Lecture 8: Introduction to Nonlinear Estimation

March 4, 2025

Note: We develop theory heuristically only for the purpose of motivating MLE and non-linear GMM.

Start with Nonlinear Regression Models

- Up until now, all estimators we have studied can be written as "closed form" functions of the data. That is, given the observed data, we have a mathematical rule for obtaining the estimate. For example, the OLS estimator is

$$\hat{\beta}_{\text{OLS}} = \left(\sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{x}_i' y_i \right).$$

- Such estimators do not cover all cases of interest, particularly when we turn to nonlinear models.

Start with Nonlinear Regression Models

- As another example, suppose for $y \geq 0$ we specify an exponential conditional mean model:

$$E(y \mid \mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta}) = \exp\left(\beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K\right).$$

- Without further assumptions, we cannot "linearize" the model by using $\log(y)$ as the dependent variable. (In fact, $\log(y)$ may not even be well defined.)

Start with Nonlinear Regression Models

- Instead, we can directly estimate $\boldsymbol{\beta}$ by nonlinear least squares (NLS):

$$\min_{\mathbf{b}\in\mathbb{R}^R} \sum_{i=1}^{N} \left[y_i - \exp\left(\mathbf{x}_i \mathbf{b}\right)\right]^2 .$$

- As in the case of LAD, we cannot present the solution in closed form. But the estimator minimizes a function that is an average of i.i.d. random functions of $\mathbf{b}$.

- For our purposes, "nonlinear" means any situation where an estimator cannot be obtained in closed form. This requires a new set of tools for asymptotic analysis.

M-Estimator

Consistency of M-estimators

- We first cover a class of estimation problems estimators known as M-estimation. (The "M" refers, for us, to "minimization". Originally, M-estimators we defined as maximization problems.)
- So $\boldsymbol{\theta}$ is a $P \times 1$ vector. The parameter space $\boldsymbol{\Theta}$ is the set of all parameters values that could be the population value.
- As an example, $m(\mathbf{x}, \boldsymbol{\theta}) = \exp(\mathbf{x}\boldsymbol{\theta}) = \exp(\theta_1 + \theta_2 x_2 + \ldots + \theta_K x_K)$ where $\mathbf{x} = (1, x_2, \ldots, x_K)$ contains unity for convenience. The parameter space is probably $\boldsymbol{\Theta} = \mathbb{R}^K$ because it is unlikely we would restrict it ahead of time.

M-Estimator

Assumption NLS.1: For some $\boldsymbol{\theta_0} \in \boldsymbol{\Theta}$,

$$E(y \mid \mathbf{x}) = m\left(\mathbf{x}, \boldsymbol{\theta_0}\right).$$

- Remember, $\boldsymbol{\theta_0}$ is just the $P \times 1$ vector of numbers we are trying to learn about. Sometimes, $\boldsymbol{\theta_0}$ is called the "true value of the parameters."
- For some purposes, it is useful to write the equation in error form:

$$\begin{aligned} y &= m\left(\mathbf{x}, \boldsymbol{\theta_0}\right) + u \\ E(u \mid \mathbf{x}) &= 0, \end{aligned}$$

  where the zero conditional mean holds by construction.
- Generally, we should avoid thinking of situations where u is independent of $\mathbf{x}$, and we should not even think $\mathrm{Var}(u \mid \mathbf{x}) = \mathrm{Var}(u)$.

$$\begin{aligned} \left[y - m(\mathbf{x}, \boldsymbol{\theta})\right]^2 &= \left[m\left(\mathbf{x}, \theta_o\right) + u - m(\mathbf{x}, \boldsymbol{\theta})\right]^2 \\ &= u^2 + 2\left[m\left(\mathbf{x}, \theta_o\right) - m(\mathbf{x}, \boldsymbol{\theta})\right] u + \left[m\left(\mathbf{x}, \theta_o\right) - m(\mathbf{x}, \boldsymbol{\theta})\right]^2 \end{aligned}$$

M-Estimator

Then

$$\begin{aligned} E\left\{[y-m(\mathbf{x},\boldsymbol{\theta})]^2\right\} =& E\left(u^2\right)+E\left\{2\left[m\left(\mathbf{x},\theta_o\right)-m(\mathbf{x},\boldsymbol{\theta})\right]u\right\} \\ &+E\left\{\left[m\left(\mathbf{x},\theta_o\right)-m(\mathbf{x},\boldsymbol{\theta})\right]^2\right\} \\ =& E\left(u^2\right)+E\left\{\left[m\left(\mathbf{x},\theta_o\right)-m(\mathbf{x},\boldsymbol{\theta})\right]^2\right\} \end{aligned}$$

because $E(u \mid \mathbf{x})=0$.

- Now $E\left(u^2\right)$ does not depend on $\boldsymbol{\theta}$ and $E\left\{\left[m\left(\mathbf{x},\theta_o\right)-m(\mathbf{x},\boldsymbol{\theta})\right]^2\right\}$ is smallest when $\boldsymbol{\theta}=\theta_o$.
- So, we have shown that

$$\boldsymbol{\theta_0}=\operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E\left\{[y-m(\mathbf{x},\boldsymbol{\theta})]^2\right\}.$$

- In other words, θ_o solves a population minimization problem.
- The *analogy principle* says to solve the sample analog of the population problem, which leads to

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N}\left[y_i-m(\mathbf{x},\boldsymbol{\theta})\right]^2.$$

Uniform convergence

- The M-estimation principle generalizes this reasoning. We assume that $\boldsymbol{\theta}_{\mathbf{0}} \in \mathbf{\Theta}$ uniquely solves
$$\min_{\boldsymbol{\theta} \in \mathbf{\Theta}} E[q(\mathbf{w}, \boldsymbol{\theta})]$$
where $q : \mathcal{W} \times \mathbf{\Theta} \to \mathbb{R}$ is a real valued function of an observable vector $\mathbf{w}$ and the parameter vector $\boldsymbol{\theta}$.

- An M-estimator of θ_o solves the sample analog,
$$\min_{\boldsymbol{\theta} = \mathbf{\Theta}} N^{-1} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right)$$

Uniform convergence

- By the law of large numbers, for each $\boldsymbol{\theta}$,

$$N^{-1} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right) \xrightarrow{p} E[q(\mathbf{w}, \boldsymbol{\theta})]$$

$\hat{\boldsymbol{\theta}}$ minimizes	θ_o minimizes
(sample average)	(population average)

So $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_{\mathbf{0}}$ (as $N \to \infty$, as always) seems reasonable.

- But pointwise convergence of the sample objective function is not sufficient for consistency. A sufficient condition is *uniform convergence in probability*:

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| N^{-1} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right) - E[q(\mathbf{w}, \boldsymbol{\theta})] \right| \xrightarrow{p} 0$$

- Means that we can bound the distance between $N^{-1} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right)$ and its expected value by something that does not depend on $\boldsymbol{\theta}$.
- In "regular" cases, the pointwise law of large numbers translates into the *uniform law of large numbers*. Sufficient is that $q(\mathbf{w}, \cdot)$ is continuous on $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}$ is closed and bounded (compact), and $|q(\mathbf{w}, \boldsymbol{\theta})| \leq b(\mathbf{w})$ for some function $b(\mathbf{w})$ with $E[b(\mathbf{w})] < \infty$.

Identification condition

- For NLS, we can write the identification as

  Assumption NLS.2: $E\left\{\left[m\left(\mathbf{x},\boldsymbol{\theta}_{\mathbf{0}}\right)-m(\mathbf{x},\boldsymbol{\theta})\right]^{2}\right\}>0$ for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $\boldsymbol{\theta}\neq\boldsymbol{\theta}_{\mathbf{0}}$.

- Assumption NLS.2 plays the role of the rank condition. In the linear case, $m(\mathbf{x},\boldsymbol{\theta})=\mathbf{x}\boldsymbol{\theta}$, and then

$$m\left(\mathbf{x},\boldsymbol{\theta}_{\mathbf{0}}\right)-m(\mathbf{x},\boldsymbol{\theta})=\left[\left(\boldsymbol{\theta}_{\mathbf{0}}-\boldsymbol{\theta}\right)\mathbf{x}\right]^{2}=\left(\boldsymbol{\theta}_{\mathbf{0}}-\boldsymbol{\theta}\right)'\mathbf{x}'\mathbf{x}\left(\boldsymbol{\theta}_{\mathbf{0}}-\boldsymbol{\theta}\right)$$

$$E\left\{\left[m\left(\mathbf{x},\boldsymbol{\theta}_{\mathbf{0}}\right)-m(\mathbf{x},\boldsymbol{\theta})\right]^{2}\right\}=\left(\boldsymbol{\theta}_{\mathbf{0}}-\boldsymbol{\theta}\right)'E\left(\mathbf{x}'\mathbf{x}\right)\left(\boldsymbol{\theta}_{\mathbf{0}}-\boldsymbol{\theta}\right)$$

  For the last expression to be positive for all $\boldsymbol{\theta}\neq\boldsymbol{\theta}_{\mathbf{0}}$, we need $E\left(\mathbf{x}'\mathbf{x}\right)$ to have full rank K, which is exactly Assumption OLS.2.

- Theorem 12.2 (Wooldridge) contains a formal consistency result for general M-estimators. Practically important restriction is continuity of $q(\mathbf{w},\cdot)$.

Identification condition

Asymptotic Distribution

- In the previous section, we showed how consistency of M-estimators can be established without having closed form solutions. Now we turn to the question of approximating the sampling distribution of $\hat{\boldsymbol{\theta}}$.
- We now add some smoothness assumptions. In particular, assume $q(\mathbf{w}, \cdot)$ is twice continuously differentiable on $\operatorname{int}(\boldsymbol{\Theta})$.
- Further, assume $\boldsymbol{\theta_0}$ is in the interior of the parameter space:

$$\boldsymbol{\theta_0} \in \operatorname{int}(\boldsymbol{\Theta}).$$

Two Basic Concepts: Score

- The gradient of $q(\mathbf{w}, \boldsymbol{\theta})$, defined on $\mathrm{int}(\boldsymbol{\Theta})$, is the $1 \times P$ row vector

  $$\nabla_\theta q(\mathbf{w}, \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial q(w,\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial q(w,\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial q(w,\boldsymbol{\theta})}{\partial \theta_P} \end{pmatrix}.$$

  The score is the transpose of the gradient:

  $$\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}) = \nabla_\theta q(\mathbf{w}, \boldsymbol{\theta})'.$$

- Because $\hat{\boldsymbol{\theta}}$ minimizes the sample objective function and is an interior solution, $\hat{\boldsymbol{\theta}}$ solves

  $$\sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right) = \mathbf{0}$$

  a set of P equations in P unknowns. (Many algorithms to actually find $\hat{\boldsymbol{\theta}}$ are based on this first order condition.) Because $q(\mathbf{w}, \cdot)$ is twice continuously differentiable, each $s_m(\mathbf{w}, \cdot), m = 1, \ldots, P$, is continuously differentiable.

Two Basic Concepts: Hessian

- By the mean value theorem (for each element of the score),

$$\sum_{i=1}^{N} s_m\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{N} s_m\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right) + \left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} s_m\left(\mathbf{w}_i, \ddot{\boldsymbol{\theta}}_m\right)\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right)$$

  where $\ddot{\boldsymbol{\theta}}_m$ is on the line segment between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$ for $m = 1, \ldots, P$. Therefore, $\ddot{\theta}_m \xrightarrow{p} \theta_o$. (In effect, $\ddot{\theta}_m$ is "trapped" beween $\hat{\theta}$ and θ_o.)

- Stack all P elements to get

$$\sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right) + \left(\sum_{i=1}^{N} \ddot{\mathbf{H}}_i\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right),$$

  where $\ddot{\mathbf{H}}_i$ is the $P \times P$ Hessian of $q(\mathbf{w}, \boldsymbol{\theta})-$ also the Jacobian of $\mathbf{s}(\mathbf{w}, \boldsymbol{\theta})$

Substitute in F.O.C.

- Back to the score representation. Because $\hat{\boldsymbol{\theta}}$ solves the FOC,

$$\mathbf{0} = \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right) + \left(\sum_{i=1}^{N} \ddot{\mathbf{H}}_i\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right)$$

so

$$\mathbf{0} = N^{-1/2} \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right) + \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i\right) \sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right).$$

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right) = \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i\right)^{-1} \left[N^{-1/2} \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right)\right]$$

Two Conditions

1. very generally the score has zero mean when evaluated at θ_0 :

$$E\left[\mathbf{s}\left(\mathbf{w}, \boldsymbol{\theta}_o\right)\right] = \mathbf{0}.$$

Why is $E\left[\mathbf{s}\left(\mathbf{w}, \boldsymbol{\theta}_o\right)\right] = 0$ important? Because then, by the central limit theorem,

$$N^{-1/2} \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right) \xrightarrow{d} \text{Normal}\left(\mathbf{0}, \mathbf{B}_o\right)$$

$$\mathbf{B}_o = \text{Var}\left[\mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right)\right] = E\left[\mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right) \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right)'\right].$$

Related to interchangeable differentiation and integral

2. Because each $\ddot{\theta}_m \xrightarrow{p} \theta_o, N^{-1} \sum_{i-1}^{N} \ddot{\mathbf{H}}_i \xrightarrow{p} E\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_o\right)\right] \equiv \mathbf{A}\left(\boldsymbol{\theta}_o\right) \equiv \mathbf{A}_0$
An assumption related to identification is that $\mathbf{A}_o$ is positive definite.

Apply Asymptotic Equivalence

- Now

$$\begin{aligned}\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o\right) &= \mathbf{A}_o^{-1}\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{s}\left(\mathbf{w}_i,\boldsymbol{\theta}_o\right)\right] \\ &\quad + \left[\left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{H}}_i\right)^{-1}-\mathbf{A}_o^{-1}\right]\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{s}\left(\mathbf{w}_i,\boldsymbol{\theta}_o\right)\right] \\ &= \mathbf{A}_o^{-1}\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{s}\left(\mathbf{w}_i,\boldsymbol{\theta}_o\right)\right]+o_p(1)\cdot O_p(1) \\ &= \mathbf{A}_o^{-1}\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{s}\left(\mathbf{w}_i,\boldsymbol{\theta}_o\right)\right]+o_p(1).\end{aligned}$$

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o\right)\xrightarrow{d}\text{Normal}\left(\mathbf{0},\mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}\right).$$

- Generally, the asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o\right)$ depends on the expected value of the Hessian and the variance of the score (both evaluated at $\boldsymbol{\theta}_o$).

Estimating the Asymptotic Variance

- Technically, we must talk about consistent estimation of $Avar\left[\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o\right)\right]$, as this is the quantity that does not depend on N. So we must consistently estimate $\mathbf{A}_o$ and $\mathbf{B}_o$.

- There are sometimes several different ways to estimate $\mathbf{A}_o$. An estimator that is always available is simply

$$N^{-1}\sum_{i=1}^{N}\mathbf{H}\left(\mathbf{w}_i,\hat{\boldsymbol{\theta}}\right)=N^{-1}\sum_{i=1}^{N}\mathbf{H}_i(\hat{\boldsymbol{\theta}}),$$

  the average of the Hessians evaluated at the estimates.

Estimating the Asymptotic Variance

- When $\mathbf{w}_i$ partitions as $(\mathbf{x}_i, \mathbf{y}_i)$, and we are correctly modeling a feature of $D(\mathbf{y}_i \mid \mathbf{x}_i)$, we can often find

$$\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) = E[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o) \mid \mathbf{x}_i].$$

  By iterated expectations, $\mathbf{A}_o = E[\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o)]$. So a second consistent estimator of $\mathbf{A}_o$ is sometimes available:

$$N^{-1} \sum_{i=1}^{N} \mathbf{A}\left(\mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) = N^{-1} \sum_{i=1}^{N} \hat{\mathbf{A}}_i.$$

- It is rarely possible to find the unconditional expected value of $\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o)$ when there are conditioning variables because we are not usually modeling $D(\mathbf{x}_i)$.

Estimating the Asymptotic Variance

- A natrual consistent estimator of $\mathbf{B}_o = E\left[\mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right)\mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_o\right)'\right]$ is

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\mathbf{s}\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right)\mathbf{s}\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right)' = N^{-1}\sum_{i=1}^{N}\mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})' = N^{-1}\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'.$$

- Called the outer product of the score.
- Therefore,

$$\begin{aligned}\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) &= N^{-1}\left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'\right)\left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1} \\ &= \left(\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'\right)\left(\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}\end{aligned}$$

- As with all other procedures, the divions by N disappear in $\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}})$.

MLE as Example

- The motivation for MLE in introductory statistics is intuitively appealing, but it does not directly lead to a verification of consistency. In fact, we will apply the M-estimation results to the objective function

$$q(\mathbf{w}_i, \theta) = -\log f(\mathbf{y}_i \mid \mathbf{x}_i; \theta)$$

- $\ell_i(\theta) \equiv \log f(\mathbf{y}_i \mid \mathbf{x}_i, \theta)$ called the log-likelihood function for observation i. It is random because it depends on $(\mathbf{x}_i, \mathbf{y}_i)$, but we are interested in it as a function of θ.

MLE as Example

- So $f(\mathbf{y} \mid \mathbf{x}; \theta_o)$ is the true density of $\mathbf{y}_i$ given $\mathbf{x}_i = \mathbf{x}$ for all $\mathbf{x} \in \mathcal{X}$.
- The (Conditional) Maximum Likelihood Estimator of θ_o, $\hat{\theta}$:

$$\max_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} \log f(\mathbf{y}_i \mid \mathbf{x}_i; \theta).$$

- Note that this is the starting point. The key is to show that the log likelihood identifies θ_o. This follows by the *Kullback-Leibler Information Inequality*. For our purposes, it implies that

$$E\left[\log f(\mathbf{y}_i \mid \mathbf{x}_i; \theta_o) \mid \mathbf{x}_i\right] \geq E\left[\log f(\mathbf{y}_i \mid \mathbf{x}_i; \theta) \mid \mathbf{x}_i\right], \text{ all } \theta \in \Theta$$

and so

$$E\left[\log f(\mathbf{y}_i \mid \mathbf{x}_i; \theta_o)\right] \geq E\left[\log f(\mathbf{y}_i \mid \mathbf{x}_i; \theta)\right], \text{ all } \theta \in \Theta$$

MLE as Example

- Provided $\ell_i(\theta) \equiv \log f\left(\mathbf{y}_i \mid \mathbf{x}_i ; \theta\right)$ is continuous in θ and that enough moments of the log likelihood are bouned across θ, the MLE is generally consistent. Just apply the M-estimation consistency result directly.

Asymptotic distribution of MLE

- Denote the score of the log likelihood as the $P \times 1$ vector

$$\mathrm{s}_i(\theta) = \mathbf{s}\left(\mathbf{x}_i, \mathbf{y}_i, \theta\right) = \nabla_\theta \log f\left(\mathbf{y}_i \mid \mathbf{x}_i; \theta\right)' = \nabla_\theta \ell_i(\theta)'$$

  Further, the Hessian is still the Jacobian of the score:

$$\mathbf{H}_i(\theta) = \mathbf{H}\left(\mathbf{x}_i, \mathbf{y}_i, \theta\right) = \nabla_\theta \mathbf{s}_i(\theta)$$

- A slight notational change from M-estimation:

$$\mathbf{A}_o = -E\left[\mathbf{H}_i\left(\theta_o\right)\right]$$

$$\mathbf{A}\left(\mathbf{x}_i, \theta_o\right) = -E\left[\mathbf{H}_i\left(\theta_o\right) \mid \mathbf{x}_i\right]$$

  so that $\mathbf{A}\left(\mathbf{x}_i, \theta_o\right)$ is postive semi-definite and $\mathbf{A}_o$ is pd.

- As before, let

$$\mathbf{B}_o = E\left[\mathbf{s}_i\left(\theta_o\right) \mathbf{s}_i\left(\theta_o\right)'\right].$$

Asymptotic distribution of MLE

Fisher consistency:

$$\max_{\theta \in \Theta} E\left[\log f\left(\mathbf{y}_i \mid \mathbf{x}_i; \theta\right) \mid \mathbf{x}_i\right]$$

the score generally satisfies

$$E\left[\mathbf{s}_i\left(\theta_o\right) \mid \mathbf{x}_i\right] = 0$$

and so

$$E\left[\ \mathrm{s}_i\left(\theta_o\right)\right] = 0.$$

unconditional information matrix equality (UIME)

$$-E\left[\mathbf{H}_i\left(\theta_o\right) \mid \mathbf{x}_i\right] = E\left[\mathbf{s}_i\left(\theta_o\right)\mathbf{s}_i\left(\theta_o\right)' \mid \mathbf{x}_i\right]$$

(Check textbook for smoothness conditions)
In the notation of M-estimation,

$$\mathbf{A}_o = \mathbf{B}_o.$$

Asymptotic distribution of MLE

- Therefore, for correctly specified (conditional) maximum likelihood problems,

$$A\,\mathrm{var}\left[\sqrt{N}\left(\hat{\theta}-\theta_o\right)\right]=\mathbf{A}_o^{-1}=\mathbf{B}_o^{-1}.$$

- So, generally, one chooses among three estimators of $\mathrm{Avar}(\hat{\theta})$:

$$\left(\sum_{i=1}^{N}-\mathbf{H}_i(\hat{\theta})\right)^{-1},\left(\sum_{i=1}^{N}\mathbf{A}_i(\hat{\theta})\right)^{-1},\left(\sum_{i=1}^{N}\mathrm{s}_i(\hat{\theta})\mathrm{s}_i(\hat{\theta})'\right)^{-1}.$$

- The outer product of the score formulation, while computationally simple, can have severe finite-sample bias; usually the standard errrors are too small on average.

- The Hessian and expected Hessian forms tend to work well. In leading cases, the expected Hessian form depends only on first derivatives.