Chapter 3

Classical regularity conditions

Preliminary draft. Please do not distribute.

The results from classical asymptotic theory typically require assumptions of pointwise differentiability of a criterion function with respect an unknown parameter. Taylor expansion about some "true" value in the parameter space then gives a quadratic approximation to the criterion function, within error terms that can be bounded using the remainder from the Taylor expansion. With appropriately small error terms, an estimator defined to minimize the criterion function will lie close to the random variable that minimizes the quadratic, a random variable that typically has a neat closed form representation. The estimator inherits the limiting behaviour of the minimizer of the quadratic.

3.1 The classical (local) regularity conditions

Classical::classical.reg

Consider once more the problem discussed heuristically in Chapter 1, but this time with the aim of proving rigorously some of the limit theory from before. The methods will be classical, a combination of smoothness and domination assumptions that together ensure good local approximaions by random quadratics. The desire for rigor compels us to pay more attention to the effect of remainder terms in Taylor expansions.

As before, start with a set of real-valued functions $\{g(x,\theta) : \theta \in \Theta\}$ indexed by a subset Θ of some Euclidean space $\mathbb{R}^k$. An estimator $\widehat{\theta}_n$ is defined by (approximate) minimization of a random criterion function,

$$G_n(\theta) = G_n(\omega,\theta) = \frac{1}{n}\sum_{i\le n} g(X_i(\omega),\theta).$$

For simplicity, begin with the assumption that the $\{X_i\}$ are independent and identically distributed under $\mathbb{P}$, each with distribution P.

Remark. It might perhaps be clearer to write $\mathbb{P}_n$ instead of $\mathbb{P}$, anticipating the situation where the marginal distributions can change with n, as will be needed for a discussion of power of tests under alternatives.

To simplify even more, et us assume that $\widehat{\theta}_n$ is consistent in probability,

assume.consistent<1>
$$\widehat{\theta}_n = \theta_0 + o_p(1) \qquad \text{under the } \mathbb{P} \text{ model.}$$

where θ_0 minimizes $G(\theta) := P^x g(x,\theta)$. Let us not carry along global assumptions as extra baggage while we study local behaviour.

Without consistency it would make little sense to be thinking about behaviour of G_n near θ_0. Only if $\widehat{\theta}_n$ has high probability of concentrating near θ_0 can we safely ignore how G_n behaves outside a neighborhood of θ_0.

Remark. There is much to recommend taking <1>, or some other *high level requirement,* as an assumption for the next step in the analysis. It reduces the clutter of notation in the statement of theorems; it helps to distinguish the roles of global and local assumptions; and it makes the proofs more "modular". For example, if someone thinks up a clever new way to establish consistency in a particular setting, we have no need to rewrite the proofs of rates of convergence or asymptotic limit behaviour. And most important of all, when we wish to modify our proofs to cover new cases—there is no such thing as the perfect or most general form of a theorem in Asymptopia, as is demonstrated, for example, by the constant stream of new and improved consistency theorems that keep journal editors busy—we will have less material to dig through.

Taylor expansion about θ_0 gives the simplest way of deriving a quadratic approximation for a smooth criterion function. The classical assumptions use two derivatives to construct the quadratic, and a third derivative (or something slightly weaker) to bound an error.

classic.reg <2> **Definition.** *Say that $\{g(x,\theta) : \theta \in \Theta\}$ satisfies the classical regularity conditions at a point θ_0 with respect to P if there is a neighborhood $\mathcal{N}$ of $\theta_0 = \operatorname{argmin}_\theta P^x g(x,\theta)$ for which:*

classical.smooth

(a) for P almost all x, the function $\theta \mapsto g(x,\theta)$ is twice differentiable in $\mathcal{N}$, with the second derivative $\ddot{g}(x,\theta)$ that is continuous at θ_0;

classical.score (b) *the components of* $\Delta(x) := \dot{g}(x, \theta_0)$ *all belong to* $\mathcal{L}^2(P)$*;*

classical.dom (c) *there exists a function* $M(x) \in \mathcal{L}^1(P)$ *for which* $\|\ddot{g}(x, \theta)\|_2 \le M(x)$ *for all* $\theta \in \mathcal{N}$*.*

Remark. *The norm* $\|\cdot\|_2$ *of a matrix* V *is defined by* $\|V\|_2 := \sup_{|u|=1} |Vu|$*. By Cauchy-Schwarz,* $|s'Vt| \le |s|\,\|V\|_2\,|t|$ *for all* s, t*. The domination assumption is equivalent to each of the* k^2 *components of the matrix* $\ddot{g}(x, \theta)$ *being dominated by some* $M(x) \in \mathcal{L}^1(P)$

classical.posdef (d) *the expected value* $J := P\ddot{g}(x, \theta_0)$ *is nonsingular.*

classical.int (e) θ_0 *is an interior point of* Θ*.*

Remark. The assumption that θ_0 is an interior point is sometimes only implicit in the literature. See Problem [2] or Section 7 for some examples of what can happen if θ_0 lies on the boundary of Θ.

For notational convenience I will express all approximations in terms of the difference $t = \theta - \theta_0$. Without some such notational trick I find my displayed equations too often overflow the line and look intimidating. Almost equivalently, I could just assume (without loss of generality) that $\theta_0 = 0$. If the notation troubles you, feel free to rewrite what follows with $\theta - \theta_0$ in place of t.

classical.quadratic <3> **Lemma.** *Under the classical regularity assumptions (a), (b), (c), and (e), the criterion function has a local approximation,*

Gn.Taylor<4>
$$G_n(\theta_0 + t) = G_n(\theta_0) + t'Z_n/\sqrt{n} + \tfrac{1}{2}t'Jt + |t|^2 R_n(t)$$

where $Z_n = \sum_{i \le n} \Delta(X_i)/\sqrt{n}$ *has a limiting* $N\big(0, P(\Delta\Delta')\big)$ *distribution and, for each deterministic sequence* $\{\delta_n\}$ *converging to zero,*

$$\sup_{|t| \le \delta_n} |R_n(t)| \to 0 \qquad \textit{in probability.}$$

Proof From Assumption (a), the function $g(x, \cdot)$ has a pointwise expansion

pwise.Taylor<5>
$$g(x, \theta_0 + t) = g(x, \theta_0) + t'\Delta(x) + \tfrac{1}{2}t'\ddot{g}(x, \theta_0)t + \tfrac{1}{2}t'r(x, t)t.$$

When t is close enough to 0, the remainder term has the representation

$$r(x, t) = \ddot{g}(x, \theta_0 + t^*) - \ddot{g}(x, \theta_0),$$

for some $\theta_0 + t^*$ on the line segment L joining θ_0 and $\theta_0 + t$.

Remark. This representation assumes that L lies wholly within Θ. If $|t|$ is small enough, we have no problem with L running outside Θ, because θ_0 is an interior point. If θ_0 were on the boundary of Θ, we would need to place further assumptions on the shape of Θ near θ_0.

Define

$$M_\delta(x) := \sup\{\|r(x,t)\|_2 : |t| \le \delta\}.$$

From the continuity and domination assumptions on $\ddot{g}$,

$$2M(x) \ge M_\delta(x) \to 0 \qquad \text{as } \delta \to 0.$$

Dominated Convergence then implies

$$P\,\|r(x,t)\|_2 \to 0 \qquad \text{as } t \to 0.$$

Integrating <5> with respect to P we get

G.Taylor<6>
$$G(\theta_0 + t) = G(\theta_0) + t'P\Delta + \tfrac{1}{2}t'Jt + o(|t|^2),$$

The coefficient $P\Delta$ of the linear term must vanish because $G(\cdot)$ has its minimum at the interior point θ_0.

Remark. If θ_0 had been on the boundary of Θ, there would be no guarantee that $P\Delta = 0$; the quadratic approximation to $G(\theta_0+t)$ might contain a nonvanishing linear term $t'P\Delta$ in the case of a boundary point. Both the rate of convergence for $\widehat{\theta}_n - \theta_0$ and the limiting distribution theory would then be affected.

For the $g(x,\theta) = -\log f(x,\theta)$ corresponding to maximum likelihood estimation, the derivative Δ becomes the score function. The equality $P\Delta = 0$ corresponds to a formal "differentiation under the integral sign" at θ_0.

The equality $P\Delta = 0$ and Assumption (b) ensure that Z_n has a limiting $N\big(0, P(\Delta\Delta')\big)$ distribution.

To get the quadratic approximation for $G_n(\theta_0 + t)$, start from <5> evaluated at each observation X_i:

$$\begin{aligned} G_n(\theta_0 + t) &= \frac{1}{n}\sum_{i\le n} g(X_i, \theta_0 + t) \\ &= \frac{1}{n}\sum_{i\le n} \big(g(X_i,\theta_0) + t\Delta(X_i) + \tfrac{1}{2}t'\ddot{g}(X_i,\theta_0)t + \tfrac{1}{2}t'r(X_i,\theta_0+t)t\big) \\ &= G_n(\theta_0) + \frac{t'Z_n}{\sqrt{n}} + \tfrac{1}{2}t'Jt + \tfrac{1}{2}|t|^2 R_n(t), \end{aligned}$$

where

$$|R_n(t)| \le \left\| \frac{1}{n} \sum\nolimits_{i \le n} \ddot{g}(X_i, \theta_0) - J \right\|_2 + \frac{1}{n} \sum\nolimits_{i \le n} \|r(X_i, \theta_0 + t)\|_2 .$$

By the (weak) law of large numbers for each component of $\ddot{g}(x, \theta_0)$, the first contribution to $R_n(t)$ converges in probability to zero, regardless of $\{\delta_n\}$ and t. Whenever $|t| \le \delta_n$, the second contribution is bounded in absolute value by $\sum_{i \le n} M_{\delta_n}(X_i)/n$, which converges in $\mathcal{L}^1(\mathbb{P})$ to zero.

□

3.2 Asymptotics via quadratic approximation

Classical::quad.comp

Once we know (or assume) that $\widehat{\theta}_n$ lies close to some θ_0 with probability tending to one, and we have a suitable quadratic approximation to the criterion function near θ_0, it takes but two comparisons to derive the limiting form for the estimator. The regularity conditions play no further role in the asymptotic arguments. The same arguments can work even when classical regularity assumptions fail, provided we are able (by whatever means) to establish consistency and construct suitable approximations to the random criterion function.

classical.CLT <7>

Theorem. *Suppose*

(i) a quadratic approximation

$$G_n(\theta_0 + t) = G_n(\theta_0) + \frac{t' Z_n}{\sqrt{n}} + \tfrac{1}{2} t' J t + o_p(|t|^2)$$

holds uniformly in $o_p(1)$ neighborhoods of $t = 0$, with J a positive definite matrix and Z_n a random vector of order $O_p(1)$.

(ii) θ_0 is an interior point of Θ

(iii) $\widehat{\theta}_n \to \theta_0$ in probability

(iv) $\widehat{\theta}_n$ comes within $o_p(1/n)$ of minimizing G_n.

Then $\widehat{\theta}_n = \theta_0 - J^{-1} Z_n/\sqrt{n} + o_p(1/\sqrt{n})$. If Z_n has a limiting $N(0, D)$ distribution then $\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow N(0, J^{-1} D J)$.

The proof breaks naturally into two steps, which I will state as two separate lemmas. First we must prove that $\widehat{\theta}_n$ lies within $O_p(1/\sqrt{n})$ of θ_0, using a

comparison between the values of G_n at $\widehat{\theta}_n$ and at θ_0. Assumption (e) plays no role in the first step. Within $O_p(1/\sqrt{n})$ neighborhoods of θ_0 the quadratic approximation reduces to a simpler form, which suggests the second-stage comparison between the values of G_n at $\widehat{\theta}_n$ and at a random θ^* that also lies within $O_p(1/\sqrt{n})$ of θ_0. Assumption (e) is needed at this step to ensure that, with probability tending to one, θ^* is a well defined point of Θ.

Both parts of the argument are variations on a single comparison technique, which is worth isolating as a deterministic result.

det.quad <8>

Lemma. *Suppose f is a real-valued function on a set T and $\kappa : T \to \mathbb{R}^+$. If t^* and t_0 are points of T such that, for some nonnegative constants ϵ, γ and η,*

lower.quad

(i) $f(t) \geq f(t_0) - \gamma - \eta\kappa(t) + \kappa(t)^2$ *for all* t

eps.max

(ii) $f(t^*) \leq f(t_0) + \epsilon$

then $\kappa(t^*) \leq \eta + \sqrt{\epsilon + \gamma}$.

PROOF Temporarily abbreviate $\kappa(t^*)$ to κ. From (i) evaluted at $t = t^*$,

$$f(t^*) - f(t_0) \geq (\kappa - \eta/2)^2 - \gamma - \eta^2/4$$

From (ii), the left-hand side is less than ϵ. Thus

$$\sqrt{\epsilon + \gamma + \eta^2/4} \geq |\kappa - \eta/2|.$$

It follows that either $\kappa \leq \eta/2$ or

$$\eta/2 + \sqrt{\epsilon + \gamma + \eta^2/4} \geq \kappa.$$

The general inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, for $a, b \geq 0$, then leads to the asserted bound for κ.

□

To be continued