}{\theta}) = \dfrac{\partial g(X, \underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}}$ $p \times 1$

Then Let $\varphi = \{y - g(x, \theta)\} \cdot \dot{g}(x, \theta) \Rightarrow \hat{\theta}$ is an M-estimator.

If we are also interested in σ^2

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - g(X_i, \hat{\underset{\sim}{\theta}})\}^2$$

↑

partial M-estimator

where if $\varphi(y, x, \underset{\sim}{\theta}, \sigma^2) = \begin{pmatrix} (y - g(x, \theta)) \cdot \dot{g}(x, \theta) \\ \sigma^2 - (y - g(x, \theta))^2 \end{pmatrix}_{(p+1) \times 1}$

Then $(\hat{\underset{\sim}{\theta}}, \hat{\sigma}^2)$ is a solution to $\underset{\sim}{0} = \sum_{i=1}^{n} \varphi(Y_i, X_i, \underset{\sim}{\theta}, \sigma^2)$

↳ a full M-estimator

Eg. Consistent roots to the likelihood equation are M-estimators

$$L(\underset{\sim}{\theta}) = \prod_{i=1}^{n} f_Y(Y_i | \underset{\sim}{\theta})$$

$$l(\underset{\sim}{\theta}) = \log L(\underset{\sim}{\theta}) = \sum_{i=1}^{n} \log f_Y(Y_i | \underset{\sim}{\theta})$$

$\hat{\theta}$ solves $\displaystyle\sum_{i=1}^{n} \frac{\partial \log f_Y(Y_i | \underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = 0$

Taking $\varphi(y, \underset{\sim}{\theta}) = \dfrac{\partial}{\partial \underset{\sim}{\theta}} \log f_Y(y | \underset{\sim}{\theta})$

we see $\hat{\theta}$ is an M-estimator.

Eg. MoM estimators are M-estimators

Suppose $Y_1, \dots, Y_n$ iid with

$$E(Y^j) = g_j(\underset{\sim}{\theta}) \quad j=1,\dots,p. \quad \dim(\underset{\sim}{\theta}) = p$$

MOM estimator $\hat{\theta}$ solves

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} Y_i = g_1(\underset{\sim}{\theta}) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} Y_i^p = g_p(\underset{\sim}{\theta}) \end{cases}$$

Taking $\varphi(y, \underset{\sim}{\theta}) = \begin{pmatrix} y - g_1(\underset{\sim}{\theta}) \\ \vdots \\ y^p - g_p(\underset{\sim}{\theta}) \end{pmatrix} \Rightarrow \hat{\theta}$ is an M-estimator.

Eg. Functions of M-estimators (or partial M-estimator) are partial M-estimators

Eg: $\hat{\theta}_3 = \frac{(\bar{Y}_n)^2}{\bar{X}_n}$ $\quad \begin{pmatrix} Y_1, \dots, Y_n \\ X_1, \dots, X_n \end{pmatrix}$

then $\varphi(y, x, \theta_1, \theta_2, \theta_3) = \begin{pmatrix} y - \theta_1 \\ x - \theta_2 \\ \theta_1^2 - \theta_2\theta_2 \end{pmatrix}$

$$\sum_{i=1}^{n} \varphi(y_i, x_i, \theta_1, \theta_2, \theta_3) = 0$$
$$\Downarrow$$
$$\hat{\theta}_1 = \bar{Y}_n, \ \hat{\theta}_2 = \bar{X}_n$$
$$\hat{\theta}_3 = \frac{(\bar{Y}_n)^2}{\bar{X}_n}$$

or $\varphi(y, x, \theta_1, \theta_2, \theta_3) = \begin{pmatrix} y - \theta_1 \\ x - \theta_2 \\ \theta_1 y - \theta_2\theta_3 \end{pmatrix}$

$$= \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \theta_1 & 0 & 1 \end{pmatrix}} \cdot \begin{pmatrix} y - \theta_1 \\ x - \theta_2 \\ \theta_1^2 - \theta_2\theta_3 \end{pmatrix}$$

non-uniqueness of φ.

The equation $\sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta}) = \underset{\sim}{0}$ as long as M^{-1} exists

will have the same roots to $M \sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta}) = 0$

or $\sum_{i=1}^{n} M\varphi(Y_i, \underset{\sim}{\theta}) = 0$ take $\varphi^* = M\varphi$. we say φ^* and φ are equivalent!

WHY? are M-estimator of interest?

1. Many estimators are M-estimators
2. M-estimators are often consistent & asymptotically normal
   (a) conditions on φ (smoothness, etc..)
   (b) Moment assumption $E(\varphi^2) < \infty$
3. Obtaining asymptotic distribution of M-estimators is almost automatic & the covariance matrix is easily calculated.
4. It is easy to study how the estimate depends on the data (Sensitivity, robustness analysis)

Properties: $Y_1, \cdots, Y_n \overset{iid}{\sim} F$

$\hat{\underset{\sim}{\theta}}$ is an M-estimator satisfying $\sum_{i=1}^{n} \varphi(Y_i, \hat{\underset{\sim}{\theta}}) = 0$ for a specific φ.

(or if $\hat{\theta}$ minimizes $\sum_{i=1}^{n} \rho(Y_i, \underset{\sim}{\theta})$ for a specific ρ)

under two different formulations. different sets of conditions on φ (or ρ) ensure consistency and asymptotic normality of $\hat{\underset{\sim}{\theta}}$

The variety of conditions can be confusing so we will deal essentially with one clean set of conditions

Huber (1964). Serfling (1980). vander Vaarts (1998)

The difficulties with the φ approach are that:

(a) the asymptotic behavior of a root of $\frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta})$ depends on the global behavior of

$$\lambda_F(\underset{\sim}{\theta}) = E(\varphi(Y, \underset{\sim}{\theta})) = \int \varphi(Y, \underset{\sim}{\theta})\, dF(Y).$$

whether $\int \varphi(Y, \underset{\sim}{\theta})\, dF(Y) = 0$ has a unique root.

(b) The equation $\int \varphi(Y, \underset{\sim}{\theta}) dF(Y) = 0$ may not have any exact roots.

(c) There can be multiple roots in which case a rule is required to select one.

if $\varphi(Y, \overset{\to \text{scalar}}{\theta})$ is continuous and strictly monotone in θ and if $\int \varphi(Y, \theta) dF(Y) = 0$ has a unique root θ_0 the $\frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i, \theta) = 0$ will have a unique root and the M-estimator is unique defined and consistent.

Monotonicity and continuity of φ are frequently assumed.

Thm. Assume that

$\hat{\underset{\sim}{\theta}}$ solves $\frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta}) = 0$

$\downarrow$

$E[\varphi(Y, \underset{\sim}{\theta})] = 0$

(i) $E\{\psi(Y, \theta)\} = 0$ has a unique root θ_0

(ii) ψ is continuous and either bounded or monotone.

Then $\sum_{i=1}^{n} \varphi(Y_i, \theta) = 0$ admits a sequence of roots $\hat{\theta}_n$ s.t. $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$

see Serfling (1980) for a proof

Alternatively, one can entirely give up the idea of identifying consistent roots of $\sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta}) = 0$

For example: Starting with an initial $\sqrt{n}$-consistent estimator $\hat{\theta}_n$ one can look at the one-step Newton-Raphson estimator.

$$\underset{\sim}{\delta}_n = \hat{\underset{\sim}{\theta}}_n - \left[\sum_{i=1}^{n} \dot{\varphi}(Y_i, \hat{\underset{\sim}{\theta}}_n)\right]^{-1} \sum_{i=1}^{n} \varphi(Y_i, \hat{\underset{\sim}{\theta}}_n)$$

Consistency (even asymptotic normality) of δ_n is automatic.

but $\underset{\sim}{\delta}_n$ is not a root of $\sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta}) = 0$

We now establish the asymptotic distribution of $\hat{\underset{\sim}{\theta}}$ for iid case.

Assumptions:

1. There exists a $\underset{\sim}{\theta}_0$ (truth) such that

$$E_{\underset{\sim}{\theta}_0}\{\varphi(Y, \underset{\sim}{\theta}_0)\} = \underset{\sim}{0}$$

2. $\hat{\underset{\sim}{\theta}} \xrightarrow{P} \underset{\sim}{\theta}_0$

3. φ has a continuous derivative in $\underset{\sim}{\theta}$ for all y

Define $\quad G_n(\underset{\sim}{\theta}) = \sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta}) \; : \mathbb{R}^p \to \mathbb{R}^p$ and $G_n(\hat{\underset{\sim}{\theta}}) = 0$

Causal Argument:

Using a Taylor Series Approximation. $\quad \left.\dfrac{\partial G_n(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}^T}\right|_{\underset{\sim}{\theta} = \tilde{\underset{\sim}{\theta}}_n}$

$$\underset{\sim}{0} = G_n(\hat{\theta}) \approx G_n(\underset{\sim}{\theta}_0) + \dot{G}_n(\theta_0)(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0)$$

$$\Rightarrow \quad \hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0 = \{-\dot{G}_n(\theta_0)\}^{-1} G_n(\underset{\sim}{\theta}_0)$$

we assume $\dot{G}_n(\theta_0)$ is invertible at least when n is large.

$$\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \theta_0) = \left\{-\frac{1}{n}\dot{G}_n(\theta_0)\right\}^{-1} \sqrt{n}\left(\frac{1}{n} G_n(\underset{\sim}{\theta}_0)\right)$$

$$-\frac{1}{n}\dot{G}_n(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n} \dot{\varphi}(Y_i, \theta_0)$$

$$\xrightarrow{P} E[-\dot{\varphi}(Y_i, \theta_0)] \doteq A(\theta_0)_{p\times p}$$

provided existence

Now. $$\sqrt{n}\left(\frac{1}{n}G_n(\theta_0)\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(Y_i, \underset{\sim}{\theta}_0) - \underset{E[\varphi(Y_i, \theta_0)]}{\underset{\uparrow}{0}}\right)$$

based on CLT:

$$\sqrt{n}\left(\frac{1}{n}G_n(\theta_0)\right) \xrightarrow{d} N(0, Var(\underline{\varphi(Y, \theta)}))$$

$$\parallel$$

$$E[\varphi(Y, \theta_0)\, \varphi(Y, \theta_0)^T] = B(\theta_0)$$

provided exists

$$\Rightarrow \sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \xrightarrow{d} N(0, V(\underset{\sim}{\theta}_0))$$

where $V(\underset{\sim}{\theta}_0) = A(\underset{\sim}{\theta}_0)^{-1} B(\underset{\sim}{\theta}_0) [A(\theta_0)^{-1}]^T$

Sandwich Matrix

Meat

Bread

Note: if the model is correctly specified for Y

then for the MLE $\hat{\theta}$ — Information Matrix

$$A(\underset{\sim}{\theta}_0) = B(\underset{\sim}{\theta}_0) = I(\underset{\sim}{\theta}_0)$$

$$V(\underset{\sim}{\theta}_0) = I^{-1}(\underset{\sim}{\theta}_0)$$

otherwise, inference should be carried out using $V(\theta_0)$

not $I^{-1}(\theta_0)$

Estimating $V(\underset{\sim}{\theta}_0)$

Define $A_n(\underset{\sim}{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \{-\dot{\varphi}(Y_i, \underset{\sim}{\theta})\}$

$B_n(\underset{\sim}{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i, \underset{\sim}{\theta}) \varphi(Y_i, \underset{\sim}{\theta})^T$

Then $\hat{\theta}_n \xrightarrow{P} \theta_0 \Rightarrow A_n(\hat{\theta}_n) \xrightarrow{P} A(\theta_0)$

$B_n(\hat{\theta}_n) \xrightarrow{P} B(\theta_n)$

$$V_n(\hat{\theta}_n) = A_n(\hat{\theta}_n)^{-1} B_n(\hat{\theta}_n) \{A_n(\hat{\theta}_n)^{-1}\}^T \xrightarrow{P} V(\theta_0)$$

Eg. Ratio estimator. (X_i, Y_i) iid

$E(X) = \mu_X$. $E(Y) = \mu_Y$ $\quad Cor(X, Y) = \sigma_{XY}$

$V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$

$$\theta = \frac{\mu_Y}{\mu_X} \qquad \hat{\theta} = \frac{\overline{Y}_n}{\overline{X}_n}$$

$\hat{\theta}$ is an M-estimator. $\varphi(y, x, \theta) = y - \theta x$ $\quad \dot{\varphi}(y, x, \theta) = -x$

$$\sum_{i=1}^{n} \varphi(Y_i, X_i, \theta) = \sum_{i=1}^{n} Y_i - \theta \sum_{i=1}^{n} X_i$$

$$A(\theta_0) = E\left[-\dot{\varphi}(Y_i, X_i, \theta_0)\right] = E[X_i] = \mu_x$$

$$\begin{aligned}
B(\theta_0) &= E[\varphi(Y_i, X_i, \theta_0)\varphi(Y_i, X_i, \theta_0)^T] \\
&= E[\varphi^2(Y_i, X_i, \theta_0)] = E(Y_i - \theta_0 X_i)^2 \\
&= E(Y_i^2) + \theta_0^2 E(X_i^2) - 2\theta_0 E(X_i Y_i) \\
&= \mu_Y^2 + \sigma_Y^2 + \theta_0^2(\mu_x^2 + \sigma_x^2) - 2\theta_0(\sigma_{xY} + \mu_x \mu_Y) \\
&= \underbrace{(\mu_Y - \theta_0 \mu_x)^2}_{0} + \sigma_Y^2 + \frac{\mu_Y^2}{\mu_x^2}\sigma_x^2 - 2\frac{\mu_Y}{\mu_x}\sigma_{xY}
\end{aligned}$$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \quad A(\theta_0)^{-1} B(\theta_0) (A_0(\theta)^{-1})^T\right)$$

$$\| $$

$$\frac{\sigma_Y^2 - 2\theta_0 \sigma_{xY} + \theta_0^2 \sigma_x^2}{\mu_x^2}$$

Delta Method:

$$\sqrt{n}\left\{\begin{pmatrix} \bar{Y}_n \\ \bar{X}_n \end{pmatrix} - \begin{pmatrix} \mu_Y \\ \mu_x \end{pmatrix}\right\} \xrightarrow{d} N\left(0, \Sigma = \begin{pmatrix} \sigma_Y^2 & \sigma_{xY} \\ \sigma_{xY} & \sigma_x^2 \end{pmatrix}\right)$$

$$g(y, x) = \frac{y}{x}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \dot{g}^T \Sigma \dot{g})$$
