19

Estimating Equations and Maximum Likelihood asymptotics

Here we give a rigorous account of the consistency and asymptotic normality for certain solutions of estimating equations, of which least squares and maximum likelihood estimation are special cases. Though the material and the proof in particular are technical, it is worthwhile to understand the conditions under which these types of estimators have such desired properties, and how those conditions can be verified, as is done in the examples that follow.

Let be given $n \in \mathbb{N}$ and a set χ, a random vector $\mathbf{X} \in \chi^n$, a parameter set $\Theta \subset \mathbb{R}^p$ with non-empty interior and a function $\mathcal{U}_n : \chi^n \times \Theta \rightarrow \mathbb{R}^p$. We consider the estimating equation

$$\mathcal{U}_n(\mathbf{X}, \theta) = 0, \quad \theta \in \Theta. \tag{19.1}$$

For least squares estimation, say pairs $(\mathbf{X}_i, Y_i), i = 1, \ldots, n$ with distribution depending on θ are observed for which

$$E[Y_i|\mathbf{X}_i] = f_i(\mathbf{X}_i; \theta)$$

for $f_i(\mathbf{x}; \theta)$ in some parametric class of functions. The least squares estimate of θ is given as the minimizer of

$$J(\theta; \mathbf{X}) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f_i(\mathbf{X}_i; \theta)\right)^2,$$

which under smoothness conditions can be obtained via (19.1) with

$$\mathcal{U}_n(\mathbf{x},\theta) = \partial_\theta J(\theta;\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \left(f_i(\mathbf{x}_i;\theta) - y_i\right)\partial_\theta f_i(\mathbf{x}_i;\theta). \tag{19.2}$$

In the following, ∂_θ as in (19.2) applied to a real valued function depending on a vector parameter θ returns a gradient vector, and likewise ∂_θ^2 returns a matrix of second partial derivatives. Further, functions appearing here in connection with estimating equations may notationally appear to depend only on the argument θ.

For maximum likelihood, under smoothness conditions on the density $p(\mathbf{x};\theta)$ of $\mathbf{X}$, the maximizer of the log likelihood $\mathcal{L}_n(\theta;\mathbf{x}) = \log p(\mathbf{x};\theta)$ is given as a solution to (19.1) for

$$\mathcal{U}_n(\mathbf{X},\theta) = \partial_\theta \mathcal{L}_n(\theta;\mathbf{X}).$$

In typical cases, one observes independent vectors $\mathbf{X}_i$ for $i = 1,\ldots,n$ which have distributions $p_i(\mathbf{x}_i;\theta)$ from given parametric families. Hence, the collection $\mathbf{X}$ of these vectors has density, or likelihood, given by the product

$$p(\mathbf{x};\theta) = \prod_{i=1}^n p_i(\mathbf{x}_i,\theta) \quad \text{with log likelihood} \quad \mathcal{L}_n(\theta) = \log p(\mathbf{x};\theta).$$

Under smoothness, the maximizer can be found by setting the derivative of the logarithm $\mathcal{L}_n(\theta)$ to zero, resulting in the estimating function

$$\mathcal{U}_n(\mathbf{x};\theta) = \sum_{i=1}^n \partial_\theta \log p_i(\mathbf{x}_i,\theta).$$

The aim of the estimating equation $\mathcal{U}_n(\mathbf{X},\theta) = 0$ is to provide a value close to the one where the function $\mathcal{U}_n(\mathbf{X},\theta)$, written also as $\mathcal{U}_n(\theta)$ for short, takes the value of 0 in some expected, or asymptotic sense. In particular, in Theorem 19.0.1 we will show that the roots of the estimating equation lie close to the value $\theta_0 \in \Theta$ for which the function $\mathcal{U}_n(\theta)$, with appropriate scaling, is zero as $n \to \infty$, or, more precisely for which there exists a sequence of real numbers a_n for which $a_n\mathcal{U}_n(\theta_0) \to_p 0$, see Condition (19.4). In Theorem 19.0.2, we will also provide a corresponding limiting distribution result for consistent solutions to the estimating equation (19.1). Let $\mathcal{U}_n(\mathbf{X},\theta)$ have components

$$\mathcal{U}_n(\mathbf{X}_n,\theta) = (\mathcal{U}_{n,j}(\mathbf{X}_n,\theta))_{1\le j\le p} \quad \text{where} \quad \mathcal{U}_{n,j} : \mathbb{R}^n \times \Theta \to \mathbb{R}.$$

In the case of maximum likelihood estimation, where the function $\mathcal{U}_n(\theta)$ is given by the derivative of the log likelihood $\mathcal{L}_n(\theta)$, and under the assumption of the existence and continuity of second derivatives in θ, by the equality of the mixed partial derivatives, we have

$$\frac{\partial \mathcal{U}_{n,j}(\theta)}{\partial\theta_a} = \frac{\partial^2\mathcal{L}_n(\theta)}{\partial\theta_a\partial\theta_j} = \frac{\partial^2\mathcal{L}_n(\theta)}{\partial\theta_j\partial\theta_a} = \frac{\partial \mathcal{U}_{n,a}(\theta)}{\partial\theta_j},$$

that is, $\mathcal{U}_n'(\theta)$ is the symmetric observed information matrix, and condition (19.5) below is equivalent to the condition that the limiting information matrix is positive definite.

As in general $\partial U_{n,j}(\theta)/\partial\theta_a$ is not necessarily equal to $\partial U_{n,a}(\theta)/\partial\theta_j$ the limiting matrix Γ in (19.4) may not be symmetric. In general condition (19.5) below is equivalent to the condition that $\Gamma + \Gamma^{\mathsf{T}}$ is positive definite. We let $|\cdot|$ denote the Euclidean norm of a vector in $\mathbb{R}^p$, and also the operator norm of a matrix in $\mathbb{R}^{p\times p}$. We also let ∂_k denote the result of taking the partial with respect to the k^{th} coordinate, and use similar notation for higher order derivatives. Further, for $\theta \in \Theta$ let j, k entry of

$$\mathcal{U}_n'(\theta) \in \mathbb{R}^{p\times p} \quad \text{be given by} \quad (\mathcal{U}_n'(\theta)_{j,k} = \partial_k \mathcal{U}_{n,j}(\theta)).$$

Over each coordinate $j = 1, \ldots, p$ we will make use of the second order Talyor expansion of $\mathcal{U}_{n,j}(\theta)$ about zero,

$$\mathcal{U}_{n,j}(\theta) = \mathcal{U}_{n,j}(0) + \sum_{k=1}^{p} \partial_k \mathcal{U}_{n,j}(0)\theta_k + \frac{1}{2}\sum_{1\le k,l\le p} \theta_k \partial_{k,l}\mathcal{U}_{n,j}(\theta_{n,j}^*)\theta_l, \quad (19.3)$$

where $\theta_{n,j}^*$ lies on the line segment connecting θ and 0.

Theorem 19.0.1 *Suppose that there exists $\theta_0 \in \Theta$, a sequence of real numbers a_n, and a matrix $\Gamma \in \mathbf{R}^{p\times p}$ such that*

$$a_n \mathcal{U}_n(\theta_0) \to_p 0 \quad \text{and} \quad a_n \mathcal{U}_n'(\theta_0) \to_p \Gamma, \quad (19.4)$$

and that $\mathcal{U}_n(\theta)$ is twice continuously differentiable in an open set $\Theta_0 \subset \Theta$ containing θ_0. Assume that for some $\gamma > 0$ that Γ satisfies

$$\inf_{|\theta|=1} \theta^{\mathsf{T}}\Gamma\theta = \gamma. \quad (19.5)$$

Further, for any $\eta \in (0,1)$, suppose there exists a K such that for all n sufficiently large,

$$P(|a_n \partial_{k,l}\mathcal{U}_{n,j}(\boldsymbol{\theta})| \le K, 1 \le k,l,j \le p, \theta \in \Theta_0) \quad \ge \quad 1-\eta. \quad (19.6)$$

Then for any given $\epsilon > 0$ and $\eta \in (0,1)$, for all n sufficiently large, with probability at least $1-\eta$ there exists $\widehat{\theta}_n \in \Theta$ satisfying $\mathcal{U}_n(\widehat{\theta}_n) = 0$ and $|\widehat{\theta}_n - \theta_0| \le \epsilon$. Thus, there exists a sequence of roots to the estimating equation (19.1) consistent for θ_0.

In addition, for any sequence $\widehat{\theta}_n \to_p \theta_0$, we have

$$a_n \mathcal{U}_n'(\widehat{\theta}_n) \to_p \Gamma, \quad (19.7)$$

that is, Γ can be consistently estimated by $a_n\mathcal{U}_n'(\widehat{\theta}_n)$ from any sequence consistent for θ_0.

Proof: By replacing $\mathcal{U}_n$ by $a_n\mathcal{U}_n$ and θ by $\theta - \theta_0$, we may assume that the conditions of Theorem 19.0.1 hold with $a_n = 1$ and $\theta_0 = 0$, and so (19.3) is in force. For $\delta > 0$ let

$$B_\delta = \{\theta : |\theta| \le \delta\}.$$

For the given $\eta \in (0,1)$, let K and n_0 be such that (19.6) holds with η replaced by $\eta/2$ for $n \ge n_0$. For the given $\epsilon > 0$, take $\delta \in (0, \epsilon)$ such that

$$B_\delta \subset \Theta_0 \quad \text{and} \quad C\delta < \gamma \quad \text{where} \quad C = 2 + \frac{1}{2}K.$$

Now by (19.4) there exists $n_1 \ge n_0$ such that for $n \ge n_1$, with probability for each event below at least $1 - \eta/2$,

$$|\mathcal{U}_n(0)| < \delta^2 \quad \text{and} \quad |\mathcal{U}_n'(0) - \Gamma| < \delta. \tag{19.8}$$

Let $R_n = (R_{n,1}, \ldots, R_{n,p})^T$ have components

$$R_{n,j} = \sum_{1 \le k,l \le p} \theta_k \partial_{k,l} \mathcal{U}_{n,j}(\theta^*_{n,j})\theta_l.$$

Then, for $n \ge n_1$ and $\theta \in B_\delta$, with probability at least $1 - \eta$, from (19.3), (19.8) and (19.6),

$$\begin{aligned} |\mathcal{U}_n(\theta) - \Gamma\theta| &\le |\mathcal{U}_n(\theta) - \mathcal{U}_n'(0)\theta| + |\mathcal{U}_n'(0)\theta - \Gamma\theta| \\ &= |\mathcal{U}_n(0) + \frac{1}{2}R_n| + |(\mathcal{U}_n'(0) - \Gamma)\theta| \\ &< \delta^2 + \frac{1}{2}K|\theta|^2 + \delta|\theta| \le C\delta^2, \end{aligned}$$

so

$$|\theta^{\mathsf{T}}\mathcal{U}_n(\theta) - \theta^{\mathsf{T}}\Gamma\theta| < C\delta^3.$$

Hence, if $|\theta| = \delta$,

$$\theta^{\mathsf{T}}\mathcal{U}_n(\theta) > \theta^{\mathsf{T}}\Gamma\theta - C\delta^3 \ge \gamma\delta^2 - C\delta^3 = \delta^2(\gamma - C\delta) > 0.$$

Now we argue as in Lemma 2 of Aitchison, John, and S. D. Silvey. "Maximum-likelihood estimation of parameters subject to restraints," Annals of Mathematical Statistics (1958): 813-828. Assume for the sake of contradiction that $\mathcal{U}_n(\theta)$ does not have a root in B_δ. Then for $\theta \in B_\delta$, the function $f(\theta) = -\delta\mathcal{U}_n(\theta)/|\mathcal{U}_n(\theta)|$ continuously maps B_δ to itself. By the Brouwer fixed point theorem, there exists $\vartheta \in B_\delta$, with $f(\vartheta) = \vartheta$. Since $|f(\theta)| = \delta$ for all $\theta \in B_\delta$, we have $|f(\vartheta)| = |\vartheta| = \delta$, which gives the contradiction $\delta^2 = |\vartheta|^2 = \vartheta^{\mathsf{T}}\vartheta = \vartheta^{\mathsf{T}} f(\vartheta) < 0$. Hence $\mathcal{U}_n(\theta)$ has a root within δ of 0, and since $\delta < \epsilon$, therefore within ϵ, as required.

To prove (19.7), a first order Talyor expansion yields

$$\begin{aligned} \partial_k \mathcal{U}_{n,j}(\widehat{\theta}_n) &= \partial_k \mathcal{U}_{n,j}(0) + \sum_{l=1}^{p} \partial_{k,l} \mathcal{U}_{n,j}(\theta^*_{n,j})\widehat{\theta}_{n,l} \\ &:= \partial_k \mathcal{U}_{n,j}(0) + Q^{\mathsf{T}}_{n,k,j}\widehat{\theta}_n \end{aligned}$$

where $Q^{\mathsf{T}}_{n,k,j} = (\partial_{k,1}\mathcal{U}_{n,j}(\theta^*_{n,j}), \ldots, \partial_{k,p}\mathcal{U}_{n,j}(\theta^*_{n,j}))$ and $\theta^*_{n,j}$ lies along the line segment connecting $\widehat{\theta}_n$ and 0. Writing this identity out in vector notation, we have

$$\mathcal{U}'_n(\widehat{\theta}) - \mathcal{U}'_n(0) = Q_n \quad \text{where} \quad (Q_n)_{k,j} = Q^{\mathsf{T}}_{n,k,j}\widehat{\theta}_n.$$

Let $\eta \in (0,1)$ and $\epsilon > 0$ be given, choose $\delta \in (0, \epsilon/Kp^{3/2})$ so that $B_\delta \subset \Theta_0$, and let K and n_2 be such that for all $n \geq n_2$, with probability at least $1-\eta$, $|\partial_{k,l}\mathcal{U}_n(\theta)| \leq K$ for all $1 \leq k, l \leq p$ and $|\widehat{\theta}_n| \leq \delta$. Then, for $n \geq n_2$ with probability at least $1-\eta$. each entry of the matrix Q_n satisfies the inequality

$$|Q^T_{n,k,j}\widehat{\theta}_n| \leq K\sqrt{p}\delta \leq \epsilon/p.$$

It follows that

$$|\mathcal{U}'_n(\widehat{\theta}) - \mathcal{U}'_n(0)| = |Q_n| \leq \epsilon.$$

The claim follows, since ϵ and η are arbitrary, and $\mathcal{U}'_n(0) \to_p \Gamma$ by assumption. ■

Theorem 19.0.2 *Suppose the sequence of solutions $\widehat{\theta}_n$ to (19.1) is consistent for θ_0, let a_n be a sequence of real numbers for which the second condition of (19.4) and (19.6) hold, assume that the matrix Γ is non-singular and that $\mathcal{U}_n(\theta)$ is twice differentiable in an open set $\Theta_0 \subset \Theta$ containing θ_0. Further, let b_n be a sequence of real numbers such that for some random variable Y,*

$$b_n\mathcal{U}_n(\theta_0) \quad \to_d \quad Y. \tag{19.9}$$

Then

$$\frac{b_n}{a_n}(\widehat{\theta}_n - \theta_0) \to_d -\Gamma^{-1}Y.$$

Proof: As in the proof of Theorem 19.0.1, without loss of generality take $a_n = 1$, and $\theta_0 = 0$. Since the limit in distribution does not depend on events of vanishingly small probability, we may assume that for all n sufficiently large $\widehat{\theta}_n \in \Theta_0$, and $|\partial_{k,j}\mathcal{U}_n(\theta)| \leq K$ for all $1 \leq j, k \leq p$ and $\theta \in \Theta_0$ for some K. For such n the expansion (19.3) holds, and substituting $\widehat{\theta}_n$ for θ and using $\mathcal{U}_n(\widehat{\theta}_n) = 0$ yields

$$-\mathcal{U}_n(0) = (\mathcal{U}'_n(0) + \epsilon_n)\widehat{\theta}_n = \Gamma_n\widehat{\theta}_n$$

where ϵ_n is the matrix with components

$$(\epsilon_n)_{j,l} = \frac{1}{2}\sum_k \widehat{\theta}_{n,k}\partial_{k,l}\mathcal{U}_{n,j}(\theta^*_{n,j}) \quad \text{and} \quad \Gamma_n = \mathcal{U}'_n(0) + \epsilon_n.$$

Since, by the Cauchy-Schwarz inequality,

$$|(\epsilon_n)_{j,l}| \leq \frac{\sqrt{p}}{2}K|\widehat{\theta}_n| \to_p 0,$$

we have $\Gamma_n \to_p \Gamma$ so that Γ_n^{-1} exists with probability tending to 1, and converges in probability to Γ^{-1}; set $\Gamma^{-1} = 0$ arbitrarily when Γ is singular. Now using (19.9) and Slutsky's theorem, on an event of probability tending to one as n tends to infinity,

$$b_n \widehat{\theta}_n = \Gamma_n^{-1} \left(b_n \Gamma_n \widehat{\theta}_n \right) = -\Gamma_n^{-1} \left(b_n \mathcal{U}_n(0) \right) \to_d -\Gamma^{-1} Y. \qquad \blacksquare$$

In the independent case, and in one dimension, distributional convergence is shown by applying the Central Limit Theorem to a sum

$$S_n = \sum_{i=1}^n X_i$$

of independent random variables with $\text{Var}(X_i) = \sigma_i^2$, and $s_n^2 = \sum_{i=1}^n \sigma_i^2$. In this case,

$$\frac{S_n - E[S_n]}{s_n} \to_d \mathcal{N}(0,1)$$

when the Lindeberg condition is satisfied, that is, when

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n E\left[(X_i - E[X_i])^2 \mathbf{1}(|X_i - E[X_i]| > \epsilon s_n) \right] = 0. \tag{19.10}$$

This condition holds, generally speaking, when all the summands contribute the same order to the total variance. It is a bit unwieldy to check, but one can obtain a stronger condition that implies (19.10) and that is easier to verify. Indeed, by Hölder's inequality with $p = 3/2, q = 3$, followed by Markov's inequality, we obtain the following bound on the summmands,

$$\begin{aligned} &E\left[(X_i - E[X_i])^2 \mathbf{1}(|X_i - E[X_i]| > \epsilon s_n) \right] \\ &\qquad \le \left(E|X_i - E[X_i]|^3 \right)^{2/3} \left(P\left(|X_i - E[X_i]| > \epsilon s_n \right) \right)^{1/3} \\ &\quad \le \left(E|X_i - E[X_i]|^3 \right)^{2/3} \left(\frac{E|X_i - E[X_i]|^3}{\epsilon^3 s_n^3} \right)^{1/3} = \frac{E|X_i - E[X_i]|^3}{\epsilon s_n}. \end{aligned}$$

Hence (19.10) is satisfied if

$$\lim_{n\to\infty} \frac{1}{s_n^3} \sum_{i=1}^n E|X_i - E[X_i]|^3 = 0. \tag{19.11}$$

One of the main applications of Theorems 19.0.1 and 19.0.2 is to maximum likelihood estimators, and to verify the conditions of these theorems for that setting it helps to have at hand ways to confirm that differentiation under the integral is allowed. In particular, the dominated convergence theorem says that if $f_n \to f$ pointwise, and $|f_n| \le g$ almost everywhere on an arbitrary measure space $(\chi, \mathcal{M}, \mu)$, and $\int g < \infty$, then

$$\int f_n \to \int f. \tag{19.12}$$

We can apply the dominated convergence theorem to derive the following lemma to verify that integration and differentiation with respect to $\theta \in \mathcal{B} \subset \mathbf{R}^p$ can be interchanged.

Lemma 19.0.1 *Let the function $f : \mathbf{R}^n \times \mathcal{B} \to \mathbf{R}$ be differentiable with respect to θ in an open set $\mathcal{B}_0 \subset \mathcal{B}$, and suppose that there exists $g : \mathbf{R}^n \to \mathbf{R}$ such that*

$$\int g(\mathbf{x})d\mathbf{x} < \infty,$$

and for all $\theta \in \mathcal{B}_0$,

$$|\frac{\partial}{\partial\theta} f(\mathbf{x};\theta)| \leq g(\mathbf{x}).$$

Then for all $\theta \in \mathcal{B}_0$,

$$\frac{\partial}{\partial\theta}\int f(\mathbf{x};\theta)d\mathbf{x} = \int \frac{\partial}{\partial\theta} f(\mathbf{x};\theta)d\mathbf{x}. \tag{19.13}$$

Proof: *Since (19.13) is true if and only if it is true componentwise, it suffices to consider a real valued function f. For $\beta \in \mathcal{B}_0$, take any $\theta_n \to \theta$. Then for all n sufficiently large β_n lies in an open ball B_0 centered at β. For such n, let*

$$f_n(\mathbf{x};\theta) = \frac{f(\mathbf{x};\theta_n) - f(\mathbf{x};\theta)}{\theta_n - \theta}.$$

By the mean value theorem, the ratio equals $\partial f(\mathbf{x};\beta_n^*)/\partial\beta$ for some β_n^* on the line segment connecting β and β_n, therefore lying in the ball B_0. Hence for all $\beta \in \mathcal{B}_0$ and large n,

$$|f_n(\mathbf{x};\theta)| \leq g(\mathbf{x}),$$

and therefore (19.13) holds by (19.12), since $f(\mathbf{x};\beta)$ is given by the limit $\lim_{n\to\infty} f_n(\mathbf{x};\beta) = \partial f(\mathbf{x};\beta)/\partial\beta$.

When f and g satisfy the conditions of Lemma 19.0.1 we say that f' is L^1 dominated in $\mathcal{B}_0$.

We now focus on the case where we have n observations from the model of the form

$$y_i = f_i(\mathbf{x}_i;\theta_0) + \epsilon_i, \tag{19.14}$$

where $\epsilon_i, i = 1, \ldots, n$ are independent, mean zero random variables with finite variance, and θ is estimated via least squares. From (19.2) we have

$$\mathcal{U}_n(\mathbf{x},\theta) = \frac{1}{n}\sum_{i=1}^{n}(f_i(x_i;\theta) - y_i)\,\partial_\theta f_i(x_i;\theta). \tag{19.15}$$

In this case, we have

$$\begin{aligned}\mathcal{U}_n(\mathbf{x},\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left(f_i(x_i;\theta) - f_i(x_i;\theta_0) - \epsilon_i\right)\partial_\theta f_i(x_i;\theta)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(f_i(x_i;\theta) - f_i(x_i;\theta_0)\right)\partial_\theta f_i(x_i;\theta)) - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\partial_\theta f_i(x_i;\theta). \qquad (19.16)\end{aligned}$$

In particular,

$$\mathcal{U}_n(\mathbf{x},\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\partial_\theta f_i(x_i;\theta_0).$$

Here is a toy example to illustrate the theorems and the conditions needed to invoke them, in the problem of the estimation of an unknown slope.

Example 19.0.1 *Consider the following one dimensional case of the linear model, where*

$$f_i(x;\theta) = \theta x \quad \textit{and we observe} \quad y_i = \theta_0 x_i + \epsilon_i, \quad i = 1,\ldots,n.$$

Suppose that $\epsilon_i, i = 1,2,\ldots$ *have mean zero, variance* σ^2 *and are uncorrelated, and that for some* $\gamma > 0$

$$\frac{1}{n}\sum_{i=1}^{n}x_i^2 \to \gamma. \qquad (19.17)$$

From (19.16) *we have*

$$\mathcal{U}_n(\mathbf{x},\theta) = \frac{1}{n}\sum_{i=1}^{n}(\theta x_i - \theta_0 x_i - \epsilon_i)x_i = (\theta-\theta_0)\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i x_i$$

In this example, though we may directly find that the solution to the estimating equation is given by

$$\widehat{\theta}_n = \theta_0 + \frac{\sum_{i=1}^{n}\epsilon_i x_i}{\sum_{i=1}^{n}x_i^2},$$

which is easily seen to be consistent and asymptotically normal under simple assumptions on the sequence $x_1,\ldots,x_n$ *and the errors* $\epsilon_1,\ldots,\epsilon_n$*. We (neverthless) explore the verification of the conditions of the theorem.*

For consistency, to verify the first condition in (19.4)*, we take* $\theta = \theta_0$ *and* $a_n = 1$*, and obtain*

$$\mathcal{U}_n(\mathbf{x},\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\epsilon_i x_i \quad \textit{and} \quad \mathrm{Var}\left(-\frac{1}{n}\sum_{i=1}^{n}\epsilon_i x_i\right) = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}x_i^2 \to 0,$$

thus showing that the first condition in (19.4) *is satisfied. For the second condition there, we have*

$$\mathcal{U}_n'(\theta) = \frac{1}{n}\sum_{i=1}^n x_i^2 \to \gamma,$$

yielding (19.5). *The second derivative condition* (19.6) *is satisfied trivially, as* $\mathcal{U}''(\theta) = 0$.

For Theorem 19.0.2, assuming now that the error terms are independent, letting $b_n = \sqrt{n}$ *we have*

$$\sqrt{n}\mathcal{U}(\theta_0) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i x_i \to \mathcal{N}(0, \sigma^2\gamma^2)$$

when Lindeberg condition, or the stronger condition (19.11) *derived above, holds for the case at hand Let* $X_i = x_i\epsilon_i$, *and suppose that* ϵ_i *are mean zero with constant variance* σ^2 *and uniformly bounded absolute third moment* $E|\epsilon|^3 \le \tau^3$. *Then*

$$E[X_i] = 0, \quad \sigma_i^2 = \sigma^2 x_i^2 \quad \text{and} \quad E|X_i|^3 \le \tau^3|x_i|^3,$$

and condition (19.11) *is satisfied when*

$$\lim_{n\to\infty} \frac{\tau^3\sum_{i=1}^n |x_i|^3}{\sigma^3\left(\sum_{i=1}^n x_i^2\right)^{3/2}} = 0.$$

If there exists $\psi \in \mathbb{R}$ *such that*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n |x_i|^3 \to \psi \quad \text{then}$$

$$\lim_{n\to\infty} \frac{\tau^3\sum_{i=1}^n |x_i|^3}{\sigma^3\left(\sum_{i=1}^n x_i^2\right)^{3/2}} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{\tau^3\frac{1}{n}\sum_{i=1}^n |x_i|^3}{\sigma^3\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)^{3/2}} = 0,$$

as desired. In particular, the condition is (trivially) satisfied when $x_i = c$, *a constant, or, say, when* $x_i = i/n, i = 1, \ldots, n$. *The conditions that* ϵ_i *have constant variance can also be somewhat relaxed.*

The next example takes on the situation where the observed value for an individual i is some function of $\mathbf{x}_i$, known to lie in a parametric class of functions, and additive noise.

Example 19.0.2 *Suppose we observe*

$$y_i = f(\mathbf{x}_i, \theta_0) + \epsilon_i \quad i = 1, \ldots, n$$

and estimate θ_0 *via least squares, minimizing*

$$J_n(\theta) = \frac{1}{2n}\sum_{i=1}^n \left(f(\mathbf{x}_i, \theta) - y_i\right)^2 = \frac{1}{2n}\sum_{i=1}^n \left(f(\mathbf{x}_i, \theta) - f(\mathbf{x}_i, \theta_0) - \epsilon_i\right)^2.$$

Taking derivative with respect to θ, we obtain the estimating equation $\mathcal{U}_n(\theta) = 0$ where

$$\mathcal{U}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{x}_i,\theta) - f(\mathbf{x}_i,\theta_0) - \epsilon_i\right)\partial_\theta f(\mathbf{x}_i,\theta).$$

We have

$$\mathcal{U}_n'(\theta) = \frac{1}{n}\sum_{i=1}^{n}\Big((\partial_\theta f(\mathbf{x}_i,\theta))^2 + \left(f(\mathbf{x}_i,\theta) - f(\mathbf{x}_i,\theta_0) - \epsilon_i\right)\partial_\theta f(\mathbf{x}_i,\theta)\Big),$$

so in particular,

$$\mathcal{U}_n(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\partial_\theta f(\mathbf{x}_i,\theta_0)\epsilon_i \quad \textit{and}$$

$$\mathcal{U}_n'(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\left(\partial_\theta f(\mathbf{x}_i,\theta_0)\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\partial_\theta f(\mathbf{x}_i,\theta_0).$$

The two conditions in (19.4)*, and condition* (19.5) *will be satisfied, as in Example* (19.0.1)*, and under the conditions on the errors as there, when there exists $\gamma > 0$ such that*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\partial_\theta f(\mathbf{x}_i,\theta_0))^2 = \gamma.$$

When the limit of this sum exists, it will be positive under the condition, say, that when x is chosen according to some distribution, then for some $\tau > 0$ we have $P(|\partial_\theta f(X,\theta)| \geq \tau) > 0$. Similarly, if x_i are deterministic, then the limit will be positive when there exists a $\tau > 0$ and a set χ such that

$$\inf_{x\in\chi}|\partial_\theta f(X,\theta)| \geq \tau \quad \textit{and} \quad \liminf_{n\to\infty}\frac{1}{n}|\{i : x_i \in \chi\}| > 0.$$

Returning to Example (19.0.1)*, for the case where x is chosen according to a distribution, when $f(x,\theta) = \theta x$ then $\partial_\theta f(x,\theta) = x$ and the condition is equivalent to $P(|x| = 0) < 1$.*

The reader may verify that Condition (19.6) *holds under the foregoing assumptions on the noise when $\partial_\theta^2 f(\mathbf{x},\theta)$ is uniformly bounded over an open set containing θ_0, now altogether yielding the existence of a consistent sequence of roots to the estimating equation.*

We leave the reader to explore sets of conditions on $f(\mathbf{x},\theta)$ under which the hypotheses of Theorem (19.0.2) *are satisfied, in particular, the Lindeberg Condition for the guarantee of asymptotic normality.*

Example 19.0.3 *Maximum Likelihood Estimation in one dimension based on an i.i.d. sample. Let the data $\mathbf{X}_n$ be composed of independent observations $X_1,\ldots,X_n$, each having density $p(x;\theta_0), \theta_0 \in \Theta \subset \mathbf{R}^p$. Let X denote a variable with the common observation distribution.*

We first address the question of consistency of the MLE. Assume that in some open set $\Theta_0 \subset \Theta$ *containing* θ_0 *the density* $p(x, \theta)$ *is twice differentiable with respect to* θ *and that* $p'(x, \theta)$ *and* $p''(x, \theta)$ *are* L^1 *dominated, that* $\mathcal{U}(X, \theta_0)$ *is a non-degenerate random variable with finite variance, and there exists a function* $R(x)$ *such that*

$$|\mathcal{U}''(X, \theta)| \le R(X) \qquad \text{for all } \theta \in \Theta_0 \text{ and } \quad E[R(X)] < \infty. \tag{19.18}$$

We show that the conditions of Theorem 19.0.1 hold with $a_n = 1/n$ *and* $\Gamma = I_X(\theta_0)$*, which we denote by* I_0 *for short.*

Since $p'(x, \theta)$ *is* L^1 *dominated in a neighborhood of* θ_0*, Proposition 19.0.1 with* $f(\mathbf{x}; \theta) = p(\mathbf{x}; \theta)$ *at* θ_0 *justifies the interchange of differentiation and integration, and hence that*

$$\begin{aligned} E[\mathcal{U}(X; \theta_0)] = E[\partial_\theta \log p(x; \theta_0)] &= \int \partial_\theta p(x; \theta_0) \\ &= \partial_\theta \int p(x, \theta) dx = \partial_\theta 1 = 0. \end{aligned}$$

Since the score function

$$\mathcal{U}_n(\mathbf{X}_n, \theta_0) = \sum_{i=1}^n \partial_\theta \log p(X_i; \theta_0)$$

is therefore the sum of mean zero i.i.d random variables, the first condition in (19.4) is satisfied with $a_n = 1/n$ *by the law of large numbers.*

Since the second derivative of $p(x, \theta)$ *is also dominated, the interchanges allowed by Proposition 19.0.1 show that the information 'matrix'* I_0*, here of dimension* 1×1*, can be obtained either as the variance of the score function* $\mathcal{U}(X, \theta)$ *or as the negative of the expected value of* $\mathcal{U}'(X, \theta)$*. It must be nonzero, by our assumption of non-degeneracy. Since* $\mathcal{U}_n'$ *is the sum of i.i.d. variables all with mean* I*, the second condition in (19.4) is satisfied with* $a_n = 1/n$ *by the law of large numbers with* $\Gamma = I_0$*.*

It remains to verify (19.6). Since $R(X)$ *is integrable, for given* $\eta \in (0, 1)$ *take* K *we may take* K *so large that* $E[R(X)]/K \le \eta$*. Now, by (19.18) and the Markov inequality*

$$\begin{aligned} P(\frac{1}{n}\sum_{i=1}^n |\mathcal{U}''(X_i, \theta)| \ge K, \theta \in \Theta_0) &\le P(\frac{1}{n}\sum_{i=1}^n R(X_i) \ge K) \\ &\le \frac{ER(X)}{K} \le \eta. \end{aligned}$$

The hypotheses of Theorem 19.0.1 are satisfied, therefore there exists a consistent sequence of roots to the likelihood equation.

To invoke Theorem 19.0.2 it is necessary only to verify (19.9). The classical CLT for i.i.d. variables gives

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathcal{U}_n(X_i, \theta_0) \to_d Y \quad \text{where} \quad Y \sim \mathcal{N}(0, I_0),$$

so that (19.9) is satisfied with $b_n = n^{-1/2}$. *Hence, since* $b_n/a_n = \sqrt{n}$ *and* $I_0^{-1}Y \sim \mathcal{N}(0, I_0^{-1})$, *Theorem 19.0.2 yields that for a consistent sequence of roots* $\widehat{\theta}_n$ *for* θ_0,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}(0, I_0^{-1}).$$

We consider the application of the conditions above to the estimation of the parameter $\theta_0 \in (0, \infty)$ *from an i.i.d. sample having the exponential distribution*

$$p(x;\theta_0) = \theta_0 \exp(-\theta_0 x), \quad x > 0.$$

Differentiating,

$$p'(x;\theta) = (1 - \theta x)\exp(-\theta x) \quad \textit{and} \quad p''(x;\theta) = (\theta x^2 - 2x)\exp(-\theta x).$$

Let $0 < \theta_L < \theta_0 < \theta_U < \infty$. *Then for all* $\theta \in \Theta_0 = (\theta_L, \theta_U)$,

$$|p'(X;\theta)| \le (1 + \theta_U X)\exp(-\theta_L X)$$

and

$$|p''(X;\theta)| \le (2X + \theta_U X^2)\exp(-\theta_L X),$$

so that p' *and* p'' *are* L^1 *dominated.*

The score function and its first two derivatives at θ_0 *are*

$$\mathcal{U}(x, \theta_0) = 1 - \theta_0 x, \quad \mathcal{U}'(x, \theta_0) = -x, \quad \textit{and} \quad \mathcal{U}''(x, \theta_0) = 0,$$

so that $\mathcal{U}'(X, \theta_0)$ *is a non-degenerate random variable with finite variance* θ_0, *and (19.18) is trivially satisfied.*

Hence, there exists a consistent sequence of roots $\widehat{\theta}_n$ *to the estimating equation, and these satisfy*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \to_d \mathcal{N}(0, 1/\theta_0).$$