2 Basic Approach (theory).

$\underbrace{\text{M-estimators}}_{\hat{\theta}}$ are solutions of the vector equation (iid case)

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

but what are they estimating? Some true parameter $\underline{\theta}_0$, where

$$(*) \quad E_F[\underline{\psi}(\underline{Y}_1; \underline{\theta}_0)] = \int \underline{\psi}(y; \underline{\theta}_0)\, dF(y) = 0 \quad \text{where} \quad Y_1 \sim F$$

Example (Sample Mean, cont'd): Recall we said $\theta = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is an M-estimator for $\psi(Y_i, \theta) = Y_i - \theta$. What is the true parameter?

The true parameter solves $\int (y - \theta)\, dF(y) = 0$

$$\Rightarrow \underbrace{\int y\, dF(y)}_{\text{population mean}} = \theta$$

Recall the 5-dimensional motivating example.

We said the α which maximizes the pairwise log likelihood seems like it would be a good estimator for α_0.

We didn't show this.

To do this, we would need to use (*)

To arrive at the sandwich estimator, assume $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{iid}{\sim} F$ and define

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{\theta}).$$

↑ depend on n.

In the likelihood case:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} \underbrace{\underline{\psi}(\underline{Y}_i; \underline{\theta})}_{\text{score contribution, deriv of log likelihood contribution}}}_{\text{mean derivative of log likelihood contributions.}}$$

Taylor expansion of $\boldsymbol{G}_n(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$ evaluated at $\hat{\boldsymbol{\theta}}$ yields

$$\underset{\text{definition of }\hat{\theta}}{0} = \underline{G}_n(\hat{\underline{\theta}}) = \underset{b\times 1}{\underline{G}_n(\underline{\theta}_0)} + \underset{b\times b \text{ Jacobian}}{\underline{G}_n'(\underline{\theta}_0)} \underbrace{(\hat{\underline{\theta}} - \underline{\theta}_0)}_{b\times 1} + \underset{\text{higher order terms "residual"}}{\underline{R}_n}$$

Rearranging: $-\underline{G}_n'(\underline{\theta}_0)(\hat{\underline{\theta}} - \underline{\theta}_0) = \underline{G}_n(\underline{\theta}_0) + \underline{R}_n$

$$\hat{\underline{\theta}} - \underline{\theta}_0 = \{-\underline{G}_n'(\underline{\theta}_0)\}^{-1} \underline{G}_n(\underline{\theta}_0) + \underbrace{\{-\underline{G}_n'(\underline{\theta}_0)\}^{-1}\underline{R}_n}_{R_n^*}$$

$$\sqrt{n}(\hat{\underline{\theta}} - \underline{\theta}_0) = \underbrace{\{-\underline{G}_n'(\underline{\theta}_0)\}^{-1}}_{1} \underbrace{\sqrt{n}\,\underline{G}_n(\underline{\theta}_0)}_{2} + \underbrace{\sqrt{n}\,\underline{R}_n^*}_{3}$$

We will look at each piece

① $* \quad -G_n'(\underline{\theta}_0) = \frac{d}{d\underline{\theta}} - G_n(\underline{\theta}_0) = \frac{d}{d\underline{\theta}}\left[-\frac{1}{n}\sum_{i=1}^{n} \underline{\psi}(Y_i, \underline{\theta}_0)\right] = \frac{1}{n}\sum_{i=1}^{n} -\underline{\psi}'(Y_i; \theta_0)$

Define $\boldsymbol{A}(\boldsymbol{\theta}_0) = \mathrm{E}_F[-\boldsymbol{\psi}'(\boldsymbol{Y}_1, \boldsymbol{\theta}_0)]$.

Then $-G_n'(\underline{\theta}_0) \xrightarrow{p} \underline{A}(\underline{\theta}_0)$ by WLLN.

In the likelihood setting, what is A? curvature! because ψ is the score function (derivative of log likelihood)

$\Rightarrow \psi'$ is the 2nd derivative of the log likelihood.

② $* \quad \sqrt{n}\, G_n(\underline{\theta}_0) = \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} \underline{\psi}(Y_i; \underline{\theta}_0) \xrightarrow{d} N(\underline{0}, B(\underline{\theta}_0))$ ← why? because we have a correctly scaled sum of iid things (CLT).

What is $B(\underline{\theta}_0)$? Should be the variance of $\underline{\psi}$.

$$B(\underline{\theta}_0) = E_F\{\psi(Y_1, \underline{\theta}_0)\psi(Y_1, \underline{\theta}_0)^T\}.$$

③ $* \quad \sqrt{n} R_n^* \xrightarrow{p} \underline{0}$

This is the "hard part" to prove. We will skip, see Huber (1967) or Serfling (1980).

So putting ①, ②, ③ together

$$\sqrt{n}(\underline{\hat{\theta}} - \underline{\theta}_0) \xrightarrow{d} \{A(\theta_0)\}^{-1} N(\underline{0}, B(\underline{\theta}_0)) \quad \text{(Slutsky's)}.$$

$$\xrightarrow{d} N(\underline{0}, A(\underline{\theta}_0)^{-1} B(\underline{\theta}_0)\{A(\underline{\theta}_0)^{-1}\}^T)$$

or $\quad \hat{\theta} \overset{\cdot}{\sim} N(\underline{\theta}_0, \frac{1}{n} A(\underline{\theta}_0)^{-1} B(\underline{\theta}_0)\{A(\underline{\theta}_0)\}^T)$

In practice, we don't know $\underline{\theta}_0 \Rightarrow$ replace w/ $\hat{\theta}$:

$$\hat{\theta} \overset{\cdot}{\sim} N\left(\underline{\theta}_0, \frac{1}{n} \underbrace{A(\hat{\theta})^{-1}}_{\substack{\text{curvature} \\ \text{bread}}} \underbrace{B(\hat{\theta})}_{\substack{\text{variance} \\ \text{meat}}} \underbrace{\{A(\hat{\theta})^{-1}\}^T}_{\substack{\text{curvature.} \\ \text{bread}}}\right)$$

bread meat bread = sandwich!

2.1 Estimators for $\boldsymbol{A}$, $\boldsymbol{B}$

If the data truly come from the assumed parametric family $f(y;\boldsymbol{\theta})$,

Then $A(\underline{\theta}_0) = B(\underline{\theta}_0) = \underbrace{I(\underline{\theta}_0)}$

Information matrix. where the 2 definitions of $I(\underline{\theta}_0)$ are used.

$\Rightarrow$ the sandwich estimator $A(\underline{\theta}_0)^{-1} B(\underline{\theta}_0) (A^{-1}(\underline{\theta}_0))^T = I(\underline{\theta}_0)^{-1}$

One of the key contributions of M-estimation theory is to point out what happens when the assumed parametric family is not correct.

Then $A(\underline{\theta}_0) \neq B(\underline{\theta}_0)$ and we should use the correct limiting dsn covariance matrix.

$$A(\underline{\theta}_0)^{-1} B(\underline{\theta}_0) \{A^{-1}(\underline{\theta}_0)\}^T$$

We can use empirical estimators of $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$A_n(\underline{Y}, \hat{\underline{\theta}}) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \{-\psi'(Y_i, \hat{\theta})\}}_{\text{average curvature evaluated at } \hat{\theta}}$$

$$B_n(\underline{Y}, \hat{\underline{\theta}}) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \underline{\Psi}(Y_i, \hat{\theta}) \underline{\Psi}(Y_i, \hat{\theta})^T}_{\text{variance estimate.}}$$

might need to use numeric differentiation to approximate.

Remember, the Hessian in code is $nA_n(\underline{Y}, \hat{\underline{\theta}})$.

Example (Coefficient of Variation): Let $Y_1, \ldots, Y_n$ be iid from some distribution with finite fourth moment. The coefficient of variation is defined at $\hat{\theta}_3 = s_n / \overline{Y}$.

How would we get a CI for the coefficient of variation, $\theta_3 = \frac{\sigma}{\mu}$?

Bootstrap? probably.

We'll try M-estimation.

Define a three dimensional $\boldsymbol{\psi}$ so that $\hat{\theta}_3$ is defined by summing the third component. What is the vector valued function $\boldsymbol{\psi}$ which yields an M-estimator for the coefficient of variation?

$$\underline{\Psi}(Y_i, \underline{\theta}) = \begin{pmatrix} Y_i - \theta_1 \\ (Y_i - \theta_1)^2 - \theta_2 \\ \theta_1\theta_3 - \sqrt{\theta_2} \end{pmatrix}$$

$$\sum_{i=1}^n \underline{\Psi}(Y_i, \underline{\theta}) = \begin{pmatrix} \sum_{i=1}^n Y_i - n\theta_1 \\ \sum_{i=1}^n (Y_i - \theta_1)^2 - n\theta_2 \\ n\theta_1\theta_3 - \sqrt{\theta_2} \end{pmatrix} \overset{set}{=} 0$$

$$\Rightarrow \theta_1 = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}$$

$$\theta_2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \theta_1)^2 = S_{(n)}^2$$

$\leftarrow$ divide by n, not n-1

$$\theta_3 = \frac{\sqrt{\theta_2}}{\theta_1}$$

What parameter vector is being estimated by the M-estimator?

$$E[\psi_1(Y_i,\underline{\theta})] = E[Y_i - \theta_1] = \mu - \theta_1 = 0 \Rightarrow \theta_1 = \mu.$$

$$E[\psi_2(Y_i,\underline{\theta})] = E[(Y_i-\theta_1)^2 - \theta_2] = 0 \Rightarrow \theta_2 = Var(Y_1) = \sigma^2$$

$$E[\psi_3(Y_i,\underline{\theta})] = E[\theta_1\theta_3 - \sqrt{\theta_2}] = \theta_1\theta_3 - \sqrt{\theta_2} \overset{set}{=} 0 \Rightarrow \theta_3 = \frac{\sqrt{\theta_2}}{\theta_1}.$$

$$\Rightarrow \begin{pmatrix} \mu \\ \sigma^2 \\ \frac{\sigma}{\mu} \end{pmatrix}.$$

What are the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$?

$$A = E[-\underline{\psi}'(Y_i,\underline{\theta}_0)] \qquad \underline{\Psi}(Y_i,\underline{\theta}) = \begin{pmatrix} Y_i - \theta_1 \\ (Y_i-\theta_1)^2 - \theta_2 \\ \theta_1\theta_3 - \sqrt{\theta_2} \end{pmatrix}.$$

$$\underline{\psi}' = \begin{pmatrix} -1 & 0 & 0 \\ -2(Y_i-\theta_1) & -1 & 0 \\ \theta_3 & -\frac{1}{2}\theta_2^{-1/2} & \theta_1 \end{pmatrix}$$

$$A = E[-\psi'(Y_i,\underline{\theta}_0)] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{\sigma}{\mu} & \frac{1}{2\sigma} & -\mu \end{pmatrix}$$

$$B = E[\underline{\Psi}(Y_1,\underline{\theta}_0)\Psi(Y,\underline{\theta}_0)^T]$$

$$= E\begin{bmatrix} (Y_i-\theta_1)^2 & (Y_i-\theta_1)[(Y_i-\theta_1)^2-\theta_2] & (Y_i-\theta_1)(\theta_1\theta_3-\sqrt{\theta_2}) \\ (Y_i-\theta_1)[(Y_i-\theta_1)^2-\theta_2] & [(Y_i-\theta_1)^2-\theta_2]^2 & [(Y_i-\theta_1)^2-\theta_2](\theta_1\theta_3-\sqrt{\theta_2}) \\ (Y_i-\theta_1)(\theta_1\theta_3-\sqrt{\theta_2}) & [(Y_i-\theta_1)^2-\theta_2](\theta_1\theta_3\sqrt{\theta_2}) & (\theta_1\theta_3-\sqrt{\theta_2}) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & \mu_3 & 0 \\ \mu_3 & \mu_4 - \sigma^4 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{where } \mu_j = E(Y_1-\theta_1)^j$$

Write out the asymptotic variance, $\boldsymbol{V}$.

$$V = A^{-1} B (A^{-1})^T$$

Using row operations (not shown):

$$A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{\sigma}{\mu^2} & \frac{1}{2\sigma\mu} & -\frac{1}{\mu} \end{pmatrix}$$

$$A^{-1}B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{\sigma}{\mu^2} & \frac{1}{2\sigma\mu} & -\frac{1}{\mu} \end{pmatrix} \begin{pmatrix} \sigma^2 & \mu_3 & 0 \\ \mu_3 & \mu_4 - \sigma^4 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma^2 & \mu_3 & 0 \\ \mu_3 & \mu_4 - \sigma^4 & 0 \\ -\frac{\sigma^3}{\mu_2} + \frac{\mu_3}{2\sigma\mu} & -\frac{\sigma\mu_3}{\mu_2} + \frac{\mu_4 - \sigma^4}{2\mu} & 0 \end{pmatrix}$$

$$\Rightarrow A^{-1}B(A^{-1})^T = \;\ldots\ldots$$

$$= \begin{pmatrix} \sigma^2 & \mu_3 & -\frac{\sigma^3}{\mu^2} + \frac{\mu_3}{2\sigma\mu} \\ \mu_3 & \mu_4 - \sigma^4 & -\frac{\sigma\mu_3}{\mu^2} + \frac{\mu_4 - \sigma^4}{2\sigma\mu} \\ -\frac{\sigma^3}{\mu_2} + \frac{\mu_3}{2\sigma\mu} & -\frac{\sigma\mu_3}{\mu_2} + \frac{\mu_4 - \sigma^4}{2\sigma\mu} & \underbrace{\frac{\sigma^4}{\mu^4} - \frac{\sigma\mu_3}{\sigma\mu^3} + \frac{\mu_4 - \sigma^4}{4\sigma^2\mu^2}}_{} \end{pmatrix}$$

Assume Y_i are iid from a normal distribution with mean 10 and standard deviation 1. Calculate $V_{3,3}$. Assume you have a same of size 25 and you get an estimated coefficient of variation of 0.11. Give the asymptotic 95% confidence interval.

$$V_{3,3} = \frac{\sigma^4}{\mu^4} - \frac{\sigma \mu_3}{2\sigma \mu^3} + \frac{\mu_4 - \sigma^4}{4\sigma^2 \mu^2}$$

If $Y \sim N(10,1)$, $\mu_3 = 0$, $\mu_4 = 3$ (looked up).

$$\Rightarrow V_{3,3} = \frac{1}{10^4} - 0 + \frac{3-1}{4\cdot 1\cdot 10^2} = \frac{1}{10000} + \frac{1}{200} = .0051.$$

$$n = 25 \Rightarrow Var(\hat{\theta}_3) = \frac{.0051}{25} = .000204.$$

CI: $0.11 \pm 1.96\sqrt{2.04e^{-4}}$

(0.082, 0.138)