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

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

In the likelihood case:

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

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

$$\underline{0} = \underbrace{G_n(\hat{\underline{\theta}})}_{b\times 1} = \underset{\uparrow\, b\times 1}{G_n(\underline{\theta}_0)} + \underbrace{G_n'(\underline{\theta}_0)}_{b\times b \text{ Jacobian}} \underbrace{(\hat{\underline{\theta}} - \underline{\theta}_0)}_{b\times 1} + R_n \quad \nwarrow \text{higher order "residual"}$$

Rearranging:

$$-G_n'(\underline{\theta}_0)(\hat{\underline{\theta}} - \underline{\theta}_0) = G_n(\underline{\theta}_0) + R_n$$

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

$$\sqrt{n}(\hat{\underline{\theta}} - \underline{\theta}_0) = \underbrace{\{-G_n'(\underline{\theta}_0)\}^{-1}}\underbrace{\sqrt{n}\, G_n(\underline{\theta}_0)} + \underbrace{\sqrt{n}\, R_n^*}$$

Let's look at each piece.

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

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 $\underline{\Psi}$ is the score function (derivative of log-likelihood)

$\Rightarrow \underline{\Psi}'$ is the 2^{nd} derivative of the log-likelihood.

$\star$ $\sqrt{n}\,\underline{G}_n(\underline{\theta}_0) = \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\underline{\Psi}(Y_i, \underline{\theta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underline{\Psi}(Y_i; \underline{\theta}_0) \xrightarrow{d} N(0, B(\underline{\theta}_0))$.

because this is a sum of correctly scaled iid things.

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

$$B(\underline{\theta}_0) = E_F\left[\underline{\Psi}(\underline{Y}_1, \underline{\theta}_0)\,\underline{\Psi}(\underline{Y}_1, \underline{\theta}_0)^T\right]$$

$\star$ $\sqrt{n}\,\underline{R}_n^* \xrightarrow{P} \underline{0}$

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

So, putting $\star$'s together,

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

$$\rightarrow N\left(\underline{0}, A(\underline{\theta}_0)^{-1} B(\underline{\theta}_0)\{A(\underline{\theta}_0)^{-1}\}^T\right)$$

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

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

$$\hat{\underline{\theta}} \overset{\circ}{\sim} N\left(\underline{\theta}_0, \frac{1}{n}A(\hat{\underline{\theta}})^{-1} B(\hat{\underline{\theta}})\{A(\hat{\underline{\theta}})^{-1}\}^T\right)$$

↑ curvature ↑ variance ↑ curvature

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 $A(\underline{\theta}_0) = B(\underline{\theta}_0) = \underbrace{I(\underline{\theta}_0)}_{\text{information matrix.}}$

where $A(\underline{\theta}_0)$ and $B(\underline{\theta}_0)$ are the 2 definitions of $I(\underline{\theta}_0)$.

$\Longrightarrow$ The sandwich estimator $A(\underline{\theta}_0)^{-1} B(\underline{\theta}_0) \{A(\underline{\theta}_0)^{-1}\}^T = I(\underline{\theta}_0)^{-1}$ ✓

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 distribution covariance matrix.

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

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

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

$$B_n(\underline{Y}, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \underline{\psi}(Y_i; \hat{\theta}) \underline{\psi}(Y_i; \hat{\theta})^T$$

variance estimate.

might need to use numeric differentiation to approximate.

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

$\Longrightarrow$ need $\{nA_n(\hat{\underline{\theta}})\}^{-1} \{n^2 B_n(\hat{\theta})\} \{\{nA_n(\hat{\theta})\}^{-1}\}^T$

Example (Coefficient of Variation): Let $Y_1, \dots, Y_n$ 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 $\theta_3 = \frac{\sigma}{\mu}$?

unknown dsn.

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?

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

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

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

division by n, not $n-1$.

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

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

note: not a function of data,

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

$$E[\psi_1(Y_1,\underline{\theta})] = E[Y_1-\theta_1] = \mu-\theta_1 \overset{\text{set}}{=} 0 \Rightarrow \theta_1=\mu$$

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

$$E[\psi_3(Y_1,\underline{\theta})] = E[\theta_1\theta_3-\sqrt{\theta_2}] = \theta_1\theta_3-\sqrt{\theta_2} \overset{\text{set}}{=} 0 \Rightarrow \theta_3 = \frac{\sqrt{\theta_2}}{\theta_1}$$

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

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

$$A = E[-\underline{\psi}'(Y_1,\underline{\theta}_0)], \qquad \underline{\psi}(Y_1,\underline{\theta}) = \begin{pmatrix} Y_1-\theta_1 \\ (Y_1-\theta_1)^2-\theta_2 \\ \theta_1\theta_3-\sqrt{\theta_2} \end{pmatrix}$$

$$\underline{\psi}' = \begin{pmatrix} -1 & 0 & 0 \\ -2(Y_1-\theta_1) & -1 & 0 \\ \theta_3 & -\frac{1}{2}\theta_2^{-1/2} & \theta_1 \end{pmatrix}$$

$$A = E[-\psi'(Y_1,\underline{\theta}_0)] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{\sigma}{\mu} & \frac{1}{2\sigma} & -\mu \end{bmatrix}$$

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

$$= E\begin{bmatrix} (Y_1-\theta_1)^2 & (Y_1-\theta_1)[(Y_1-\theta_1)^2-\theta_2] & (Y_1-\theta_1)(\theta_1\theta_3-\sqrt{\theta_2}) \\ (Y_1-\theta_1)[(Y_1-\theta_1)^2-\theta_2] & [(Y_1-\theta_1)^2-\theta_2]^2 & [(Y_1-\theta_1)^2-\theta_2](\theta_1\theta_3-\sqrt{\theta_2}) \\ (Y_1-\theta_1)(\theta_1\theta_3-\sqrt{\theta_2}) & [(Y_i-\theta_1)^2-\theta_2][\theta_1\theta_3-\sqrt{\theta_2}] & (\theta_1\theta_3-\sqrt{\theta_2})^2 \end{bmatrix}$$

$$= \begin{pmatrix} \sigma^2 & \mu_3 & 0 \\ \mu_3 & \mu_4-\sigma^4 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{where} \quad \mu_j = E[(Y_i-\theta_1)^j].$$

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

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

$$A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{\sigma}{\mu^2} & \frac{1}{2\sigma\mu} & -\frac{1}{\mu} \end{pmatrix}$$ using row operations (not shown).

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

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

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

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

$V_{3,3} =$ from prev. page.

If $Y_1 \sim N(10,1)$, $\mu_3 = 0$, $\mu_4 = 3$ (properties of moments of N dsn).

$\Rightarrow V_{3,3} = .0051$

$n = 25 \Rightarrow \text{var}(\hat{\theta}_3) = \dfrac{.0051}{25} = .000204$

CI: $.11 \pm 1.96\sqrt{2.04e^{-4}}$

⋮

$(.082, .138)$