STAT 460/560 Class 21: One-Step Estimators and Examples

Ben Bloem-Reddy

Reading: Chapter 5.3 and 5.7, [van98].

1. Example of Z-estimation: Nonlinear least squares

Suppose we have a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution P_{θ_0}, which follows

$$Y = f_{\theta_0}(X) + \epsilon \ , \quad E[\epsilon \mid X] = 0 \ ,$$

where f_{θ_0} belongs to a parametric family of regression functions, for example $f_\theta(x) = \theta_1 + \theta_2 e^{\theta_3 x}$. To estimate θ, the least squares estimator minimizes

$$\theta \mapsto \sum_{i=1}^n (Y_i - f_\theta(X_i))^2 \ .$$

Maximizing the negative of this leads to the M-estimator for $m_\theta(x, y) = -(y - f_\theta(x))^2$. van der Vaart [van98] analyzes the M-estimator in Example 5.27.

Instead, assuming that $\theta \mapsto f_\theta(x)$ is differentiable for each x, we can analyze the corresponding Z-estimator. For simplicity, assume that $Y_i \in \mathbb{R}$, and let $\dot{f}_\theta = \nabla_\theta f_\theta$. Then the Z-estimator $\hat{\theta}_n$ solves

$$\Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^n (Y_i - f_\theta(X_i)) \dot{f}_\theta(X_i) = 0 \ . \tag{21.1}$$

Assuming that $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$, let's apply our asymptotic normality theorem.

Exercise 21.1. Formulate conditions on f_θ under which the local Lipschitz property is satisfied by $\psi_\theta(x, y) = (y - f_\theta(x)) \dot{f}_\theta(x)$.

Activity 21.1. Assume that f_θ is such that the local Lipschitz property is satisfied, that $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$, and that $\hat{P}_n \psi_{\hat{\theta}_n} = o_P(n^{-1/2})$. Find the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. For simplicity, assume that $E[\epsilon^2 | X] = \sigma^2$ almost surely.

Solution: In this case,

$$\begin{aligned} P\psi_\theta &= E[(f_{\theta_0}(X) + \epsilon - f_\theta(X)) \dot{f}_\theta(X)] \\ &= E[(f_{\theta_0}(X) - f_\theta(X)) \dot{f}_\theta(X)] \ . \end{aligned}$$

Clearly, this has a zero at θ_0. Moreover,

$$\begin{aligned} P\psi_{\theta_0}^2 &= E[(f_{\theta_0}(X) + \epsilon - f_{\theta_0}(X))^2 \dot{f}_{\theta_0}(X)^2] \\ &= E[\epsilon^2 \dot{f}_{\theta_0}(X)^2] \\ &= \sigma^2 P \dot{f}_\theta^2 \ . \end{aligned}$$

The derivative "matrix" is $V_\theta = P(\ddot{f}_\theta f_{\theta_0} - \dot{f}_\theta^2 - \ddot{f}_\theta f_\theta)$, which at θ_0 yields $V_{\theta_0} = -P \dot{f}_\theta^2$.

Putting this all together, we find that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{P\dot{f}_\theta^2}\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i - f_{\theta_0}(X_i))\dot{f}_{\theta_0}(X_i) + o_P(1)$$
$$= \frac{1}{P\dot{f}_\theta^2}\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i \dot{f}_{\theta_0}(X_i) + o_P(1) ,$$

so that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, \sigma^2/P\dot{f}_\theta^2) .$$

2. One-step estimators

As van der Vaart points out, the method of Z-estimators can have two disadvantages. First, it may be difficult to find the solutions of the estimating equations. Second, consistency requires that the estimating equations are well-behaved over the entire parameter set. Multiple roots, ill-conditioned numerical problems, etc., can cause issues.

Suppose that we have an estimator $\tilde{\theta}_n$ (presumably obtained my means other than Z-estimation with Ψ_n), and that $\Psi_n(\tilde{\theta}_n) \neq 0$. Perhaps it could be improved upon by following the gradient of Ψ_n at $\tilde{\theta}_n$ towards zero. That is, we can solve the following equation for θ

$$\Psi_n(\tilde{\theta}_n) + \dot{\Psi}_n(\tilde{\theta}_n)(\theta - \tilde{\theta}_n) = 0 \quad \Rightarrow \quad \hat{\theta}_n = \tilde{\theta}_n - \dot{\Psi}(\tilde{\theta}_n)^{-1}\Psi_n(\tilde{\theta}_n) .$$

$\hat{\theta}_n$ is called a **one-step estimator** because it is one iteration (one step) of the Newton–Raphson method for root-finding, as illustrated in Fig. 1. (Refer back to section 7 in Class 9 for more information, including how the algorithm can be used for finding function optima.) As a practical matter, this could be iterated multiple times. It may improve finite-sample performance, but it won't change the asymptotic analysis that follows.

Note, of course, that if $\tilde{\theta}_n$ is found by using Newton–Raphson to solve Ψ_n then this method may or may not be advantageous. One must show that solving the estimating equations with Newton–Raphson produces a consistent estimator. But in some situations, solving the estimating equations is not the only way to obtain a consistent estimator. For example, method-of-moments estimators are consistent under relatively weak conditions. (More on this below.) According to the theorem below, forming a one-step estimator then yields the good asymptotic normality properties effectively separately from achieving consistency.

In order for the theory to work out, we need Ψ_n to satisfy the following. For every constant $M > 0$ and a given nonsingular matrix $\dot{\Psi}_0$,

$$\sup_{\sqrt{n}\|\theta - \theta_0\| < M} \|\sqrt{n}(\Psi_n(\theta) - \Psi_n(\theta_0)) - \dot{\Psi}_0\sqrt{n}(\theta - \theta_0))\| \xrightarrow{\text{P}} 0 . \tag{21.2}$$

This looks like differentiability of Ψ_n at θ_0, but it's weaker than that. As long as there is a sequence of nonsingular (random) matrices $\dot{\Psi}_{n,0}$ that converge in probability to $\dot{\Psi}_0$, then things will work out. Of course, if Ψ_n are differentiable and the derivatives converge to $\dot{\Psi}_0$ then the condition (21.2) will be satisfied. With that, define the one-step estimator by

$$\hat{\theta}_n = \tilde{\theta}_n - \dot{\Psi}_{n,0}^{-1}\Psi_n(\tilde{\theta}_n) . \tag{21.3}$$

Recall that a sequence of estimators $\tilde{\theta}_n$ is called $\sqrt{n}$**-consistent** if $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is bounded in probability. If that is the case then as n gets large, $\tilde{\theta}_n$ is within $n^{-1/2}$ of θ_0 with high probability.

Theorem 21.1. *Let $\sqrt{n}\Psi_n(\theta_0) \rightsquigarrow Z$, for some random variable Z. Suppose that* (21.2) *holds. For a sequence of $\sqrt{n}$-consistent estimators $\tilde{\theta}_n$ and $\dot{\Psi}_{n,0} \xrightarrow{p} \dot{\Psi}_0$, the corresponding one-step estimator $\hat{\theta}_n$ satisfies*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\dot{\Psi}_0^{-1}\sqrt{n}\Psi_n(\theta_0) + o_P(1) .$$

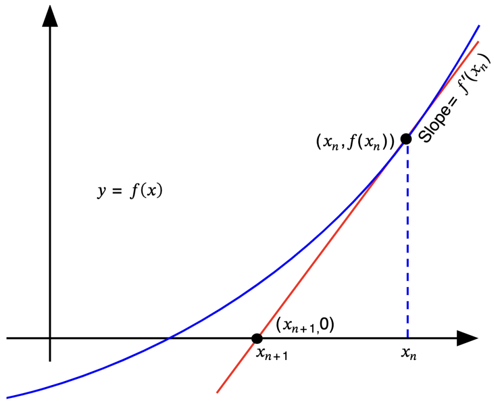

Figure 1: Illustration of one Newton–Raphson iteration for finding the solution to $f(x) = 0$. (Source: https://commons.wikimedia.org/wiki/File:Newton_iteration.svg.)

Proof. The estimator $\dot\Psi_{n,0}\sqrt{n}(\hat\theta_n - \theta_0)$ satisfies

$$\begin{aligned}\dot\Psi_{n,0}\sqrt{n}(\hat\theta_n - \theta_0) &= \dot\Psi_{n,0}\sqrt{n}(\tilde\theta_n - \dot\Psi_{n,0}^{-1}\Psi_n(\tilde\theta_n) - \theta_0) \\ &= \dot\Psi_{n,0}\sqrt{n}(\tilde\theta_n - \theta_0) - \sqrt{n}(\Psi_n(\tilde\theta_n) - \Psi_n(\theta_0)) - \sqrt{n}\Psi_n(\theta_0)\ .\end{aligned}$$

The middle term, by (21.2), can by replaced by $\dot\Psi_0\sqrt{n}(\tilde\theta_n - \theta_0) + o_P(1)$, yielding

$$\begin{aligned}\dot\Psi_{n,0}\sqrt{n}(\hat\theta_n - \theta_0) &= \dot\Psi_{n,0}\sqrt{n}(\tilde\theta_n - \theta_0) - \dot\Psi_0\sqrt{n}(\tilde\theta_n - \theta_0) - \sqrt{n}\Psi_n(\theta_0) + o_P(1) \\ &= \underbrace{(\dot\Psi_{n,0} - \dot\Psi_0)}_{o_P(1)}\underbrace{\sqrt{n}(\tilde\theta_n - \theta_0)}_{O_P(1)} - \sqrt{n}\Psi_n(\theta_0) + o_P(1) \\ &= -\sqrt{n}\Psi_n(\theta_0) + o_P(1)\ .\end{aligned}$$

Therefore, by Slutsky's lemma,

$$\begin{aligned}\sqrt{n}(\hat\theta_n - \theta_0) &= -\dot\Psi_0^{-1}\sqrt{n}\Psi_n(\theta_0) + o_P(1) \\ &\rightsquigarrow \mathcal{N}(0, \dot\Psi_0^{-1}P(\psi_{\theta_0}\psi_{\theta_0}^\top)(\dot\Psi_0^{-1})^\top)\ .\end{aligned}$$

□

3. Method-of-moments estimators (briefly)

These are discussed in more detail in Chapter 4.1 of [van98], but they're pretty much what they sound like. Suppose that for a vector of functions $f = (f_1, \ldots, f_k)$ the function $e\colon \Theta \to \mathbb{R}^k$ is $e(\theta) - P_\theta f$. The **moment estimator** $\hat\theta_n$ satisfies

$$\hat P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i) = P_{\hat\theta_n} f = e(\hat\theta_n)\ . \tag{21.4}$$

Setting $f_j(x) = x^j$ is the method of moments in its simplest form.

If the function e is one-to-one and continuous (and the usual regularity conditions for the LLN apply to f and P_{θ_0}), then it is not hard to see that the moment estimators are consistent because

$$\hat\theta_n = e^{-1}(\hat P_n f) \xrightarrow{\text{P}} e^{-1}(P_{\theta_0} f) = e^{-1}(e(\theta_0)) = \theta_0\ .$$

If e^{-1} is also differentiable and $\hat P_n f$ is asymptotically normal, then so is $\sqrt{n}(\hat\theta_n - \theta_0)$, by the delta method.

A downside is that moment estimators often have high variance. However, they are known to be useful as initial conditions to maximum likelihood estimation; the one-step procedure above gives that some theoretical justification.

Exercise 21.2. Derive the method-of-moment estimators for $X_1, \dots, X_n$ sampled from $\mathrm{Gamma}(\alpha, \beta)$. What is the corresponding asymptotic covariance matrix from Theorem 4.1 of [van98]? How does it compare to the asymptotic covariance matrix of the corresponding one-step estimator?

References

[van98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.