Joint Confidence Regions

We have seen how to find confidence intervals for each individual β_j, $1 \leq j \leq p$. We next consider the problem of finding a confidence region C in R^p that covers the vector $\boldsymbol{\beta}$ with prescribed probability $(1-\alpha)$. This can be done by inverting the likelihood ratio test or equivalently the F test. That is, we let C be the collection of $\boldsymbol{\beta}_0$ that is accepted when the level $(1-\alpha)$ F test is used to test $H : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. Under H, $\boldsymbol{\mu} = \boldsymbol{\mu}_0 = \mathbf{Z}\boldsymbol{\beta}_0$; and the numerator of the F statistic (6.1.22) is based on

$$|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0|^2 = |\mathbf{Z}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\boldsymbol{\beta}_0|^2 = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T(\mathbf{Z}^T\mathbf{Z})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

Thus, using (6.1.22), the simultaneous confidence region for $\boldsymbol{\beta}$ is the ellipse

$$C = \left\{ \boldsymbol{\beta}_0 : \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T(\mathbf{Z}^T\mathbf{Z})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{rs^2} \leq f_{r,n-r}\left(1 - \tfrac{1}{2}\alpha\right) \right\} \tag{6.1.31}$$

where $f_{r,n-r}\left(1 - \frac{1}{2}\alpha\right)$ is the $1 - \frac{1}{2}\alpha$ quantile of the $\mathcal{F}_{r,n-r}$ distribution.

Example 6.1.2. *Regression (continued).* We consider the case $p = r$ and as in (6.1.26) write $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ and $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$, where $\boldsymbol{\beta}_2$ is a vector of main effect coefficients and $\boldsymbol{\beta}_1$ is a vector of "nuisance" coefficients. Similarly, we partition $\widehat{\boldsymbol{\beta}}$ as $\widehat{\boldsymbol{\beta}}^T = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)$ where $\widehat{\boldsymbol{\beta}}_1$ is $q \times 1$ and $\widehat{\boldsymbol{\beta}}_2$ is $(p-q) \times 1$. By Corollary 6.1.1, $\sigma^2(\mathbf{Z}^T\mathbf{Z})$ is the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$. It follows that if we let $\mathbf{S}$ denote the lower right $(p-q) \times (p-q)$ corner of $(\mathbf{Z}^T\mathbf{Z})^{-1}$, then $\sigma^2\mathbf{S}$ is the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}_2$. Thus, a joint $100(1-\alpha)\%$ confidence region for $\boldsymbol{\beta}_2$ is the $p-q$ dimensional ellipse

$$C = \left\{ \boldsymbol{\beta}_{02} : \frac{(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{02})^T S^{-1}(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{02})}{(p-q)s^2} \leq f_{p-q,n-p}\left(1 - \tfrac{1}{2}\alpha\right) \right\}.$$

□

Summary. We consider the classical Gaussian linear model in which the resonse Y_i for the ith case in an experiment is expressed as a linear combination $\mu_i = \sum_{j=1}^{p} \beta_j z_{ij}$ of covariates plus an error ϵ_i, where $\epsilon_i, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. By introducing a suitable orthogonal transformation, we obtain a canonical model in which likelihood analysis is straightforward. The inverse of the orthogonal transformation gives procedures and results in terms of the original variables. In particular we obtain maximum likelihood estimates, likelihood ratio tests, and confidence procedures for the regression coefficients $\{\beta_j\}$, the response means $\{\mu_i\}$, and linear combinations of these.

6.2 ASYMPTOTIC ESTIMATION THEORY IN p DIMENSIONS

In this section we largely parallel Section 5.4 in which we developed the asymptotic properties of the MLE and related tests and confidence bounds for one-dimensional parameters. We leave the analogue of Theorem 5.4.1 to the problems and begin immediately generalizing Section 5.4.2.

6.2.1 Estimating Equations

Our assumptions are as before save that everything is made a vector: $X_1, \ldots, X_n$ are i.i.d. P where $P \in \mathcal{Q}$, a model containing $\mathcal{P} \equiv \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ such that

(i) Θ open $\subset R^p$.

(ii) Densities of P_θ are $p(\cdot, \boldsymbol{\theta})$, $\theta \in \Theta$.

The following result gives the general asymptotic behavior of the solution of estimating equations.

A0. $\boldsymbol{\Psi} \equiv (\psi_1, \ldots, \psi_p)^T$ where $\psi_j = \frac{\partial \rho}{\partial \theta_j}$ is well defined and

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(X_i, \bar{\boldsymbol{\theta}}_n) = \mathbf{0}. \tag{6.2.1}$$

A solution to (6.2.1) is called an *estimating equation estimate* or an *M-estimate*.

A1. The parameter $\boldsymbol{\theta}(P)$ given by the solution of (the nonlinear system of p equations in p unknowns):

$$\int \boldsymbol{\Psi}(x, \boldsymbol{\theta}) dP(x) = \mathbf{0} \tag{6.2.2}$$

is well defined on $\mathcal{Q}$ so that $\boldsymbol{\theta}(P)$ is the unique solution of (6.2.2). Necessarily $\boldsymbol{\theta}(P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ because $\mathcal{Q} \supset \mathcal{P}$.

A2. $E_P|\Psi(X_1, \boldsymbol{\theta}(P))|^2 < \infty$ where $|\cdot|$ is the Euclidean norm.

A3. $\psi_i(\cdot, \boldsymbol{\theta})$, $1 \leq i \leq p$, have first-order partials with respect to all coordinates and using the notation of Section B.8,

$$E_P|D\boldsymbol{\Psi}(X_1, \boldsymbol{\theta})| < \infty$$

where

$$E_P D\boldsymbol{\Psi}(X_1, \boldsymbol{\theta}) = \left\| E_P \frac{\partial \psi_i}{\partial \theta_j}(X_1, \boldsymbol{\theta}) \right\|_{p \times p}$$

is nonsingular.

A4. $\sup\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(D\boldsymbol{\Psi}(X_i, \mathbf{t}) - D\boldsymbol{\Psi}(X_i, \boldsymbol{\theta}(P)))\right| : |\mathbf{t} - \boldsymbol{\theta}(P)| \leq \epsilon_n\right\} \xrightarrow{P} 0$ if $\epsilon_n \to 0$.

A5. $\bar{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}(P)$ for all $P \in \mathcal{Q}$.

Theorem 6.2.1. *Under* A0–A5 *of this section*

$$\bar{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(P) + \frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{\Psi}}(X_i, \boldsymbol{\theta}(P)) + o_p(n^{-1/2}) \tag{6.2.3}$$

where

$$\tilde{\boldsymbol{\Psi}}(x, \boldsymbol{\theta}(P)) = -[E_P D\boldsymbol{\Psi}(X_1, \boldsymbol{\theta}(P))]^{-1}\boldsymbol{\Psi}(x, \boldsymbol{\theta}(P)). \tag{6.2.4}$$

Hence,

$$\mathcal{L}_p(\sqrt{n}(\bar{\theta}_n - \theta(P))) \to \mathcal{N}_p(0, \Sigma(\Psi, P)) \tag{6.2.5}$$

where

$$\Sigma(\Psi, P) = J(\theta, P)E\Psi\Psi^T(X_1, \theta(P))J^T(\theta, P) \tag{6.2.6}$$

and

$$J^{-1}(\theta, P) = -E_P D\Psi(X_1, \theta(P)) = \left\| -E_P \frac{\partial \psi_i}{\partial \theta_j}(X_1, \theta(P)) \right\|.$$

The proof of this result follows precisely that of Theorem 5.4.2 save that we need multivariate calculus as in Section B.8. Thus,

$$-\frac{1}{n}\sum_{i=1}^{n} \Psi(X_i, \theta(P)) = \frac{1}{n}\sum_{i=1}^{n} D\Psi(X_i, \theta_n^*)(\bar{\theta}_n - \theta(P)). \tag{6.2.7}$$

Note that the left-hand side of (6.2.7) is a $p \times 1$ vector, the right is the product of a $p \times p$ matrix and a $p \times 1$ vector.

The rest of the proof follows essentially exactly as in Section 5.4.2 save that we need the observation that the set of nonsingular $p \times p$ matrices, when viewed as vectors, is an open subset of R^{p^2}, representable, for instance, as the set of vectors for which the determinant, a continuous function of the entries, is different from zero. We use this remark to conclude that A3 and A4 guarantee that with probability tending to 1, $\frac{1}{n}\sum_{i=1}^{n} D\Psi(X_i, \theta_n^*)$ is nonsingular.

Note. This result goes beyond Theorem 5.4.2 in making it clear that although the definition of $\bar{\theta}_n$ is motivated by $\mathcal{P}$, the behavior in (6.2.3) is guaranteed for $P \in \mathcal{Q}$, which can include $P \notin \mathcal{P}$. In fact, typically $\mathcal{Q}$ is essentially the set of P's for which $\theta(P)$ can be defined uniquely by (6.2.2).

We can again extend the assumptions of Section 5.4.2 to:

A6. If $l(\cdot, \theta)$ is differentiable

$$\begin{aligned} E_\theta D\Psi(X_1, \theta) &= -E_\theta \Psi(X_1, \theta) Dl(X_1, \theta) \\ &= -\mathrm{Cov}_\theta(\Psi(X_1, \theta), Dl(X_1, \theta)) \end{aligned} \tag{6.2.8}$$

defined as in B.5.2. The heuristics and conditions behind this identity are the same as in the one-dimensional case. Remarks 5.4.2, 5.4.3, and Assumptions **A4′** and **A6′** extend to the multivariate case readily.

Note that consistency of $\bar{\theta}_n$ is assumed. Proving consistency usually requires different arguments such as those of Section 5.2. It may, however, be shown that with probability tending to 1, a root-finding algorithm starting at a consistent estimate θ_n^* will find a solution $\bar{\theta}_n$ of (6.2.1) that satisfies (6.2.3) (Problem 6.2.10).

6.2.2 Asymptotic Normality and Efficiency of the MLE

If we take $\rho(x,\boldsymbol{\theta}) = l(x,\boldsymbol{\theta}) \equiv \log p(x,\boldsymbol{\theta})$, and $\boldsymbol{\Psi}(x,\boldsymbol{\theta})$ obeys A0–A6, then (6.2.8) becomes

$$\begin{aligned} -\|E_{\boldsymbol{\theta}} D^2 l(X_1,\boldsymbol{\theta})\| &= E_{\boldsymbol{\theta}} Dl(X_1,\boldsymbol{\theta}) D^T l(X_1,\boldsymbol{\theta})) \\ &= \mathrm{Var}_{\boldsymbol{\theta}} Dl(X_1,\boldsymbol{\theta}) \end{aligned} \tag{6.2.9}$$

where

$$\mathrm{Var}_{\boldsymbol{\theta}} Dl(X_1,\boldsymbol{\theta}) = \left\| E_{\boldsymbol{\theta}} \left(\frac{\partial l}{\partial \theta_i}(X_1,\boldsymbol{\theta}) \frac{\partial l}{\partial \theta_j}(X_1,\boldsymbol{\theta}) \right) \right\|$$

is the *Fisher information matrix* $I(\boldsymbol{\theta})$ introduced in Section 3.4. If $\rho : \boldsymbol{\theta} \to R$, $\boldsymbol{\theta} \subset R^d$, is a scalar function, the matrix $\left\| \frac{\partial^2 \rho}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}) \right\|$ is known as the *Hessian* or curvature matrix of the surface ρ. Thus, (6.2.9) states that the expected value of the Hessian of l is the negative of the Fisher information.

We also can immediately state the generalization of Theorem 5.4.3.

Theorem 6.2.2. *If* A0–A6 *hold for* $\rho(x,\boldsymbol{\theta}) \equiv \log p(x,\boldsymbol{\theta})$, *then the MLE* $\widehat{\boldsymbol{\theta}}_n$ *satisfies*

$$\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta} + \frac{1}{n} \sum_{i=1}^{n} I^{-1}(\boldsymbol{\theta}) Dl(X_i,\boldsymbol{\theta}) + o_p(n^{-1/2}) \tag{6.2.10}$$

so that

$$\mathcal{L}(\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})) \to \mathcal{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta})). \tag{6.2.11}$$

If $\bar{\boldsymbol{\theta}}_n$ *is a minimum contrast estimate with* ρ *and* ψ *satisfying* A0–A6 *and corresponding asymptotic variance matrix* $\Sigma(\boldsymbol{\Psi}, P_{\boldsymbol{\theta}})$, *then*

$$\Sigma(\boldsymbol{\Psi}, P_{\boldsymbol{\theta}}) \geq I^{-1}(\boldsymbol{\theta}) \tag{6.2.12}$$

in the sense of Theorem 3.4.4 *with equality in* (6.2.12) *for* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ *iff, under* $\boldsymbol{\theta}_0$,

$$\bar{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n + o_p(n^{-1/2}). \tag{6.2.13}$$

Proof. The proofs of (6.2.10) and (6.2.11) parallel those of (5.4.33) and (5.4.34) exactly. The proof of (6.2.12) parallels that of Theorem 3.4.4. For completeness we give it. Note that by (6.2.6) and (6.2.8)

$$\Sigma(\boldsymbol{\Psi}, P_{\boldsymbol{\theta}}) = \mathrm{Cov}_{\boldsymbol{\theta}}^{-1}(\mathbf{U},\mathbf{V}) \mathrm{Var}_{\boldsymbol{\theta}}(\mathbf{U}) \mathrm{Cov}_{\boldsymbol{\theta}}^{-1}(\mathbf{V},\mathbf{U}) \tag{6.2.14}$$

where $\mathbf{U} \equiv \boldsymbol{\Psi}(X_1,\boldsymbol{\theta})$, $\mathbf{V} = Dl(X_1,\boldsymbol{\theta})$. But by (B.10.8), for any $\mathbf{U},\mathbf{V}$ with $\mathrm{Var}(\mathbf{U}^T,\mathbf{V}^T)^T$ nonsingular

$$\mathrm{Var}(\mathbf{V}) \geq \mathrm{Cov}(\mathbf{U},\mathbf{V}) \mathrm{Var}^{-1}(\mathbf{U}) \mathrm{Cov}(\mathbf{V},\mathbf{U}). \tag{6.2.15}$$

Taking inverses of both sides yields

$$I^{-1}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}^{-1}(\mathbf{V}) \leq \Sigma(\boldsymbol{\Psi},\boldsymbol{\theta}). \tag{6.2.16}$$

Equality holds in (6.2.15) by (B.10.2.3) iff for some $\mathbf{b} = b(\boldsymbol{\theta})$

$$\mathbf{U} = \mathbf{b} + \mathrm{Cov}(\mathbf{U}, \mathbf{V})\mathrm{Var}^{-1}(\mathbf{V})\mathbf{V} \tag{6.2.17}$$

with probability 1. This means in view of $E_\theta \Psi = E_{\boldsymbol{\theta}} Dl = 0$ that

$$\Psi(X_1, \boldsymbol{\theta}) = \mathbf{b}(\boldsymbol{\theta}) Dl(X_1, \boldsymbol{\theta}).$$

In the case of identity in (6.2.16) we must have

$$-[E_{\boldsymbol{\theta}} D\Psi(X_1, \boldsymbol{\theta})]^{-1} \Psi(X_1, \boldsymbol{\theta}) = I^{-1}(\boldsymbol{\theta}) Dl(X_1, \boldsymbol{\theta}). \tag{6.2.18}$$

Hence, from (6.2.3) and (6.2.10) we conclude that (6.2.13) holds. □

We see that, by the theorem, the MLE is *efficient* in the sense that for any $\mathbf{a}_{p\times 1}$, $\mathbf{a}^T\widehat{\boldsymbol{\theta}}_n$ has asymptotic bias $o(n^{-1/2})$ and asymptotic variance $n^{-1}\mathbf{a}^T I^{-1}(\boldsymbol{\theta})\mathbf{a}$, which is no larger than that of any competing minimum contrast estimate. Further any competitor $\bar{\boldsymbol{\theta}}_n$ such that $\mathbf{a}^T\bar{\boldsymbol{\theta}}_n$ has the same asymptotic behavior as $\mathbf{a}^T\widehat{\boldsymbol{\theta}}_n$ for *all* $\mathbf{a}$ in fact agrees with $\widehat{\boldsymbol{\theta}}_n$ to order $n^{-1/2}$.

A special case of Theorem 6.2.2 that we have already established is Theorem 5.3.6 on the asymptotic normality of the MLE in canonical exponential families. A number of important new statistical issues arise in the multiparameter case. We illustrate with an example.

Example 6.2.1. *The Linear Model with Stochastic Covariates.* Let $X_i = (\mathbf{Z}_i^T, Y_i)^T$, $1 \le i \le n$, be i.i.d. as $X = (\mathbf{Z}^T, Y)^T$ where $\mathbf{Z}$ is a $p \times 1$ vector of explanatory variables and Y is the response of interest. This model is discussed in Section 2.2.1 and Example 1.4.3. We specialize in two ways:

(i)

$$Y = \alpha + \mathbf{Z}^T\boldsymbol{\beta} + \epsilon \tag{6.2.19}$$

where ϵ is distributed as $\mathcal{N}(0, \sigma^2)$ independent of $\mathbf{Z}$ and $E(\mathbf{Z}) = \mathbf{0}$. That is, given $\mathbf{Z}$, Y has a $\mathcal{N}(\alpha + \mathbf{Z}^T\boldsymbol{\beta}, \sigma^2)$ distribution.

(ii) The distribution H_0 of $\mathbf{Z}$ is known with density h_0 and $E(\mathbf{Z}\mathbf{Z}^T)$ is nonsingular.

The second assumption is unreasonable but easily dispensed with. It readily follows (Problem 6.2.6) that the MLE of $\boldsymbol{\beta}$ is given by (with probability 1)

$$\widehat{\boldsymbol{\beta}} = [\widetilde{\mathbf{Z}}_{(n)}^T \widetilde{\mathbf{Z}}_{(n)}]^{-1} \widetilde{\mathbf{Z}}_{(n)}^T \mathbf{Y}. \tag{6.2.20}$$

Here $\widetilde{\mathbf{Z}}_{(n)}$ is the $n \times p$ matrix $\|Z_{ij} - Z_{\cdot j}\|$ where $Z_{\cdot j} = \frac{1}{n}\sum_{i=1}^n Z_{ij}$. We used subscripts (n) to distinguish the use of $\mathbf{Z}$ as a vector in this section and as a matrix in Section 6.1. In the present context, $\mathbf{Z}_{(n)} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^T$ is referred to as the *random design matrix*. This example is called the *random design case* as opposed to the *fixed design case* of Section 6.1. Also the MLEs of α and σ^2 are

$$\widehat{\alpha} = \bar{Y} - \sum_{j=1}^p Z_{\cdot j}\widehat{\beta}_j,\ \widehat{\sigma}^2 = \frac{1}{n}|\mathbf{Y} - (\widehat{\alpha} + \mathbf{Z}_{(n)}\widehat{\boldsymbol{\beta}})|^2. \tag{6.2.21}$$

Note that although given $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, $\widehat{\boldsymbol{\beta}}$ is Gaussian, this is not true of the marginal distribution of $\widehat{\boldsymbol{\beta}}$.

It is not hard to show that A0–A6 hold in this case because if H_0 has density h_0 and if $\boldsymbol{\theta}$ denotes $(\alpha, \boldsymbol{\beta}^T, \sigma^2)^T$, then

$$\begin{aligned} l(X, \boldsymbol{\theta}) &= -\frac{1}{2\sigma^2}[Y - (\alpha + \mathbf{Z}^T\boldsymbol{\beta})]^2 - \frac{1}{2}(\log \sigma^2 + \log 2\pi) + \log h_0(\mathbf{z}) \\ Dl(X, \boldsymbol{\theta}) &= \left(\frac{\epsilon}{\sigma^2}, \mathbf{Z}\frac{\epsilon}{\sigma^2}, \frac{1}{2\sigma^4}(\epsilon^2 - 1)\right) \end{aligned} \tag{6.2.22}$$

and

$$I(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^{-2} & 0 & 0 \\ \mathbf{0} & \sigma^{-2}E(\mathbf{Z}\mathbf{Z}^T) & \mathbf{0} \\ 0 & 0 & \frac{1}{2\sigma^4} \end{pmatrix} \tag{6.2.23}$$

so that by Theorem 6.2.2

$$\mathcal{L}(\sqrt{n}(\widehat{\alpha} - \alpha, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \widehat{\sigma}^2 - \sigma^2)) \to \mathcal{N}(0, \operatorname{diag}(\sigma^2, \sigma^2[E(\mathbf{Z}\mathbf{Z}^T)]^{-1}, 2\sigma^4)). \tag{6.2.24}$$

This can be argued directly as well (Problem 6.2.8). It is clear that the restriction of H_0 known plays no role in the limiting result for $\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2$. Of course, these will only be the MLEs if H_0 depends only on parameters other than $(\alpha, \boldsymbol{\beta}, \sigma^2)$. In this case we can estimate $E(\mathbf{Z}\mathbf{Z}^T)$ by $\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^T$ and give approximate confidence intervals for β_j, $j = 1, \ldots, p$.

An interesting feature of (6.2.23) is that because $I(\boldsymbol{\theta})$ is a block diagonal matrix so is $I^{-1}(\boldsymbol{\theta})$ and, consequently, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are asymptotically independent. In the classical linear model of Section 6.1 where we perform inference conditionally given $\mathbf{Z}_i = \mathbf{z}_i$, $1 \leq i \leq n$, we have noted this is exactly true.

This is an example of the phenomenon of *adaptation*. If we knew σ^2, the MLE would still be $\widehat{\boldsymbol{\beta}}$ and its asymptotic variance optimal for this model. If we knew α and $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\sigma}}^2$ would no longer be the MLE. But its asymptotic variance would be the same as that of the MLE and, by Theorem 6.2.2, $\widehat{\sigma}^2$ would be asymptotically equivalent to the MLE. To summarize, estimating either parameter with the other being a nuisance parameter is no harder than when the nuisance parameter is known. Formally, in a model $\mathcal{P} = \{P_{(\theta,\eta)} : \theta \in \Theta, \eta \in \mathcal{E}\}$ we say we can estimate θ *adaptively* at η_0 if the asymptotic variance of the MLE $\widehat{\theta}$ (or more generally, an efficient estimate of θ) in the pair $(\widehat{\theta}, \widehat{\eta})$ is the same as that of $\widehat{\theta}(\eta_0)$, the efficient estimate for $\mathcal{P}_{\eta_0} = \{P_{(\theta,\eta_0)} : \theta \in \Theta\}$. The possibility of adaptation is in fact rare, though it appears prominently in this way in the Gaussian linear model. In particular consider estimating β_1 in the presence of α, $(\beta_2 \ldots, \beta_p)$ with

(i) $\alpha, \beta_2, \ldots, \beta_p$ known.

(ii) $\boldsymbol{\beta}$ arbitrary.

In case (i), we take, without loss of generality, $\alpha = \beta_2 = \cdots = \beta_p = 0$. Let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$, then the efficient estimate in case (i) is

$$\widehat{\beta}_1^0 = \frac{\sum_{i=1}^n Z_{i1}Y_i}{\sum_{i=1}^n Z_{i1}^2} \tag{6.2.25}$$

with asymptotic variance $\sigma^2[EZ_1^2]^{-1}$. On the other hand, $\widehat{\beta}_1$ is the first coordinate of $\widehat{\boldsymbol{\beta}}$ given by (6.2.20). Its asymptotic variance is the $(1,1)$ element of $\sigma^2[E\mathbf{Z}\mathbf{Z}^T]^{-1}$, which is strictly bigger than $\sigma^2[EZ_1^2]^{-1}$ unless $[E\mathbf{Z}\mathbf{Z}^T]^{-1}$ is a diagonal matrix (Problem 6.2.3). So in general we cannot estimate β_1 adaptively if $\beta_2, \ldots, \beta_p$ are regarded as nuisance parameters. What is happening can be seen by a representation of $[\mathbf{Z}_{(n)}^T\mathbf{Z}_{(n)}]^{-1}\mathbf{Z}_{(n)}^T\mathbf{Y}$ and $I^{11}(\boldsymbol{\theta})$ where $I^{-1}(\boldsymbol{\theta}) \equiv \|I^{ij}(\boldsymbol{\theta})\|$. We claim that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (Z_{i1} - \widehat{Z}_i^{(1)})Y_i}{\sum_{i=1}^n (Z_{i1} - \widehat{Z}_i^{(1)})^2} \tag{6.2.26}$$

where $\widehat{\mathbf{Z}}^{(1)}$ is the regression of $(Z_{11}, \ldots, Z_{1n})^T$ on the linear space spanned by $(Z_{j1}, \ldots, Z_{jn})^T$, $2 \leq j \leq p$. Similarly,

$$I^{11}(\boldsymbol{\theta}) = \sigma^2 / E(Z_{11} - \Pi(Z_{11} \mid Z_{21}, \ldots, Z_{p1}))^2 \tag{6.2.27}$$

where $\Pi(Z_{11} \mid Z_{21}, \ldots, Z_{p1})$ is the projection of Z_{11} on the linear span of $Z_{21}, \ldots, Z_{p1}$ (Problem 6.2.11). Thus, $\Pi(Z_{11} \mid Z_{21}, \ldots, Z_{p1}) = \sum_{j=2}^p a_j^* Z_{j1}$ where $(a_2^*, \ldots, a_p^*)$ minimizes $E(Z_{11} - \sum_{j=2}^p a_j Z_{j1})^2$ over $(a_2, \ldots, a_p) \in R^{p-1}$ (see Sections 1.4 and B.10). What (6.2.26) and (6.2.27) reveal is that there is a price paid for not knowing $\beta_2, \ldots, \beta_p$ when the variables $Z_2, \ldots, Z_p$ are in any way correlated with Z_1 and the price is measured by

$$\frac{[E(Z_{11} - \Pi(Z_{11} \mid Z_{21}, \ldots, Z_{p1})^2]^{-1}}{E(Z_{11}^2)} = \left(1 - \frac{E(\Pi(Z_{11} \mid Z_{21}, \ldots, Z_{p1}))^2}{E(Z_{11}^2)}\right)^{-1}. \tag{6.2.28}$$

In the extreme case of perfect collinearity the price is ∞ as it should be because β_1 then becomes unidentifiable. Thus, adaptation corresponds to the case where $(Z_2, \ldots, Z_p)$ have no value in predicting Z_1 linearly (see Section 1.4). Correspondingly in the Gaussian linear model (6.1.3) conditional on the $\mathbf{Z}_i$, $i = 1, \ldots, n$, $\widehat{\beta}_1$ is undefined if the denominator in (6.2.26) is 0, which corresponds to the case of collinearity and occurs with probability 1 if $E(Z_{11} - \Pi(Z_{11} \mid Z_{21}, \ldots, Z_{p1}))^2 = 0$. □

Example 6.2.2. *M Estimates Generated by Linear Models with General Error Structure.* Suppose that the ϵ_i in (6.2.19) are i.i.d. but not necessarily Gaussian with density $\frac{1}{\sigma} f_0\left(\frac{\cdot}{\sigma}\right)$, for instance,

$$f_0(x) = \frac{e^{-x}}{(1 + e^{-x})^2},$$

the logistic density. Such error densities have the often more realistic, heavier tails[1] than the Gaussian density. The estimates $\widehat{\boldsymbol{\beta}}_0, \widehat{\sigma}_0$ now solve

$$\sum_{i=1}^n Z_{ij}\psi\left(\widehat{\sigma}^{-1}\left(Y_i - \sum_{k=1}^p \widehat{\beta}_{k0} Z_{ik}\right)\right) = 0$$

and

$$\sum_{i=1}^n \frac{1}{\widehat{\sigma}}\chi\left(\widehat{\sigma}^{-1}\left(Y_i - \sum_{k=1}^p \widehat{\beta}_{k0} Z_{ik}\right)\right) = 0$$

where $\psi = -\frac{f_0'}{f_0}$, $\chi(y) = -\left(y\frac{f_0'}{f_0}(y) + 1\right)$, $\widehat{\boldsymbol{\beta}}_0 \equiv (\widehat{\beta}_{10}, \ldots, \widehat{\beta}_{p0})^T$. The assumptions of Theorem 6.2.2 may be shown to hold (Problem 6.2.9) if

(i) $\log f_0$ is strictly concave, i.e., $\frac{f_0'}{f_0}$ is strictly decreasing.

(ii) $(\log f_0)''$ exists and is bounded.

Then, if further f_0 is symmetric about 0,

$$\begin{aligned} I(\boldsymbol{\theta}) &= \sigma^{-2} I(\boldsymbol{\beta}^T, 1) \\ &= \sigma^{-2}\begin{pmatrix} c_1 E(\mathbf{Z}^T\mathbf{Z}) & \mathbf{0} \\ \mathbf{0} & c_2 \end{pmatrix} \end{aligned} \tag{6.2.29}$$

where $c_1 = \int \left(\frac{f_0'}{f_0}(x)\right)^2 f_0(x)dx$, $c_2 = \int \left(x\frac{f_0'}{f_0}(x) + 1\right)^2 f_0(x)dx$. Thus, $\widehat{\boldsymbol{\beta}}_0, \widehat{\sigma}_0$ are optimal estimates of $\boldsymbol{\beta}$ and σ in the sense of Theorem 6.2.2 if f_0 is true.

Now suppose f_0 generating the estimates $\widehat{\beta}_0$ and $\widehat{\sigma}_0^2$ is symmetric and satisfies (i) and (ii) but the true error distribution has density f possibly different from f_0. Under suitable conditions we can apply Theorem 6.2.1 with

$$\boldsymbol{\Psi}(\mathbf{Z}, Y, \boldsymbol{\beta}, \sigma) = (\psi_1, \ldots, \psi_p, \psi_{p+1})^T(\mathbf{Z}, Y, \boldsymbol{\beta}, \sigma)$$

where

$$\begin{aligned} \psi_j(\mathbf{z}, y, \boldsymbol{\beta}, \sigma) &= \frac{z_j}{\sigma}\psi\left(\frac{y - \sum_{k=1}^p z_k\beta_k}{\sigma}\right),\ 1 \le j \le p \\ \psi_{p+1}(\mathbf{z}, y, \boldsymbol{\beta}, \sigma) &= \frac{1}{\sigma}\chi\left(\frac{y - \sum_{k=1}^p z_k\beta_k}{\sigma}\right) \end{aligned} \tag{6.2.30}$$

to conclude that

$$\begin{aligned} \mathcal{L}_0(\sqrt{n}(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0)) &\to \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\Psi}, P)) \\ \mathcal{L}(\sqrt{n}(\widehat{\sigma} - \sigma_0) &\to \mathcal{N}(0, \sigma^2(P)) \end{aligned}$$

where $\boldsymbol{\beta}_0, \sigma_0$ solve

$$\int \boldsymbol{\Psi}(\mathbf{y} - \mathbf{z}^T\boldsymbol{\beta}_0)dP = \mathbf{0} \tag{6.2.31}$$

and $\Sigma(\boldsymbol{\Psi}, P)$ is as in (6.2.6). What is the relation between $\boldsymbol{\beta}_0$, σ_0 and $\boldsymbol{\beta}, \sigma$ given in the Gaussian model (6.2.19)? If f_0 is symmetric about 0 and the solution of (6.2.31) is unique, then $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$. But $\sigma_0 = c(f_0)\sigma$ for some $c(f_0)$ typically different from one. Thus, $\widehat{\boldsymbol{\beta}}_0$ can be used for estimating $\boldsymbol{\beta}$ although if the true distribution of the ϵ_i is $\mathcal{N}(\mathbf{0}, \sigma^2)$ it should perform less well than $\widehat{\boldsymbol{\beta}}$. On the other hand, $\widehat{\sigma}_0$ is an estimate of σ only if normalized by a constant depending on f_0. (See Problem 6.2.5.) These are issues of robustness, that is, to have a bounded sensitivity curve (Section 3.5, Problem 3.5.8), we may well wish to use a nonlinear bounded $\boldsymbol{\Psi} = (\psi_1, \ldots, \psi_p)^T$ to estimate $\boldsymbol{\beta}$ even though it is suboptimal when $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and to use a suitably normalized version of $\widehat{\sigma}_0$ for the same purpose. One effective choice of ψ_j is the Huber function defined in Problem 3.5.8. We will discuss these issues further in Section 6.6 and Volume II. □

Testing and Confidence Bounds

There are three principal approaches to testing hypotheses in multiparameter models, the likelihood ratio principle, Wald tests (a generalization of pivots), and Rao's tests. All of these will be developed in Section 6.3. The three approaches coincide asymptotically but differ substantially, in performance and computationally, for fixed n. Confidence regions that parallel the tests will also be developed in Section 6.3.

Optimality criteria are not easily stated even in the fixed sample case and not very persuasive except perhaps in the case of testing hypotheses about a real parameter in the presence of other nuisance parameters such as $H : \theta_1 \leq 0$ versus $K : \theta_1 > 0$ where $\theta_2, \ldots, \theta_p$ vary freely.

6.2.3 The Posterior Distribution in the Multiparameter Case

The asymptotic theory of the posterior distribution parallels that in the one-dimensional case exactly. We simply make θ a vector, and interpret $|\cdot|$ as the Euclidean norm in conditions A7 and A8. Using multivariate expansions as in B.8 we obtain

Theorem 6.2.3. *If the multivariate versions of* A0–A3, A4(a.s.), A5(a.s.) *and* A6–A8 *hold then, if $\widehat{\theta}$ denotes the MLE,*

$$\mathcal{L}(\sqrt{n}(\theta - \widehat{\theta}) \mid X_1, \ldots, X_n) \to \mathcal{N}(\mathbf{0}, I^{-1}(\theta)) \tag{6.2.32}$$

a.s. under P_θ for all θ.

The consequences of Theorem 6.2.3 are the same as those of Theorem 5.5.2, the equivalence of Bayesian and frequentist optimality asymptotically.

Again the two approaches differ at the second order when the prior begins to make a difference. See Schervish (1995) for some of the relevant calculations.

A new major issue that arises is computation. Although it is easy to write down the posterior density of θ, $\pi(\theta)\prod_{i=1}^n p(X_i, \theta)$, up to the proportionality constant $\int_\Theta \pi(\mathbf{t})\prod_{i=1}^n p(X_i, \mathbf{t})d\mathbf{t}$, the latter can pose a formidable problem if $p > 2$, say. The problem arises also when, as is usually the case, we are interested in the posterior distribution of some of the parameters, say (θ_1, θ_2), because we then need to integrate out $(\theta_3, \ldots, \theta_p)$. The asymptotic theory we have developed permits approximation to these constants by the procedure used in deriving (5.5.19) (Laplace's method). We have implicitly done this in the calculations leading up to (5.5.19). This approach is refined in Kass, Kadane, and Tierney (1989). However, typically there is an attempt at "exact" calculation. A class of Monte Carlo based methods derived from statistical physics loosely called Markov chain Monte Carlo has been developed in recent years to help with these problems. These methods are beyond the scope of this volume but will be discussed briefly in Volume II.

Summary. We defined minimum contrast (MC) and M-estimates in the case of p-dimensional parameters and established their convergence in law to a normal distribution. When the estimating equations defining the M-estimates coincide with the likelihood

equations, this result gives the asymptotic distribution of the MLE. We find that the MLE is asymptotically efficient in the sense that it has "smaller" asymptotic covariance matrix than that of any MD or M-estimate if we know the correct model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and use the MLE for this model. We use an example to introduce the concept of adaptation in which an estimate $\widehat{\theta}$ is called adaptive for a model $\{P_{\theta,\eta} : \theta \in \Theta,\ \eta \in \mathcal{E}\}$ if the asymptotic distribution of $\sqrt{n}(\widehat{\theta} - \theta)$ has mean zero and variance matrix equal to the smallest possible for a general class of regular estimates of θ in the family of models $\{P_{\theta,\eta_0} : \theta \in \Theta\}$, η_0 specified. In linear regression, adaptive estimation of β_1 is possible iff Z_1 is uncorrelated with every linear function of $Z_2, \ldots, Z_p$. Another example deals with M-estimates based on estimating equations generated by linear models with non-Gaussian error distribution. Finally we show that in the Bayesian framework where given θ, $X_1, \ldots, X_n$ are i.i.d. P_θ, if $\widehat{\theta}$ denotes the MLE for P_θ, then the posterior distribution of $\sqrt{n}(\theta - \widehat{\theta})$ converges a.s. under P_θ to the $\mathcal{N}(0, I^{-1}(\theta))$ distribution.

6.3 LARGE SAMPLE TESTS AND CONFIDENCE REGIONS

In Section 6.1 we developed exact tests and confidence regions that are appropriate in regression and anaysis of variance (ANOVA) situations when the responses are normally distributed. We shall show (see Section 6.6) that these methods in many respects are also approximately correct when the distribution of the error in the model fitted is not assumed to be normal. However, we need methods for situations in which, as in the linear model, covariates can be arbitrary but responses are necessarily discrete (qualitative) or nonnegative and Gaussian models do not seem to be appropriate approximations. In these cases exact methods are typically not available, and we turn to asymptotic approximations to construct tests, confidence regions, and other methods of inference. We present three procedures that are used frequently: likelihood ratio, Wald and Rao large sample tests, and confidence procedures. These were treated for θ real in Section 5.4.4. In this section we will use the results of Section 6.2 to extend some of the results of Section 5.4.4 to vector-valued parameters.

6.3.1 Asymptotic Approximation to the Distribution of the Likelihood Ratio Statistic

In Sections 4.9 and 6.1 we considered the likelihood ratio test statistic,

$$\lambda(\mathbf{x}) = \frac{\sup\{p(\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}}{\sup\{p(\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_0\}}$$

for testing $H : \boldsymbol{\theta} \in \Theta_0$ versus $K : \boldsymbol{\theta} \in \Theta_1$, $\Theta_1 = \Theta - \Theta_0$, and showed that in several statistical models involving normal distributions, $\lambda(\mathbf{x})$ simplified and produced intuitive tests whose critical values can be obtained from the Student t and $\mathcal{F}$ distributions.

However, in many experimental situations in which the likelihood ratio test can be used to address important questions, the exact critical value is not available analytically. In such