6.S951 Modern Mathematical Statistics Fall 2024

Lecture 23 — December 5, 2024

Prof. Stephen Bates

Scribe: Charlie Cowen-Breen

Outline

Today: Asymptotic Normality of M-estimators

- Main goal
  - Heuristic derivation
  - Formal Statements
- Examples
  - Linear regression
  - Robust standard error (CIs)
  - Logistic regression

Recap (asymptotics):

- CLT, Slutsky
- Delta method
- Moment estimators
- M-estimator consistency
  - Uniform convergence + separation - IDS.160

1 Main results

Setting:

$$X_1 \dots X_n \overset{i.i.d.}{\sim} P \qquad \text{on } \mathcal{X}$$

$$m_\theta(x_i) : \Theta \times \mathcal{X} \longrightarrow \overline{\mathbb{R}}$$

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(x_i)$$

$$M(\theta) = \mathbb{E}[m_\theta(x_i)]$$

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} M_n(\theta)$$
$$\theta_0 = \text{argmax}_{\theta \in \Theta} M(\theta)$$

Assume

$$\Psi_\theta(x) := \dot{m}_\theta(x) = \frac{\partial}{\partial \theta} m_\theta(x)$$

exists. Let $\hat{\theta}$ solve

$$\frac{1}{n}\sum_{i=1}^n \Psi_\theta(x_i) = 0$$

and θ_0 solve

$$\mathbb{E}\Psi_\theta(x_i) = 0.$$

Then we have

Theorem 1 (Main result, informal statement)**.** *Assume $\hat{\theta} \xrightarrow{p} \theta_0$ and regularity conditions. Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Gamma)$$

where

$$\Gamma = V_{\theta_0}^{-1} \mathbb{E}\left[\Psi_{\theta_0} \Psi_{\theta_0}^\top\right] \left(V_{\theta_0}^{-1}\right)^\top$$

and V_{θ_0} is the derivative of the map $\theta \mapsto \mathbb{E}\Psi_\theta(x_i)$.

<u>*Heuristic derivation*</u>*.* Main idea: Taylor expansion, and apply CLT and LNN. Set

$$\tilde{\Psi}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \Psi_\theta(x_i)$$
$$\tilde{\Psi}(\theta) = \mathbb{E}\Psi_\theta(x_i)$$

Now, $\hat{\theta}$ is near θ_0 by consistency, so we Taylor expand around θ_0:

$$0 = \tilde{\Psi}_n(\hat{\theta}) = \tilde{\Psi}_n(\theta_0) + (\hat{\theta} - \theta_0)\dot{\tilde{\Psi}}_n(\theta_0) + \underbrace{\frac{1}{2}(\hat{\theta} - \theta_0)^2 \ddot{\tilde{\Psi}}_n(\tilde{\theta})}_{\text{lower order } O_p(1/n), \text{ for some } \tilde{\theta} \in [\theta_0, \hat{\theta}]}$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = \underbrace{\left[\dot{\tilde{\Psi}}_n(\theta_0)\right]^{-1}}_{\text{LLN}} \cdot \underbrace{\left(-\sqrt{n}\tilde{\Psi}(\theta_0)\right)}_{\text{CLT}}$$
$$\to \mathbb{E}[\dot{\Psi}_\theta(x_i)]$$

□

Theorem 2 (5.4 of vdV: Main result, formal statement)**.** *Suppose $\theta \mapsto \Psi_\theta(x)$ is twice continuously differentiable for $\theta \in \mathbb{R}$. Suppose θ_0 satisfies*

$$\mathbb{E}\Psi_{\theta_0}(x_i) = 0 \qquad \mathbb{E}\Psi_{\theta_0}(x_i)^2 < \infty$$

and $\mathbb{E}\dot{\Psi}_{\theta_0}(x_i)$ *exists and is nonzero for all* θ *in a neighborhood of* θ_0*. Further suppose* $|\ddot{\Psi}_\theta(x)| < f(x)$ *for some integrable function* f*, and* $\tilde{\theta} \xrightarrow{p} \theta_0$*.*[1] *Then the conclusions of the main result hold:*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Gamma)$$

where

$$\Gamma = V_{\theta_0}^{-1}\mathbb{E}\left[\Psi_{\theta_0}\Psi_{\theta_0}^\top\right]\left(V_{\theta_0}^{-1}\right)^\top$$

and V_{θ_0} *is the derivative of the map* $\theta \mapsto \mathbb{E}\Psi_\theta(x_i)$*.*

Proof. Following the heuristic derivation,

$$\sqrt{n}\tilde{\Psi}_n(\theta_0) = \sqrt{n}(\hat{\theta} - \theta_0)\left(\dot{\tilde{\Psi}}_n(\theta_0) + \underbrace{\frac{1}{2}(\hat{\theta} - \theta_0)\ddot{\tilde{\Psi}}(\tilde{\theta})}_{o_p(1)}\right)$$

from which the result follows. **NB:** there could be issues with the above equation somewhere—inconclusive from class. □

We now state a generalized result which does not require twice differentiability.

Theorem 3 (5.21 of vdV: Generalization to non-twice-differentiable $\theta \mapsto \Psi_\theta$)**.** *For* θ *in an open set* $\Theta \subset \mathbb{R}^d$*, suppose that the map* $x \mapsto \Psi_\theta(x)$ *satisfies*

$$\|\Psi_{\theta_1}(x) - \Psi_{\theta_2}(x)\| \leq f(x)\|\theta_1 - \theta_2\|$$

for all θ_1 *and* θ_2 *in a neighborhood of* θ_0*, for some square-integrable* f*, i.e.* $\mathbb{E}f(x)^2 < \infty$*. Assume that* $\mathbb{E}\|\Psi_{\theta_0}\|^2 < \infty$*, and* $\theta \mapsto \mathbb{E}\Psi_\theta(x_i)$ *differentiable at* θ_0 *with derivative* V_{θ_0}*. Then if* $\hat{\theta} \xrightarrow{p} \theta_0$*, the conclusions of the main theorem hold:*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Gamma)$$

where

$$\Gamma = V_{\theta_0}^{-1}\mathbb{E}\left[\Psi_{\theta_0}\Psi_{\theta_0}^\top\right]\left(V_{\theta_0}^{-1}\right)^\top.$$

2 Examples

2.1 Linear regression

Suppose $(X_i, Y_i) \in \mathbb{R}^{d+1}$. In linear regression, we consider M-estimation with

$$m_\theta((x, y)) = -(y - \theta^\top x)^2$$

Thus we have

$$\hat{\theta} = \text{argmax}_\theta -\frac{1}{n}\sum_{i=1}^n (Y_i - \theta^\top X_i)^2$$

[1] Is this necessary?

and

$$\Psi_\theta((x,y)) = 2(y - \theta^\top x)\cdot x$$

We now verify that the conditions of the theorem hold. Suppose that the sample space is bounded, $\|\theta\| < C$. In this case, Ψ is

- Lipschitz ✓
- $\mathbb{E}\left[\Psi_{\theta_0}\Psi_{\theta_0}^\top\right] = \mathbb{E}\left[4(Y_i - \theta_0^\top X_i)^2 X_i X_i^\top\right] < \infty$
- $\mathbb{E}\Psi_\theta(X_i) = \mathbb{E}2YX - \mathbb{E}2\theta^\top X\cdot X$
  $\implies V_{\theta_0} = -2\mathbb{E}X_i X_i^\top$, which we shall denote Σ_{xx}.

Since the conditions of the theorem hold, we conclude that

$$\sqrt{n}(\hat\theta - \theta_0) \xrightarrow{d} \mathcal{N}(0,\Gamma)$$

where

$$\Gamma = \Sigma_{xx}^{-1}\underbrace{\mathbb{E}\left[(Y_i - X_i^\top\theta)^2 X_i X_i^\top\right]}_{W}\Sigma_{xx}^{-1}.$$

2.2 Robust standard errors

Plug in estimates for Σ_{xx} and W to get confidence intervals for θ_0. The plug-ins are

$$\begin{aligned}
\widehat{\Sigma}_{xx} &= \frac{1}{n}\sum_{i=1}^n X_i X_i^\top \\
\widehat{W} &= \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\top\theta)^2 X_i X_i^\top \\
\widehat{\Gamma} &= \widehat{\Sigma}_{xx}^{-1}\widehat{W}\widehat{\Sigma}_{xx}^{-1}
\end{aligned}$$

C.f. bootstrap here. Then

$$\sqrt{n}\widehat{\Gamma}^{-1/2}(\hat\theta - \theta_0) \xrightarrow{d} \mathcal{N}(0,I) \quad\implies\quad \text{CI for } \theta_{0,j} : \hat\theta_j \pm q(\Gamma_{j,j})^{1/2}/\sqrt{n}$$

For fixed X, the bound works only when the model is actually correct (i.e. linear), with

$$\sigma^2(X^\top X)^{-1}_{jj}$$