STAT 460/560 Class 14: One-Step Estimators and Examples

Ben Bloem-Reddy

Reading: Chapter 5.3 and 5.7, [van98].

1. Example: 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

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

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 $m_{\theta}(x, y) = -(y - f_{\theta}(x))^2$. It should converge to the maximizer of

$$\theta \mapsto P_{\theta_0} m_{\theta} = -P_{\theta_0} (f_{\theta_0} - f_{\theta})^2 - E_{\theta_0} [\epsilon^2],$$
(14.1)

which at $\theta = \theta_0$ equals $-E_{\theta_0}[\epsilon^2]$.

Activity 14.1. Show that (14.1) holds.

As long as $\theta \neq \theta_0$ implies that $P_{\theta_0}(f_{\theta}(X) \neq f_{\theta_0}(X)) > 0$ then θ_0 is identifiable and we can ask whether Theorem 5.23 holds. Suppose that our regression model is regular at θ_0 in the sense that

$$P_{\theta_0}(f_{\theta_0} - f_{\theta})^2 = P_{\theta_0}((\theta - \theta_0)^\top \dot{f}_{\theta_0})^2 + o(\|\theta - \theta_0\|^2)$$
(14.2)

holds in a neighborhood of θ_0 .

Activity 14.2. Show that if (14.2) holds, then with $V_{\theta_0} = -2P_{\theta_0}\dot{f}_{\theta_0}\dot{f}_{\theta_0}^{\top}$ and $\dot{m}_{\theta_0}(x,y) = 2(y - f_{\theta_0}(x))^{\top}\dot{f}_{\theta_0}(x)$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}(X_i, Y_i) + o_P(1)$$

Show that therefore, if $\epsilon \perp \!\!\!\perp X$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, E_{\theta_0}[\epsilon^2] P_{\theta_0}(\dot{f}_{\theta_0}(X) \dot{f}_{\theta_0}(X)^\top)^{-1})$$

2. One-step estimators

As van der Vaart points out, the method of Z-estimators has 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 a Z-estimator $\tilde{\theta}_n$ for the estimating equation 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 8 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 $\hat{\theta}_n$ 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 $\dot{\Psi}_0$,

$$\sup_{\sqrt{n}\|\theta-\theta_0\| < M} \|\sqrt{n}(\Psi_n(\theta) - \Psi_n(\theta_0)) - \dot{\Psi}_0 \sqrt{n}(\theta-\theta_0))\| \xrightarrow{\mathbf{p}} 0.$$
(14.3)

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 (14.3) will be satisfied. With that, define the one-step estimator by

$$\hat{\theta}_n = \tilde{\theta}_n - \dot{\Psi}_{n,0}^{-1} \Psi_n(\tilde{\theta}_n) .$$
(14.4)

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 withing $\sqrt{n}^{-1/2}$ of θ_0 with high probability.

Theorem 14.1. Let $\sqrt{n}\Psi_n(\theta_0) \rightsquigarrow Z$, for some random variable Z. Suppose that (14.3) 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

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

Proof. The estimator $\dot{\Psi}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_0)$ satisfies

$$\begin{split} \dot{\Psi}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_0) &= \dot{\Psi}_{n,0}\sqrt{n}(\tilde{\theta}_n - \dot{\Psi}_{n,0}^{-1}\Psi_n(\tilde{\theta}_n) - \theta_0) \\ &= \dot{\Psi}_{n,0}\sqrt{n}(\tilde{\theta}_n - \theta_0) - \sqrt{n}(\Psi_n(\tilde{\theta}_n) - \Psi_n(\theta_0)) - \sqrt{n}\Psi_n(\theta_0) \;. \end{split}$$

The middle term, by (14.3), can by replaced by $\dot{\Psi}_0 \sqrt{n}(\tilde{\theta}_n - \theta_0) + o_P(1)$, yielding

$$\begin{split} \dot{\Psi}_{n,0}\sqrt{n}(\hat{\theta}_n - \theta_0) &= \dot{\Psi}_{n,0}\sqrt{n}(\tilde{\theta}_n - \theta_0) - \dot{\Psi}_0\sqrt{n}(\tilde{\theta}_n - \theta_0) - \sqrt{n}\Psi_n(\theta_0) + o_P(1) \\ &= \underbrace{(\dot{\Psi}_{n,0} - \dot{\Psi}_0)}_{o_P(1)}\sqrt{n}(\tilde{\theta}_n - \theta_0) - \sqrt{n}\Psi_n(\theta_0) + o_P(1) \\ &= -\sqrt{n}\Psi_n(\theta_0) + o_P(1) \;. \end{split}$$

Therefore, by Slutsky's lemma,

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



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.)

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 $e: \Theta \to \mathbb{R}^k$ is $e(\theta) - P_{\theta}f$. The **moment** estimator $\hat{\theta}_n$ satisfies

$$\hat{P}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i) = P_{\hat{\theta}_n} f = e(\hat{\theta}_n) .$$
(14.5)

Setting $f_i(x) = x^j$ 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

$$\hat{\theta}_n = e^{-1}(\hat{P}_n f) \xrightarrow{\mathbf{p}} e^{-1}(P_{\theta_0} f) = e^{-1}(e(\theta_0)) = \theta_0 .$$

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 14.1. Derive the method-of-moment estimators for X_1, \ldots, X_n sampled from Gamma (α, β) . 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.