M-Estimators

1.1 Definition

Definition 1.1.1. For any function $\boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta})$, define the *functional* $T(F) = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is a solution to the equation:

$$\boldsymbol{\Psi}_F(\boldsymbol{\theta}_0) \equiv \int \boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta}_0)dF(\boldsymbol{y}) = 0. \tag{1.1.1}$$

Suppose $(\boldsymbol{y}_i)_{i=1}^n$ is a random sample from F with empirical distribution function $\mathbb{F}_n$. The **M-estimator** corresponding to $\boldsymbol{\psi}$ is $\hat{\boldsymbol{\theta}}_n = T(\mathbb{F}_n)$, which is a solution to the equation:

$$\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n) \equiv \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}}_n) = \mathbf{0}.$$

■

Discussion 1.1.1. Often, $\boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta})$ arises as the gradient of an objective function $q(\boldsymbol{y};\boldsymbol{\theta})$:

$$\boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta}_0) = \frac{\partial}{\partial\boldsymbol{\theta}_0}q(\boldsymbol{y};\boldsymbol{\theta}_0),$$

and $\boldsymbol{\theta}_0$ solving (1.1.1) is a solution to the minimization problem:

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int q(\boldsymbol{y};\boldsymbol{\theta})dF(\boldsymbol{y}).$$

For example, suppose F is a parametric distribution $F(\boldsymbol{y};\boldsymbol{\theta}_0)$ and the goal is to estimate $\boldsymbol{\theta}_0$. Consider specifying the negative log likelihood $-\ln f(\boldsymbol{y};\boldsymbol{\theta})$ as the objective $q(\boldsymbol{y};\boldsymbol{\theta})$. The gradient $\boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta})$ of $q(\boldsymbol{y};\boldsymbol{\theta})$ is the parametric score equation:

$$\boldsymbol{\Psi}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\boldsymbol{\theta}}\ln f(\boldsymbol{y}_i;\boldsymbol{\theta}).$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$, which satisfies $\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$, is therefore an example of an M-estimator. See section (1.3) for additional examples. ♠

1.2 Asymptotics

Theorem 1.1.1 (Consistency). Suppose $(\boldsymbol{y}_i)_{i=1}^n$ is a random sample from F, and that $\boldsymbol{\theta}$ belongs to a compact parameter space $\boldsymbol{\Theta}$. Assume that:

i. $\boldsymbol{\Psi}_F(\boldsymbol{\theta})$ exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and that $\boldsymbol{\theta}_0$ is the unique zero of $\boldsymbol{\Psi}_F(\boldsymbol{\theta})$.

ii. Each component of $\boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta})$ is continuous and $\boldsymbol{\theta}$ and bounded by an integrable function of $\boldsymbol{y}$, not depending on $\boldsymbol{\theta}$.

If $\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{as} \mathbf{0}$, then $\hat{\boldsymbol{\theta}}_n \xrightarrow{as} \boldsymbol{\theta}_0$. □

Remark 1.1.1. See Boos and Stefanski (2013), theorem 7.1. ♦

Theorem 1.1.2 (Asymptotic Normality). Suppose $(\boldsymbol{y}_i)_{i=1}^n$ is a random sample from F, and $\boldsymbol{\theta}$ belongs to a compact parameter space $\boldsymbol{\Theta}$. Assume that:

i. $\boldsymbol{\Psi}_F(\boldsymbol{\theta})$ exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and that $\boldsymbol{\theta}_0$ is the unique zero of $\boldsymbol{\Psi}_F(\boldsymbol{\theta})$.

ii. $\boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta})$ is continuous and twice differentiable with respect to $\boldsymbol{\theta}$ for all $\boldsymbol{y}$ in the support of F and $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$.

iii. For $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, there exists an integrable function $g(\boldsymbol{y})$ such that:

$$\left|\frac{\partial^2}{\partial\theta_{j_1}\partial\theta_{j_2}}\psi_{j_3}(\boldsymbol{y};\boldsymbol{\theta})\right| \leq g(\boldsymbol{y})$$

for $\forall(j_1, j_2, j_3) \in \{1, \cdots, \dim(\boldsymbol{\theta})\}^3$, and $\int g(y)dF(y)$.

iv. $\boldsymbol{A}(\boldsymbol{\theta}_0)$ exists and is non-singular, where:

$$\boldsymbol{A}(\boldsymbol{\theta}_0) = -E_F\{\dot{\boldsymbol{\psi}}(\boldsymbol{y};\boldsymbol{\theta}_0)\} = -\int \dot{\boldsymbol{\psi}}(\boldsymbol{y};\boldsymbol{\theta}_0)dF(\boldsymbol{y}).$$

v. $\boldsymbol{B}(\boldsymbol{\theta}_0)$ exists and is finite, where:

$$\boldsymbol{B}(\boldsymbol{\theta}_0) = E_F\{\boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta}_0) \otimes \boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta}_0)\} = \int \boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta}_0) \otimes \boldsymbol{\psi}(\boldsymbol{y};\boldsymbol{\theta}_0)dF(\boldsymbol{y}).$$

If $\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n) = o_p(n^{-1/2})$, then:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, \boldsymbol{\Omega}(\boldsymbol{\theta}_0)\},$$

$\boldsymbol{\Omega}(\boldsymbol{\theta}_0) = \boldsymbol{A}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{B}(\boldsymbol{\theta}_0)\boldsymbol{A}^{-T}(\boldsymbol{\theta}_0)$. □

Remark 1.1.2. See Boos and Stefanski (2013), theorem 7.2. ♦

Proof. By Taylor expansion:

$$o_p(n^{-1/2}) = \boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) + \dot{\boldsymbol{\Psi}}_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \boldsymbol{R}_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0),$$

where the remainder:

$$\boldsymbol{R}_n = \sum_{j=1}^{\dim(\boldsymbol{\theta})} \frac{\partial \dot{\boldsymbol{\Psi}}_n}{\partial \theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*}(\hat{\theta}_{n,j} - \theta_{0,j})$$

and $||\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0|| \leq ||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||$. By condition (iii.), the matrix of partial derivatives appearing in the remainder is bounded by a function not depending on $\boldsymbol{\theta}$, hence:

$$||\boldsymbol{R}_n|| = \mathcal{O}_p\big(||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0||\big).$$

Since the conditions for consistency are contained within the conditions for asymptotic normality, $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = o_p(1)$, which implies $\boldsymbol{R}_n = o_p(1)$. Now:

$$o_p(n^{-1/2}) = \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) + \big\{\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\theta}_0) + o_p(1)\big\}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0),$$
$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \big\{-\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\theta}_0) + o_p(1)\big\}^{-1}\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0).$$

By the LLN:

$$-\dot{\boldsymbol{\Psi}}_n(\boldsymbol{\theta}_0) + o_p(1) \xrightarrow{p} -E\big\{\dot{\boldsymbol{\psi}}(\boldsymbol{y};\boldsymbol{\theta}_0)\big\} = \boldsymbol{A}(\boldsymbol{\theta}_0)$$

By continuous mapping:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \sqrt{n}\boldsymbol{A}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) + o_p(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big\{\boldsymbol{A}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\psi}(\boldsymbol{y}_i;\boldsymbol{\theta}_0)\big\} + o_p(1).$$

Identify $\boldsymbol{\varphi}(\boldsymbol{y}_i;\boldsymbol{\theta}_0) = \boldsymbol{A}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\psi}(\boldsymbol{y}_i;\boldsymbol{\theta}_0)$ as the *influence function* of $\hat{\boldsymbol{\theta}}_n$. By the CLT,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\big\{\mathbf{0}, \boldsymbol{\Omega}(\boldsymbol{\theta}_0)\big\},$$

where:

$$\boldsymbol{\Omega}(\boldsymbol{\theta}_0) = \mathrm{Var}\big\{\boldsymbol{\varphi}(\boldsymbol{y}_i;\boldsymbol{\theta}_0)\big\} = \boldsymbol{A}^{-1}(\boldsymbol{\theta}_0)E\big\{\boldsymbol{\psi}(\boldsymbol{y}_i;\boldsymbol{\theta}_0) \otimes \boldsymbol{\psi}(\boldsymbol{y}_i;\boldsymbol{\theta}_0)\big\}\boldsymbol{A}^{-T}(\boldsymbol{\theta}_0).$$

$\blacksquare$

1.3 Examples

Example 1.1.1 (Non-linear Least Squares). Consider the model:

$$Y_i = g(\boldsymbol{x}_i, \boldsymbol{\beta}) + \epsilon_i,$$

where g is a known differentiable function, and $(\epsilon_i)_{i=1}^n$ are independent, random residuals with $E(\epsilon_i|\boldsymbol{x}_i) = 0$ and $\text{Var}(\epsilon_i|\boldsymbol{x}_i) = \sigma_i^2$. Define the squared error objective function:

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\big\{Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})\big\}^2.$$

Differentiating to obtain the estimating equation:

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\big\{Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})\big\}\dot{g}(\boldsymbol{x}_i, \boldsymbol{\beta}).$$

Identify $\boldsymbol{\psi}(Y_i, \boldsymbol{x}_i; \boldsymbol{\beta}) = \big\{Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})\big\}\dot{g}(\boldsymbol{x}_i, \boldsymbol{\beta})$. Let $\hat{\boldsymbol{\beta}}$ denote a solution to $\boldsymbol{\Psi}_n(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$, then under the assumptions of theorem (1.1.2), $\hat{\boldsymbol{\beta}}$ is an M-estimator.

$\boldsymbol{A}(\boldsymbol{\beta}_0)$ takes the form:

$$\boldsymbol{A}(\boldsymbol{\beta}_0) = -E_0\big\{\dot{\boldsymbol{\psi}}(Y_i, \boldsymbol{x}_i; \boldsymbol{\beta}_0)\big\} = E_0\big\{\dot{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) \otimes \dot{g}(\boldsymbol{x}_i, \boldsymbol{\beta})\big\}.$$

The empirical estimate of $\boldsymbol{A}(\boldsymbol{\beta}_0)$ is:

$$\hat{\boldsymbol{A}} = \frac{1}{n}\sum_{i=1}^{n}\dot{g}(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}) \otimes \dot{g}(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}).$$

$\boldsymbol{B}(\boldsymbol{\beta}_0)$ takes the form:

$$\boldsymbol{B}(\boldsymbol{\beta}_0) = E_0\Big[\big\{Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})\big\}^2\dot{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) \otimes \dot{g}(\boldsymbol{x}_i, \boldsymbol{\beta})\Big] = \sigma_i^2\dot{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) \otimes \dot{g}(\boldsymbol{x}_i, \boldsymbol{\beta}).$$

The empirical estimate of $\boldsymbol{B}(\boldsymbol{\beta}_0)$ is:

$$\hat{\boldsymbol{B}} = \frac{1}{n}\sum_{i=1}^{n}\big\{Y_i - g(\boldsymbol{x}_i, \boldsymbol{\beta})\big\}^2\dot{g}(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}) \otimes \dot{g}(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}).$$

The asymptotic approximation to the distribution of $\hat{\boldsymbol{\beta}}$ is:

$$\hat{\boldsymbol{\beta}} \mathrel{\dot\sim} N\big(\boldsymbol{\beta}_0, n^{-1}\hat{\boldsymbol{\Omega}}\big),$$

where $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^{-T}$. ♠

Example 1.1.2 (Robust Regression). Consider the model:

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

where $(\epsilon_i)_{i=1}^n$ are independent, random residuals with $E(\epsilon_i|\boldsymbol{x}_i) = 0$ and $\text{Var}(\epsilon_i|\boldsymbol{x}_i) = \sigma_i^2$. The standard least squares estimating equations are:

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\big)\boldsymbol{x}_i.$$

Consider the robust estimating equations:

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}\psi_\tau\big(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\big)\boldsymbol{x}_i.$$

Where ψ_τ is a bounded loss function, such as Huber's function:

$$\psi_\tau(t) = \begin{cases} -\tau, & t < -\tau, \\ x, & -\tau < t < \tau, \\ \tau, & t > \tau. \end{cases}$$

or Tukey's biweight function:

$$\psi_\tau(t) = \begin{cases} t\left(1 - \frac{t^2}{\tau^2}\right)^2, & |t| < \tau, \\ 0, & |t| > \tau. \end{cases}$$

Let $\hat{\boldsymbol{\beta}}$ denote a solution to $\boldsymbol{\Psi}_n(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$, then under the assumptions of theorem (1.1.2), $\hat{\boldsymbol{\beta}}$ is an M-estimator.

$\boldsymbol{A}(\boldsymbol{\beta}_0)$ takes the form:

$$A(\boldsymbol{\beta}_0) = -E_0\big\{\dot{\psi}_\tau(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i \otimes \boldsymbol{x}_i\big\}.$$

The empirical estimate of $\boldsymbol{A}(\boldsymbol{\beta}_0)$ is:

$$\hat{\boldsymbol{A}} = \frac{1}{n}\sum_{i=1}^{n}\dot{\psi}_\tau(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i \otimes \boldsymbol{x}_i.$$

$\boldsymbol{B}(\boldsymbol{\beta}_0)$ takes the form:

$$\boldsymbol{B}(\boldsymbol{\beta}_0) = E_0\big\{\psi_\tau^2(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i \otimes \boldsymbol{x}_i\big\}.$$

The empirical estimate of $\boldsymbol{B}(\boldsymbol{\beta}_0)$ is:

$$\hat{\boldsymbol{B}} = \frac{1}{n}\sum_{i=1}^{n}\psi_\tau^2(Y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i \otimes \boldsymbol{x}_i.$$

The asymptotic approximation to the distribution of $\hat{\boldsymbol{\beta}}$ is:

$$\hat{\boldsymbol{\beta}} \dot{\sim} N\big(\boldsymbol{\beta}_0, n^{-1}\hat{\boldsymbol{\Omega}}\big),$$

where $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^{-T}$.

♠