Chapter I

INTRODUCTION AND EXAMPLES

I.0 Basic Ideas and Conventions

Recall from Volume I that in the field of statistics we represent important data-related problems and questions in terms of questions about distributions and their parameters. Thus our goal is to use data $X \in \mathcal{X}$ to estimate or draw conclusions about aspects of the probability distribution P of X. The probability distribution P is assumed to belong to a class $\mathcal{P}$ of distributions called the *model*. Examining what models are *useful* for answering data-related questions is an important part of statistics. In Volume I we considered three cases with a focus on the first:

(1) P is a member of a parametric class of distributions $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^d$, and our interest is in θ or some vector $q(\theta)$.

(2) P is arbitrary except for regularity conditions, such as finite second moments or continuity of the distribution function, and our interest is in functionals $\nu(P)$ that may be real valued, vectors, or functions.

(3) Our class of distributions is neither smoothly parametrizable by a Euclidean parameter nor essentially unrestricted.

In Volume I we focussed mostly[(1)] on parametric cases and on situations where the number of parameters we were dealing with was small in at least one of two ways:

(i) The complexity of the regular model $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^d$, as measured by the dimension d of the parametrization, was small in relation to the amount of information, as measured by the sample size n of the data. In particular, when examining the properties of statistical procedures, d does not increase with n.

(ii) The procedures we considered, estimation of low dimensional Euclidean parameters, testing, and confidence regions, corresponded to simple (finite or low dimensional) action spaces $\mathcal{A}$, where $\mathcal{A}$ is the range of the statistical decision procedure.

In this volume we will focus on inference in non- and semiparametric models. In doing so, we will not only reexamine the procedures introduced in Volume I from a more sophisticated point of view but also come to grips with new problems originating from our analysis

of estimation of functions and other complex decision procedures that appear naturally in these contexts. The mathematics needed for this work is often of a higher level than that used in Volume I. But, as before, we present what is needed in the appendices with proofs or references.

Modeling Conventions

The guiding principle of modern statistics was best formulated by George Box (see Section 1.1.):

"Models, of course, are never true, but fortunately it is only necessary that they be useful"

One implication of this statement is that the parameters we deal with are the parameters of the distribution in our model class closest to the unknown true distribution. See Sections 2.2.2, 5.4.2, and 6.2.1. For instance, a linear regression model can detect linear trends that provide useful information even if the true population relationship is not linear. See Figure 1.4.1. This leads to an interesting dilemma and accompanying research questions: The more general a class we postulate the more closely we will be able to approximte the true population distribution. However, using a very general class of models means more parameters and more variability of statistical methods. Achieving a balance leads to useful models. One approach is to use a nested sequence of "regular" parametric models (sieves) that become more general as we add parameters and then select the number of parameters by minimizing estimated prediction error (cross validation). See Chapter 12.

As in Volume I (see Section 1.1.3), except for Bayesian models, our parametric models are restricted to be *regular parametric models*. $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^d$, where P_θ is either continuous or discrete, and in the discrete case $\{x : p(x;\theta) > 0\}$ does not involve θ. But see Section I.5 for a general concept of regular and irregular parameters that includes semiparametric and nonparametric models.

As in our discussion of Bayesian models in Section 1.2, conditioning of continuous variables by discrete variables and vice versa generally preserves the interpretation of the conditional p as being a continuous or discrete case density, respectively. If $\mathbf{X} = (I, Y)^T$ where I is discrete and Y is continuous, then $p(i, y)$ is a density if it satisfies $P(I = i, Y \leq y) = \int_{-\infty}^{y} p(i, t)dt$. Readers familiar with measure theory will recognize that all results remain meaningful when $p = dP/d\mu$, where μ is a σ-finite measure dominating all P under discussion, and conditional densities are interpreted as being with respect to the appropriate conditional measure. All of the proofs can be converted to this general case, subject only to minor technicalities. Finally, we will write $h = 0$ when $h = 0$ a.s. (almost surely). More generally, a.s. equality is denoted as equality.

As in Volume I, throughout this volume, for $x \in R^d$ we shall use $p(x)$ interchangeably for frequency functions $p(x) = P[X = x]$ and for continuous case density functions $p(x)$. We will call $p(\cdot)$ a density function in both cases. When we write $\int h(x)dP(x)$ we will mean $\sum_x h(x)P[X = x]$ or $\int h(x)p(\mathbf{x})d\mathbf{x}$. Unless we indicate otherwise, statements which can be interpreted under either interpretation are valid under both, although proofs in the text will be given under one formalism or the other. That is, when we write $\int h(x)p(x)dx$, for instance, we really mean $\int h(x)dP(x)$ as interpreted above. When $d = 1$, we let $F(x) = P(-\infty, x]$ and often write $\int h(x)dF(x)$ for $\int h(x)dP(x)$.

Selected Topics

Statistical methods in Volume II include the bootstrap, Markov Chain Monte Carlo (MCMC), Steinian shrinkage, sieves, cross-validation, censored data analysis, Cox proportional hazard regression, nonparametric curve (kernel) estimation, model selection, classification, prediction, classification and regression trees (CART), penalty estimation such as the Lasso, and Bayesian procedures. The effectiveness of statistical methods is examined using classical concepts such as risk, Bayes risk, mean squared error, power, minimaxity, admissibility, invariance, and equivariance. Statistical methods that are optimal based on such criteria are obtained in a finite sample context in Chapter 8. However, most of the book is concerned with asymptotic theory including empirical process theory and efficient estimation in semiparametric models as well as the development of the asymptotic properties of the statistical methods listed above. We present in Section I.1–I.7 a few more details of some of the topics in Volume II.

Notation

$X \sim F$, X is distributed according to F

statistic, a function of observable data $\mathbf{X}$ only

$\mathcal{L}(X)$, the distribution, or law, of X

$X_n \Longrightarrow X$, $\mathcal{L}(X_n) \to \mathcal{L}(X)$ X_n converges weakly (in law) to X

df, distribution function

J, identity matrix = diag$(1, \ldots, 1)$

$\bar{F} = 1 - F$, the survival function

$[t]$, greatest integer less than or equal to t

i.i.d., independent identically distributed

sample, $X_1, \ldots, X_n$ i.i.d. as $X \sim F$

$\widehat{P}$ and $\widehat{P}_n$, empirical probability of a sample $X_1, \ldots, X_n$

$\mathcal{B}(n, \theta)$, binomial distribution with parameters n and θ

Ber(θ), Bernoulli distribution = $\mathcal{B}(1, \theta)$

$\mathcal{E}(\lambda)$, exponential distribution with parameter λ (mean $1/\lambda$)

$\mathcal{H}(D, N, n)$, hypergeometric distribution with parameters D, N, n

$\mathcal{M}(n, \theta_1, \ldots, \theta_q)$, multinomial distribution with parameters $n, \theta_1, \ldots, \theta_q$

$\mathcal{N}(\mu, \sigma^2)$, normal (Gaussian) distribution with mean μ and variance σ^2

$\varphi, \mathcal{N}(0,1)$ density

$\Phi, \mathcal{N}(0,1)$ df

z_α, αth quantile of $\Phi : z_\alpha = \Phi^{-1}(\alpha)$

$\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, bivariate normal (Gaussian) distribution

$\mathcal{N}(\mu, \Sigma)$, multivariate normal (Gaussian) distribution

$\mathcal{P}(\lambda)$, Poisson distribution with parameter λ

$\mathcal{U}(a,b)$, uniform distribution on the interval (a,b)

d.f., degrees of freedom

χ_k^2, chi-square distribution with k d.f.

$\equiv$, defined to be equal to

$\perp$, orthogonal to, uncorrelated

$1(\cdot)$, indicator function

The $O_P, \asymp_P$, and o_P Notation

The following asymptotic order in probability notation is from Section B.7. Let $\mathbf{U}_n$ and $\mathbf{V}_n$ be random vectors in R^d and let $|\cdot|$ denote Euclidean distance.

$$
\begin{aligned}
\mathbf{U}_n = o_P(1) \quad &\text{iff} \quad \mathbf{U}_n \xrightarrow{P} 0, \text{ that is, } \forall \epsilon > 0, P(|\mathbf{U}_n| > \epsilon) \to 0 \\
\mathbf{U}_n = O_P(1) \quad &\text{iff} \quad \forall \epsilon > 0, \exists M < \infty \text{ such that } \forall n \quad P[|\mathbf{U}_n| \geq M] \leq \epsilon \\
\mathbf{U}_n = o_P(\mathbf{V}_n) \quad &\text{iff} \quad \frac{|\mathbf{U}_n|}{|\mathbf{V}_n|} = o_p(1) \\
\mathbf{U}_n = O_P(\mathbf{V}_n) \quad &\text{iff} \quad \frac{|\mathbf{U}_n|}{|\mathbf{V}_n|} = O_P(1) \\
\mathbf{U}_n \asymp_P \mathbf{V}_n \quad &\text{iff} \quad \mathbf{U}_n = O_P(\mathbf{V}_n) \quad \text{and} \quad \mathbf{V}_n = O_P(\mathbf{U}_n) \\
\mathbf{U}_n = \Omega_P(\mathbf{V}_n) \quad &\text{iff } \mathbf{U}_n \asymp_P \mathbf{V}_n
\end{aligned}
$$

Note that

$$
O_P(1)o_P(1) = o_P(1), \; O_P(1) + o_P(1) = O_P(1), \tag{I.1}
$$

and $\mathbf{U}_n \xrightarrow{L} \mathbf{U} \Rightarrow \mathbf{U}_n = O_P(1)$.

Suppose $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are i.i.d. as $\mathbf{Z}$ with $E|\mathbf{Z}| < \infty$. Set $\boldsymbol{\mu} = E(\mathbf{Z})$, then $\bar{\mathbf{Z}}_n = \boldsymbol{\mu} + o_p(1)$ by the weak law of large numbers. If $E|\mathbf{Z}|^2 < \infty$, then $\bar{\mathbf{Z}}_n = \boldsymbol{\mu} + O_p(n^{-\frac{1}{2}})$ by the central limit theorem.

I.1 Tests of Goodness of Fit and the Brownian Bridge

Let $X_1, ..., X_n$ be i.i.d. as X with distribution P. For one dimensional observations, the distribution function (df) $F(\cdot) = P[X \leq \cdot]$ is a natural infinite dimensional parameter to consider. In Example 4.1.5 we showed how one could use the Kolmogorov statistic $T(\widehat{F}, F_0)$, where $T(\widehat{F}, F) \equiv \sup_{t\in R} |\widehat{F}(t) - F(t)|$ and $\widehat{F}$ is the empirical df, to construct a test of the hypothesis $H : F = F_0$. The test is designed so that one can expect it to be consistent against all alternatives, so that our viewpoint is fully nonparametric. In Example 4.4.6 we showed how to construct a simultaneous confidence band for $F(\cdot)$ using the pivot $T(\widehat{F}, F)$. In both cases we noted that the critical values needed for the test and confidence band could be obtained by determining the distribution of $T(\widehat{F}, \mathcal{U})$, where $\mathcal{U}$ is the $Unif[0, 1]$ distribution function under $F = Unif[0, 1]$, and stated that these values could be determined by Monte Carlo simulation.

How does $T(\widehat{F}, F)$ behave qualitatively? We will show in Section 7.1 that, although infinite dimensional, $F(\cdot)$ is a "regular" parameter. In this case, what "regular" means is that the stochastic process,

$$\mathcal{E}_n(x) \equiv \sqrt{n}\,(\widehat{F}(x) - F(x)),\ x \in R\,, \tag{I.2}$$

converges in law in a strong sense (called "weak convergence!") to a Gaussian process $W^0(F(\cdot))$. Here $W^0(u),\ 0 \leq u \leq 1$, is a Gaussian process called the "Brownian bridge" with mean 0 and covariance structure given by,

$$\text{Cov}(W^0(u_1), W^0(u_2)) = u_1(1 - u_2),\ u_1 \leq u_2\ .$$

By "Gaussian" we mean that the distribution of $W^0(u_1), \ldots, W^0(u_k)$ is multivariate normal for all $u_1, \ldots, u_k$. Note that

$$\text{Cov}\left(W^0(F(x_1)), W^0(F(x_2))\right) = \text{Cov}(\mathcal{E}_n(x_1), \mathcal{E}_n(x_2)) = F(x_1)(1-F(x_2)), \quad x_1 \leq x_2.$$

The weak convergence of $\mathcal{E}_n(\cdot)$ to $W^0\big(F(\cdot)\big)$, to be established in Section 7.1, will enable us to derive the Kolmogorov theorem, that when $F = Unif[0, 1]$, $T(\widehat{F}, F)$ converges in law to $\mathcal{L}(\sup\{|W^0(u)| : 0 \leq u \leq 1\})$, which is known analytically. This approach is based on heuristics due to Doob (1949) and developed in Donsker (1952). See also Doob (1953). We will discuss the heuristics in Section 7.1 and apply them to this and other examples in Section 7.2.

These results will provide approximate size α critical values for the Kolmogorov statistics and other interesting functionals of distribution functions. The critical values yield confidence regions for distribution functions and related parameters. See Examples 4.4.6, 4.4.7 and Problems 4.4.17–4.4.19, 4.5.14–4.5.16. □

I.2 Testing Goodness of Fit to Parametric Hypotheses

In Examples 4.1.6 and 4.4.6 we considered the important problem of testing goodness-of-fit to a Gaussian distribution $H : F(\cdot) = \Phi(\frac{\cdot-\mu}{\sigma})$ for some μ, σ. We deduced that the

goodness-of-fit statistic

$$\sup_x |\widehat{G}(x) - \Phi(x)|,$$

where $\widehat{G}$ is the empirical distribution function of $(Z_1, \ldots, Z_n)$ with $Z_i = (X_i - \bar{X})/\widehat{\sigma}$, has a null distribution which does not depend on μ and σ, so that critical values can be calculated by simulating from $\mathcal{N}(0,1)$. In Section 8.2, we will consider other classes of hypothesis models which admit reasonable tests whose critical values can be specified without knowledge of which particular hypothesized distribution is true. However, when we consider the general problem of testing $H : X \sim P \in \mathcal{P}$ where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a regular parametric model, we quickly come to situations where the methods of Chapter 8 will not apply. For instance, suppose that in the Gaussian goodness-of-fit problem above, our observations $X_1, \ldots, X_n$ which are i.i.d. as $F(x) = \Phi([x-\mu]/\sigma)$ are truncated at 0, that is, we assume that we observe $Y_1, \ldots, Y_n$ i.i.d. distributed as $X|X \geq 0$ where $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $P[0,\infty) = 1$ and H is

$$\begin{aligned} P[Y \leq t] \equiv G_{\mu,\sigma^2}(t) = \tfrac{P(0 \leq X \leq t)}{P(X \geq 0)} &= 1 - \left(\Phi\left(\tfrac{\mu - t}{\sigma}\right) \Big/ \Phi\left(\tfrac{\mu}{\sigma}\right)\right), \quad t \geq 0 \\ &= 0, \qquad\qquad\qquad\qquad\quad t < 0. \end{aligned}$$

The only promising approach here is to estimate μ and σ^2 consistently using, for instance, maximum likelihood or the method of moments (Problem I.2.1) by $\widehat{\mu}$ and $\widehat{\sigma}^2$ and estimate the null distribution of

$$T_n \equiv \sup_{t \geq 0} \sqrt{n} |\widehat{F}(t) - G_{\widehat{\mu},\widehat{\sigma}^2}(t)|$$

by simulating samples of size n from $\mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2)$ truncated at 0, i.e., keep as observations only the nonnegative ones. But can this method, called the "parametric bootstrap," be justified? To answer this question we need to consider asymptotics: It turns out that, under H, T_n converges in law to a limit. This helps us little in approximating the null distribution of T_n since an analytic form for its limiting distribution is not available. But, as we show in Section 9.4, such results are essential in justifying the parametric bootstrap.

The more important "nonparametric bootstrap" and other methods for simulating or approximately simulating observations from complicated distributions, often dependent on the data, such as Markov Chain Monte Carlo, are developed in Chapter 10.

I.3 Regular Parameters. Minimum Distance Estimates

We have seen in Chapters 5 and 6 how to establish asymptotic normality and approximate linearity of estimates that are solutions to estimation equations (M estimates) and then used these results to establish efficiency of the MLE under suitable conditions.

There are many types of estimates which cannot be characterized as solutions of estimating equations. Examples we have discussed are the linear combinations of order statistics, such as the trimmed mean introduced in Section 3.5,

$$\bar{X}_\alpha = (n - 2[n\alpha])^{-1} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)}$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the ordered X_i.

Here is another important class of such estimates. Suppose the data $X \in \mathcal{X}$ have probability distribution P. Let $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta, \Theta \subset R^q\}$ be a regular parametric model. There may be a unique $\boldsymbol{\theta}$ such that $P_{\boldsymbol{\theta}} = P$, or if not, we choose the $\boldsymbol{\theta}$ that makes $P_{\boldsymbol{\theta}}$ closest to P in some metric d on the space of probability distributions on $\mathcal{X}$. For instance, if $\mathcal{X} = R$, examples of such metrics are

$$d_\infty(P, Q) = \sup_t |P(-\infty, t] - Q(-\infty, t]| \tag{I.3}$$

and

$$d_2^2(P, Q) = \int_{-\infty}^{\infty} (P(-\infty, t] - Q(-\infty, t])^2 \psi(t) dt \tag{I.4}$$

where $\psi(\cdot)$ is a nonnegative weight function of our choice with $\int_{-\infty}^{\infty} \psi(t)dt < \infty$.

A *minimum distance* estimate $\boldsymbol{\theta}(\widehat{P})$ is obtained by plugging the empirical distribution distribution $\widehat{P}$ into the parameter

$$\boldsymbol{\theta}(P) = \arg\min\{d(P, P_{\boldsymbol{\theta}}) : \boldsymbol{\theta} \in \Theta\}$$

where we assume that $d(Q, P_{\boldsymbol{\theta}})$ is well defined for Q in $\mathcal{M}$, a general class of distributions containing all distributions with finite support. Thus $\mathcal{M}$ contains the probability distribution P generating X and the empirical probability $\widehat{P}$. See Problems 7.2.10 and 7.2.18 for examples and properties of minimum distance estimates $\boldsymbol{\theta}(\widehat{P})$. These problems show $\sqrt{n}$ consistency of $\boldsymbol{\theta}(\widehat{P})$. They also show that $\boldsymbol{\theta}(\widehat{P})$ may not have a linear approximation in the sense of Section 7.2.1, and they may not be asymptotically normally distributed. Note that the minimum contrast estimates of Section 2.1 are of this form but, in this case, $d(Q, P_{\boldsymbol{\theta}}) = \int \rho(x, \boldsymbol{\theta}) dQ(x)$, which is not a metric, but is linear in Q, whereas metrics are not.

Can minimum distance estimates $\boldsymbol{\theta}(\widehat{P})$ be linearized in the sense of (6.2.3), and are they asymptotically Gaussian as we have shown M estimates to be in Section 6.2.1 and 6.2.2? When this is true asymptotic inference is simple as we have seen in Section 6.3. We have effectively studied this question for $\mathcal{X}$ finite in Theorem 5.4.1. To do the general case, we need to extend the notion of Taylor expansion to function spaces, and apply so called maximal inequalities discussed in Section 7.1. In fact, we shall go further and examine function valued estimates such as the quantile function and study conditions under which these can be linearized and shown to be asymptotically Gaussian in the sense of weak convergence which will be rigorously defined in Section 7.1. Moreover, we want to conclude that asymptotic Gaussianity holds uniformly in a suitable sense. This is an important issue which we did not focus on in Volume I. As the Hodges Example 5.4.2 shows, it is possible to have estimates whose asymptotic behavior is not a good guide to the finite n case because of lack of uniformity of convergence when we vary the underlying distribution.

In Section 9.3 we will be concerned with regular parameters $\boldsymbol{\theta}(P)$, ones whose plug in estimates $\boldsymbol{\theta}(\widehat{P})$ converge to $\boldsymbol{\theta}(P)$ at rate $n^{-\frac{1}{2}}$ uniformly over a suitable subset $\mathcal{M}_0$ of a nonparametric family $\mathcal{M}$ of probability distributions.

Definition I.1. *$\theta(\widehat{P})$ converges to $\theta(P)$ at rate δ_n over $\mathcal{M}_0 \subset \mathcal{M}$ iff for all $\varepsilon > 0$ there exists $c < \infty$ such that*

$$\sup\{P[\,|\theta(\widehat{P}) - \theta(P)| \geq c\delta_n] : P \in \mathcal{M}_0\} \leq \varepsilon\,.$$

There is another important set of questions having to do with the extension of the notion of efficient estimation from parametric to non- and semiparametric models. For instance, consider the parameter corresponding to a minimum contrast estimate

$$\nu(P) = \text{argmin}\left\{\int \rho(x,\boldsymbol{\theta})dP(x) : \boldsymbol{\theta} \in \Theta\right\}.$$

Suppose that, as usual, we assume ν is defined for all $P \in \mathcal{M}$, a nonparametric model. Is there any estimate of $\nu(P)$ which behaves regularly and yet is able to achieve smaller asymptotic variance than the minimum contrast estimate $\nu(\widehat{P})$ at some P in $\mathcal{M}$? Note that this is not the same question as asking whether $\nu(\widehat{P})$ is improvable by another such estimator $\eta(\widehat{P})$ such that $\eta(P_{\boldsymbol{\theta}}) = \nu(P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ on a parametric model $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. Intuitively, $\nu(\widehat{P})$ is to first order the only regular estimate of $\nu(P)$ on all of $\mathcal{M}$ and should not be and indeed, is not improvable. Regularity is essential here to exclude the Hodges Example phenomena. We develop this theme further in the context of semiparametric as well as nonparametric models in Section 9.3.

I.4 Permutation Tests

In Section 4.9.3 we considered the Gaussian two-sample problem. We want to compare two samples $X_1, \ldots, X_{n_1}$ (control) and $Y_1, \ldots, Y_{n_2}$ (treatment) from distributions F and G and, in particular, want to test the hypothesis $H : F = G$ of no treatment effect. We studied the classical case in which F and G were $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{N}(\mu + \Delta, \sigma^2)$, respectively. Then H becomes $\Delta = 0$ and we arrived at the classical two-sample t-test. In Example 5.3.7 we showed that, even if $F = G$ was not Gaussian, if $\int x^2 dF(x) < \infty$, the asymptotic level of the test is preserved as $n_1, n_2 \to \infty$. This can be thought of as a result for testing the hypothesis that the semiparametric model $\{(F,G) : F = G\}$ holds within the full nonparametric model $\{(F,G) : F, G \text{ arbitrary}, \int x^2 dF(x) < \infty, \int y^2 dG(y) < \infty\}$. But is it possible to construct tests with reasonable properties which have level $0 < \alpha < 1$ for fixed $n_1,\ n_2$ and all F, G? The answer is yes. We shall study such *permutation tests* and their simple special subclass, the *rank tests*, in Sections 8.2 and 8.3 in the context of classes of composite semiparametric and parametric hypotheses $H : P \in \mathcal{P}_0 \subset \mathcal{P}$ which allow the construction of tests whose null distribution does not depend on where we are in $\mathcal{P}_0$. We have, in fact, already seen examples of such parametric hypotheses in the Gaussian one- and two-sample problems with unknown variance.

I.5 Estimation of Irregular Parameters

We now show that the phenomena we encounter with irregular parameters are not simply a function of infinite dimension but rather manifest themselves as soon as the parameter

space complexity p, as measured naively by parameter space dimension, is comparable to the information in the data as measured naively by the sample size n. Although the distribution function is a reasonable object to study for $\mathcal{X} = R$, a much more visually informative parameter, even for this dimension and certainly for $\mathcal{X} = R^d$ with $d > 1$, is the density function, assuming that we postulate that only P which have continuous case densities p (in the usual sense) are considered. If $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a regular parametric model and P_θ has density $p(\cdot, \theta)$, we can estimate the density by plugging in, say, $\widehat{p} = p(\cdot, \widehat{\theta})$ where $\widehat{\theta}$ is the MLE. If $\mathcal{P}$ is essentially nonparametric, there is no natural extension of the function valued parameter $\nu(P) \equiv p(\cdot)$ to all of the nonparametric class $\mathcal{M}$ since, if $\widehat{P}$ denotes the empirical probability (2.1.15), $\nu(\widehat{P})$, the density of $\widehat{P}$, has no meaning. This leads to a strategy of "regularization" by which we first approximate $\nu(P)$ on $\mathcal{P}$ by $\nu_n(P)$, which extends smoothly to $\widetilde{\nu}_n$ on $\mathcal{M}$, i.e., for which $\widetilde{\nu}_n(\widehat{P})$ makes sense and yet $\nu_n(P) - \nu(P) \to 0$ in some uniform sense as $n \to \infty$. "Regularization" refers precisely to the change of estimation from the "irregular" ν to the "regular" ν_n. In essence, there is now a natural decomposition of the estimation error,

$$\widetilde{\nu}_n(\widehat{P}) - \nu(P) = (\widetilde{\nu}_n(\widehat{P}) - \nu_n(P)) + (\nu_n(P) - \nu(P)). \tag{I.5}$$

The first term in parenthesis can be interpreted as the source of random variability, and the second as that of deterministic "bias." We will loosely refer to this as the "bias-variance decomposition." For instance, consider the usual *histogram estimate* of a one dimensional density $p(\cdot)$,

$$\widehat{p}_h(t) = \widehat{P}[I_j(t)]/h$$

where $I_j = (jh, (j+1)h]$, $-\infty < j < \infty$, and $I_j(t)$ is the unique I_j which contains t. Now $\widehat{p}_h(t)$ is the plug-in estimate for the parameter,

$$p_h(t) \equiv P[I_j(t)]/h.$$

Of course, $p_h \neq p$ for $h > 0$ but if $h = h_n \downarrow 0$, then $p_h(t) \to p(t)$ for all t and $p_{h_n}(\cdot)$ is a sequence of approximating parameters. Now,

$$E(\widehat{p}_h(t) - p_h(t)) = 0,$$

$$\text{Var}\, \widehat{p}_h(t) = p_h(t)(1 - hp_h(t))/hn.$$

Thus, the "variance" part of the decomposition tends to 0 only if $h \to 0$ slower than n^{-1}. On the other hand, the rate of convergence of the "bias" part to 0

$$\text{BIAS}(h) \equiv \frac{1}{h} \int_{jh}^{(j+1)h} (p(s) - p(t))ds$$

is fastest when $h \to 0$ fastest. In fact, typically, at best $h^{-1}\text{BIAS}(h) \to c \neq 0$ (Problem I.5.1). So we see a tension present here in choosing the rate at which $h \to 0$, which is a new phenomenon whose consequences we shall investigate in Chapter 11 and 12.

Irregular parameters play a critical role in nonparametric regression, classification, and prediction which we shall also study in Chapters 11 and 12, as well as even in some aspects

of regular parameter estimation in semiparametric models. As mentioned in Section I.3, the distinction between regular and irregular is loose, roughly corresponding to the distinction between parameters which can at least asymptotically be estimated unbiasedly at rate $n^{-\frac{1}{2}}$ in the sense that $\mathrm{MSE}(\widehat{\theta}) \asymp n^{-\frac{1}{2}}$ for some $\widehat{\theta}$, and for which the usual asymptotic Gaussian-based methods of inference apply straightforwardly; and those which cannot be treated in this way.

I.6 Stein and Empirical Bayes Estimation

For conceptual reasons, we consider the analysis of variance p-sample Gaussian model (Example 6.1.3), specified by $\mathbf{X} = \{X_{ij} : 1 \le i \le n,\ 1 \le j \le p\}$ where the X_{ij} are independent, $\mathcal{N}(\mu_j, \sigma_0^2)$, with $\boldsymbol{\mu}_p = (\mu_1, \ldots, \mu_p)^T$ unknown. We write $\mathcal{P}(n,p)$ for this class of distributions for $\mathbf{X}$. The MLE of $\boldsymbol{\mu}_p$ is

$$\bar{\mathbf{X}}_p \equiv (X_{\cdot 1}, \ldots, X_{\cdot p})^T$$

where $X_{\cdot j} \equiv n^{-1}\sum_{i=1}^n X_{ij}$. Evidently, $\bar{\mathbf{X}}_p \sim \mathcal{N}(\boldsymbol{\mu}_p, (\sigma_0^2/n)J)$ where $J_{p\times p}$ is the identity. Let our loss function be $l(\boldsymbol{\mu}_p, \mathbf{d}) = |\boldsymbol{\mu}_p - \mathbf{d}|^2/p$, where $|\mathbf{t}|$ is the Euclidean norm of the vector $\mathbf{t}$. Then, the MSE is,

$$R(\boldsymbol{\mu}_p, \bar{\mathbf{X}}_p) = \frac{1}{p}\sum_{j=1}^p E(X_{\cdot j} - \mu_j)^2 = \frac{\sigma_0^2}{n}\,. \tag{I.6}$$

We can show $\bar{\mathbf{X}}_p$ is minimax (Problem I.6.1) for each n and p and asymptotically efficient as $n \to \infty$ for p fixed. But, even if p is only ≥ 3, a remarkable phenomenon discovered by Stein (1956(b)) occurs: $\bar{\mathbf{X}}_p$ is not admissible, whatever be n. That is, there exist minimax estimates $\boldsymbol{\delta}^*(\bar{\mathbf{X}}_p)$ such that $R(\boldsymbol{\mu}_p, \boldsymbol{\delta}^*(\bar{\mathbf{X}}_p)) < \sigma_0^2/n$ for all $\boldsymbol{\mu}_p$. A famous simple and intuitively reasonable estimate is *Stein's positive part estimate*,

$$\boldsymbol{\delta}^*(\bar{\mathbf{X}}_p) = \left(1 - \frac{p-2}{|\bar{\mathbf{X}}_p|^2}\right)_+ \bar{\mathbf{X}}_p \tag{I.7}$$

where if y is a scalar, $y_+ \equiv \max(y, 0)$. This estimate shrinks $\bar{\mathbf{X}}_p$ towards $\mathbf{0}$ and if the distance of $\bar{\mathbf{X}}_p$ from $\mathbf{0}$ is smaller than $\sqrt{p-2}$ declares the estimate to be $\mathbf{0}$.

This result can be made more plausible by considering why $\bar{\mathbf{X}}_p$ becomes a poor estimate as $p \to \infty$ for fixed n. Suppose n is fixed. Because $\bar{\mathbf{X}}_p$ is sufficient and normally distributed with independent components we can without loss of generality set $n = 1$. In this case we write $X_1, \ldots, X_p$ for the data. Put a prior distribution Π on R^p according to which the μ_i are i.i.d. with density π_0 on R. If π_0 is known, the posterior mean of $(\mu_1, \ldots, \mu_p)$ given $(X_1, \ldots, X_p)$, which minimizes the Bayes risk, is

$$\boldsymbol{\delta}_0(X_1, \ldots, X_p) = (\delta_0(X_1), \ldots, \delta_0(X_p))$$

where

$$\delta_0(x) = \frac{\int_{-\infty}^{\infty} \mu\ \phi\left(\frac{x-\mu}{\sigma_0}\right)\pi_0(\mu)d\mu}{\int_{-\infty}^{\infty} \phi\left(\frac{x-\mu}{\sigma_0}\right)\pi_0(\mu)d\mu} = x + \sigma_0^2\frac{f_0'(x)}{f_0(x)} \tag{I.8}$$

and

$$f_0(x) = \frac{1}{\sigma_0}\int_{-\infty}^{\infty} \phi\left(\frac{x-\mu}{\sigma_0}\right)\pi_0(\mu)d\mu\ ,$$

the marginal density of X_1 (Problem I.6.2). The Bayes risk of this estimate is just

$$r(\pi_0, \delta_0) \equiv \sigma_0^2[1 - \sigma_0^2 I(f_0)] < \sigma_0^2 \tag{I.9}$$

where $I(f_0) = \int\{[f_0'(x)]^2/f_0(x)\}\,dx$, the *Fisher information for location of* f_0 (Problem I.6.4).

Suppose now that π_0 is unknown. We shall show, in Chapter 12, following Robbins (1956), that we can construct estimates $\widehat{f}_0'/\widehat{f}_0$ using $X_1, \ldots, X_p$ which, when plugged in to (I.8), yield a purely data dependent estimate $\widehat{\delta}_0$, such that if $r(\pi_0, \delta)$ denotes Bayes risk (see Section 3.3), then $r(\pi_0, \widehat{\delta}_0) \to r(\pi_0, \delta_0)$ as $p \to \infty$ for all π_0. This strictly improves the performance of $\bar{\mathbf{X}}_p$ for $n = 1$, for *all* π_0 (but not uniformly). Here $\widehat{\delta}_0$ is an example on an *empirical Bayes* estimate.

What if both p and n tend to ∞? There is no change in our conclusion that δ_0 is optimal if π_0 is fixed and known. An alternative and more informative analysis leads us to consider priors π_{0n} such that $\sqrt{n}\mu/\sigma_0$, the signal to noise ratio, stabilizes. The extent to which we can or want to try to estimate π_{0n} now depends on the family of priors. We shall consider this as well as related questions in the context of the so called Gaussian white noise model in Chapter 12.

The next section looks at this situation from a different point of view.

I.7 Model Selection

How complex should our model be? Usually this question can be reduced to asking how many parameters should be included in the model. We continue to consider the model $\mathcal{P}(n,p)$ of Section I.6. We identify the model with its parameter space R^p for $(\mu_1, \ldots, \mu_p)$. Consider nested submodels $\mathcal{P}_t(n,p)$, $0 \le t \le p$, specified by $\omega_t = \{\boldsymbol{\mu}_p :\ \mu_{t+1} = \cdots = \mu_p = 0\}$. Here t is unknown and is to be selected on the basis of the data $\mathbf{X}$. Consider the problem of estimating $\boldsymbol{\mu}_p$ as a vector with quadratic loss $l(\boldsymbol{\mu}_p, d)$. If we knew t and, in fact, that $\boldsymbol{\mu}_p \in \omega_t,\ t < p$ we could use $\delta_t(\mathbf{X}) = (X_{\cdot 1}, \ldots, X_{\cdot t}, 0, \ldots, 0)^T$ and obtain

$$R(\boldsymbol{\mu}_p, \delta_t) = \frac{t\sigma_0^2}{n} < \frac{p\sigma_0^2}{n} = R(\boldsymbol{\mu}_p, \delta_p),\quad \boldsymbol{\mu}_p \in \omega_t$$

where R is the risk function

$$R(\boldsymbol{\mu}, \delta) = E|\delta(\boldsymbol{X}) - \boldsymbol{\mu}|^2$$

and $|\cdot|$ denotes Euclidean distance.

This seemingly artificial appearing situation is, in fact, a canonical model for a general linear model replicated n times:

$$Y_{ki} = \sum_{j=1}^{p} z_{kj}\beta_j + e_{ki},\ 1 \le k \le p,\ 1 \le i \le n$$

where the e_{ki} are i.i.d. $\mathcal{N}(0, \sigma_0^2)$ and the z_{kj} are distinct covariate (predictor) values. Then, if $\widehat{\boldsymbol{\beta}}_p$ denotes the MLE,

$$\widehat{\boldsymbol{\beta}}_p \equiv (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^T \sim \mathcal{N}_p(\boldsymbol{\beta}_p, \frac{\sigma_0^2}{n}[\mathbf{Z}_p^T \mathbf{Z}_p]^{-1})$$

where $\boldsymbol{\beta}_p = (\beta_1, \ldots, \beta_p)^T, \mathbf{Z}_p = \|z_{kj}\|_{p\times p}$, and our model $\mathcal{P}(n,p)$ is the special case $\mathbf{Z}_p^T\mathbf{Z}_p = J_{p\times p}$. Our submodels $\{\boldsymbol{\beta}_p : \beta_{t+1} = \cdots = \beta_p = 0\}$ are natural if we think the p predictors or covariates $Z_1, \ldots, Z_p$ whose influence on the distribution of Y is governed by the β_j can be ordered in terms of importance and we expect $(p-t)$ of them to be independent of the response Y. In the context of the model $\mathcal{P}(n,p)$, the questions we address briefly now and more extensively in Chapter 12 are

(1) Suppose $p = \infty$ (the possible number of predictors we can measure is "very large") and we believe that $\boldsymbol{\mu} \in \omega_t$ for some $t < \infty$. That is, $\boldsymbol{\mu} \in \{\boldsymbol{\mu} : \mu_j = 0,\ j > t\}$.

Can we, without knowledge of t, estimate t by $\widehat{t}$ so that, as $n \to \infty$,

$$E|\delta_{\widehat{t}}(\mathbf{X}) - \boldsymbol{\mu}|^2 = \frac{t\sigma_0^2}{n}(1 + o(1)) \tag{I.10}$$

where $|\mathbf{a}|^2 = \sum_{j=1}^{\infty} a_j^2$ for $\mathbf{a} = (a_1, a_2, \ldots)^T$, that is, can we asymptotically do as well not knowing t as knowing it? The optimal procedure for the case where t is known is called the *oracle* solution.

(2) Suppose that all μ_j can be nonzero but $\sum_{j=1}^{\infty} \mu_j^2 < \infty$. Then, we can write the risk as

$$V_n(t, \boldsymbol{\mu}) \equiv E|\boldsymbol{\delta}_t(\mathbf{X}) - \boldsymbol{\mu}|^2 = \frac{t\sigma_0^2}{n} + \sum_{j=t+1}^{\infty} \mu_j^2 \tag{I.11}$$

There is clearly a best $t(\boldsymbol{\mu}, n)$, one such that

$$t(\boldsymbol{\mu}, n) = \arg\min_t V_n(t, \boldsymbol{\mu}).$$

Note that $t(\boldsymbol{\mu}, n) = \infty$, that is, estimating each μ_j by $X_{\cdot j}$ is always a bad idea! Putting $t = 0$ will always do better. Can we select $\widehat{t}(n)$ such that

$$\frac{E|\boldsymbol{\delta}_{\widehat{t}(n)}(\mathbf{X}) - \boldsymbol{\mu}|^2}{V_n(t(\boldsymbol{\mu}, n), \boldsymbol{\mu})} \longrightarrow 1 \tag{I.12}$$

as $n \to \infty$?

For question (1) we want (I.10) and (I.12) to hold uniformly over "moderately large" sets of $\boldsymbol{\mu}$. It is natural to consider $\widehat{t}$ of the form: $\widehat{t}$ is the largest k such that for a suitable decreasing sequence $\{c_n\}$ of positive numbers, $|X_{\cdot j}| \le c_n$ for all j such that $k+1 \le j \le n$. Suppose $p \le n$. Finding the c_n such that this $\widehat{t}$ solves (I.10) then turns out to lead to a special case of the solution to Schwarz's Bayes criterion (SBC) (1978) which also is called

the "Bayes information criterion" (BIC). Let $P_{\boldsymbol{\mu}}$ denote computation under $\boldsymbol{\mu}$; then take c_n such that

$$P_{\mathbf{0}}[\max\{|X_{\cdot j}| : 1 \le j \le n\} \ge c_n] \to 0 \tag{I.13}$$

and

$$P_{\boldsymbol{\mu}}[|X_{\cdot j}| \le c_n] \to 0\ . \tag{I.14}$$

Since, for $j \ge t+1$, the $X_{\cdot j}$ are i.i.d. $\mathcal{N}(0, \sigma_0^2/n)$ it is easy to see (Problem I.7.1) that

$$c_n = \sigma_0\sqrt{(2\log n)/n}$$

will achieve (I.10) without knowledge of t. To see this compute for $\boldsymbol{\mu} \in \omega_t$,

$$\begin{aligned} P_{\boldsymbol{\mu}}[\widehat{t} \ne t] &= P_{\boldsymbol{\mu}}[\widehat{t} < t] + P_{\boldsymbol{\mu}}[\widehat{t} > t] \\ &\le P_{\boldsymbol{\mu}}[|X_t| \le c_n] + P_{\boldsymbol{\mu}}[\max\{|X_{\cdot j}| : t+1 \le j \le n\} > c_n] \end{aligned} \tag{I.15}$$

(Problem I.7.2). So, (I.13), (I.14), and (I.15) establish our answer to question (1).

The solution to question (2) is subtler and somewhat different and was first proposed in a time series context by Akaike (1969) and in the regression context by Mallows (1973). We will show that $\sum_{j=t+1}^n X_{\cdot j}^2$ can be used to obtain an unbiased estimate of $V_n(t, \boldsymbol{\mu})$. Evidently

$$E_{\boldsymbol{\mu}}\left(\sum_{j=t+1}^n X_{\cdot j}^2\right) = (n-t)\frac{\sigma_0^2}{n} + \sum_{j=t+1}^n \mu_j^2\ .$$

Therefore,

$$E_{\boldsymbol{\mu}}\left(\sum_{j=t+1}^n X_{\cdot j}^2 + 2t\frac{\sigma_0^2}{n} - \sigma_0^2\right) = V_n(t, \boldsymbol{\mu}),$$

and it seems plausible since σ_0^2 doesn't depend on t that minimizing the contrast

$$\rho_n(\mathbf{X}, k) \equiv \sum_{j=k+1}^n X_{\cdot j}^2 + 2k\frac{\sigma_0^2}{n}, \quad 1 \le k \le n$$

will yield $\widehat{\widehat{t}}$ which achieves (I.12). This can be shown under suitable conditions (Shibata (1981)). However, it is interesting to note that this solution, which is called Mallows' C_p, is quite different from the SBC solution. Because

$$\rho_n(\mathbf{X}, j) - \rho_n(\mathbf{X}, j-1) = -X_{\cdot j}^2 + 2\frac{\sigma_0^2}{n}$$

then $\rho_n(X, j) \le \rho_n(X_n, j-1)$ iff $(nX_{\cdot j}^2/\sigma_0^2) \ge 2$. That is, we prefer j parameters to $j-1$ iff $\sqrt{n}|X_{\cdot j}| \ge \sqrt{2}$. Thus, Mallows' C_p essentially uses a threshold of $\sqrt{2}$ on $\sqrt{n}|X_{\cdot j}|/\sigma_0$

rather than the corresponding threshold $\sqrt{2 \log n} \to \infty$ for SBC. These methods and others will be discussed and contrasted with oracle solutions in Chapter 12. □

Remark I.7.1. Mallows' C_p is usually discussed in the context of squared prediction error rather than squared estimation error. These two criteria are equivalent under general conditions; see Problem I.7.5.

Remark I.7.2. More generally we can consider risks other than those based on squared error. We shall do so in Chapter 12.

Remark I.7.3. This introductory chapter has barely touched on some of the main topics of Chapter 12. These topics include what is referred to as "Statistical Learning," See Section 12.1 for an introduction to this topic.

Summary We have in this chapter introduced, through examples, some of the main issues, topics, and procedures to be considered in this volume. One issue is the level of accuracy possible in the estimation of a parameter. In Section I.3, we illustrated *regular parameters*, which are parameters that are relatively easy to estimate in the sense that $\sqrt{n}$ times the estimation error is well behaved as $n \to \infty$. We also discussed, in Section I.1, *Doob's heuristic*, which suggests how to approximate the distribution of a functional of a sample-based stochastic process by the distribution of the functional of the limiting stochastic process. Other methods for obtaining approximations, *Monte Carlo Methods* and *the bootstrap*, are illustrated in Section I.2. We also gave, in Section I.3, important parameters that are regular but nevertheless do not fall under the framework of Vol. I and discussed efficient estimation of parameters for a given model. In Section I.4 we discussed situations where we can construct tests, called *permutations tests*, whose significance level α remain the same for every member of a general class of distributions $\mathcal{P}_0$. An important subclass of the permutation tests is the *rank tests*. In Section I.5 we illustrated *irregular parameters*, and the notion of plugging into approximations of irregular parameters such as densities. In this context we introduced the bias variance tradeoff. In Section I.6 we introduced Stein estimation, which, in our illustration with p Gaussian population means, consists of providing vector estimates with smaller average mean squared error than the vector of p sample means. We then connected this approach to the *empirical Bayes* approach where we solve the Bayes estimation problem with p Gaussian means assuming the prior is known and then use the data to estimate the unknowns in the Bayes procedure. Finally, in Section I.7 we considered model selection where the question is how many parameters should be included in a model used to analyze the results of an experiment. A complex model with many parameters may represent the true distribution of the data better than a simpler model with fewer parameters. However, we illustrated that trying to estimate too many parameters for the given sample size may result in worse inference than fitting a simpler model providing a less adequate approximation to the truth. This is another reflection of the bias-variance tradeoff discussed in Section I.5. We used a simple p-sample framework to outline the derivations of two procedures that give the most accurate result in the sense of minimizing average mean squared error for the problem of estimating a vector of means. The first method, which applies to the situation where all but a finite number of parameters are zero, turns out to be a special case of the method generated by *Schwarz's Bayes criterion* (also called BIC), while the

second method, which allows for an infinite sequence of positive means, yields a special case of *Mallows' C_p*.

I.8 PROBLEMS AND COMPLEMENTS

Problems for Section I.1

1. Let $X_1, \ldots, X_n$ be i.i.d. as $X \sim F$, let $\widehat{F}(x) = n^{-1} \sum_{i=1}^{n} 1[X_i \leq x]$ be the empirical distribution function, and let $\mathcal{E}_n(x) = \sqrt{n}[\widehat{F}(x) - F(x)],\ x \in R$.

(a) Show that for fixed x_0, $\mathcal{E}_n(x_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(x_0)[1 - F(x_0)])$

(b) Find $\text{Cov}(\mathcal{E}_n(x_1), \mathcal{E}_n(x_2))$ for $x_1 \leq x_2$.

(c) Find the limiting law of $(\mathcal{E}_n(x_1), \mathcal{E}_n(x_2))$ for $x_1 \leq x_2$ using the bivariate central limit theorem.

2. Let $X_1, \ldots, X_m$ be i.i.d. as $X \sim F$, $Y_1, \ldots, Y_n$ i.i.d. as $Y \sim G$, and let the X's and Y's be independent (see Example 1.1.3). Let $\widehat{F}(\cdot)$ and $\widehat{G}(\cdot)$ be the empirical df's, define $N = m + n$, and

$$\mathcal{E}_N(t) = \sqrt{\frac{mn}{N}}\{\widehat{G}(t) - \widehat{F}(t) - [G(t) - F(t)]\}$$

Assume that $\frac{m}{N} \to \lambda$ as $N \to \infty$ with $0 < \lambda < 1$.

(a) Show that for fixed t_0, as $N \to \infty$,

$$\mathcal{E}_N(t_0) \xrightarrow{\mathcal{L}} \sqrt{\lambda} Z_1(t_0) + \sqrt{(1-\lambda)} Z_2(t_0)$$

where $Z_1(t_0)$ and $Z_2(t_0)$ are independent with $Z_1(t_0) \sim \mathcal{N}(0, G(t_0)[1 - G(t_0)])$ and $Z_2(t_0) \sim \mathcal{N}(0, F(t_0)[1 - F(t_0)])$.

(b) Describe the weak limit of $\mathcal{E}_N(t)$.

Problems for Section I.2

1. Let $X_1, \ldots, X_n$ be i.i.d. as $X \sim F(x) = \Phi([x-\mu]/\sigma)$. Suppose we observe $Y_1, \ldots, Y_m$ i.i.d. as $Y \sim (X|X \geq 0)$. Let $\mu = E(X),\ \sigma^2 = \text{Var}(X)$, and $\boldsymbol{\theta} = (\mu, \sigma^2)^T$.

(a) Express $E(Y)$ and $\text{Var}(Y)$ as functions of $\boldsymbol{\theta}$. By replacing $E(Y)$ and $\text{Var}(Y)$ by $\overline{Y}$ and $\widehat{\sigma}_Y^2$ in these two equations and solving them for $\boldsymbol{\theta}$ numerically, we have method of moment estimates of μ and σ^2.

(b) Use Proposition 5.2.1 to outline an argument for the consistency of the estimates of μ and σ^2 in part (a).

(c) Use Theorem 5.2.3 to outline an argument for the consistency of the MLE of $\boldsymbol{\theta}$.

Problems for Section I.5

1. Suppose p is positive and Lipschitz continuous over I_j, that is, for some $\gamma > 0$ and all $x, z \in I_j$, $|p(x) - p(z)| \leq \gamma|x - z|$. Show that

(a) BIAS $\widehat{p}_h(t) = O(h)$, all $t \in I_j$.

(b) Var $\widehat{p}_h(t) = O((nh)^{-1})$, all $t \in I_j$.

(c) If $p(x)$ is not constant on I_j, then for some constant $c > 0$,

$$c < \inf_{h>0} h^{-1}|\text{BIAS}\widehat{p}_h(t)| \leq \sup h^{-1}|\text{BIAS}\widehat{p}_h(t)| \leq c^{-1} .$$

(d) Let t be as in (c) above. Assume that p' exists and that $0 < |p'(x)| \leq M < \infty$ for x in a neighbourhood of t. Show that

$$\inf_h \text{MSE}\ \widehat{p}_h(t) = \text{Var}\ \widehat{p}_h(t) + [\text{BIAS}\ \widehat{p}_h(t)]^2 \asymp (n^{-\frac{2}{3}}) ,$$

where $A_n \asymp B_n$ iff $A_n = O(B_n)$, $B_n = O(A_n)$.

Hint (a) and (b). By the mean value theorem, for some $x_0 \in I_j$,

$$P[I_j(t)] = \int_{I_j} p(x)d(x) = hp(x_0) .$$

Now show that

$$|\text{BIAS}\ \widehat{p}_h(t)| \leq \gamma|x_0 - t| \leq \gamma h,\ \text{Var}\ \widehat{p}_h(t) \leq \frac{p(x_0)}{nh} .$$

Hint (d). Show that MSE $\widehat{p}_n(t) = b(nh)^{-1} + ch^2$ plus smaller order terms for some constants b and c. Now minimize $A(h) \equiv b(nh)^{-1} + ch^2$ with respect to h.

Problems for Section I.6

1. Show that $\bar{\mathbf{X}}_P$ is the minimax estimate for the model $\mathcal{P}(n, p)$ and the risk (I.6).
Hint. See Example 3.3.3.

2. Derive formula (I.8).
Hint. The first equality is from (3.2.6).

3. Show that the Bayes risk for estimating $q(\boldsymbol{\theta})$ with quadratic loss when $\boldsymbol{\theta}$ has prior π is

$$E\text{Var}(q(\boldsymbol{\theta}|\boldsymbol{X}) = \int \left[q(\boldsymbol{\theta}) - \delta^*(\boldsymbol{x})\right]^2 \pi(\boldsymbol{\theta})p(\boldsymbol{x}, \theta)d\boldsymbol{\theta}d\boldsymbol{x},$$

where $\delta^*(\boldsymbol{x})$ is the Bayes estimate of $q(\boldsymbol{\theta})$.

4. Derive formula (I.9).
Hint. Use (I.8), (5.4.32), and Problem I.6.3.

Problems for Section I.7

1. Show that if $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(0,1)$, then

$$P[\max(X_1, \ldots, X_n) \leq \sqrt{2 \log n}] \to 1 \, .$$

Hint. Use the following refinement of (7.1.9) (Feller (1968), p. 175)

$$(x^{-1} - x^{-3})\phi(x) \leq (1 - \Phi(x)) \leq x^{-1}\phi(x) \qquad \text{for all } \; x > 0.$$

2. Verify (I.15).
Hint. If $Z_1, \ldots, Z_n$ are i.i.d. $\mathcal{N}(0,1)$, $P[\max\{|Z_i| \; : \; 1 \leq i \leq n\} \geq z] \leq n \, P[|Z_1| \geq z]$. Use $1 - \Phi(z) \leq \phi(z)/z$ for all $z > 0$.

3. *Selecting the model to minimize MSE. The t-test revisited.* Consider the simple linear regression model (e.g. Example 2.2.2)

$$Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i, \; i = 1, \ldots, n \tag{I.8.1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. as $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and the z_i's are not all equal. Without loss of generality we assume $\sum z_i = 0$ and $\beta_0 = E(\overline{Y})$. We are interested in estimating

$$\mu_i \equiv \mu(z_i) \equiv E(Y_i) = \beta_0 + \beta_1 z_i, \; i = 1, \ldots, n.$$

In this problem we assume that (I.8.1) is the true model. Even so, it may be that the MLE $\widehat{\mu}_{10} = \widehat{\beta}_0 = \overline{Y}$ based on the submodel with $\beta_1 = 0$ has smaller risk than the MLE $\widehat{\mu}_{1i} \equiv \overline{Y} + \widehat{\beta}_1 z_i$ based on the full model. For any estimate $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)^T$ we define the risk for estimating $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ as

$$R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}) = \frac{1}{n} E|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}|^2 = \frac{1}{n} \sum_{i=1}^{n} E|\widehat{\mu}_i - \mu_i|^2$$

where the expectation is computed for the full model (I.8.1).

(a) Show that for $\widehat{\boldsymbol{\mu}}_j = (\widehat{\mu}_{j1}, \ldots, \widehat{\mu}_{jn})^T$,

$$R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_j) = \frac{(1+j)\sigma^2}{n} + \frac{(1-j)\beta_1^2 \sum_{i=1}^n z_i^2}{n}, \; j = 0, 1 \tag{I.8.2}$$

Hint. See Example 6.1.2, pages 381–382.

(b) Show that unbiased estimates of $R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_j), \; j = 0, 1$, are

$$\widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0) \; \equiv \; n^{-1}[RSS_0 - RSS_1] = n^{-1}|\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0|^2 = \widehat{\beta}_1^2 \left(\frac{\sum_{i=1}^n z_i^2}{n} \right)$$

$$\widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_1) \; = \; 2s^2/n$$

where $RSS_j = \sum_{i=1}^n [Y_i - \widehat{\mu}_{ji}]^2, \; j = 0, 1$, and $s^2 = RSS_1/(n-2)$.
Hint. See Section 6.1.3.

(c) Show that the model selection rule that decides to keep β_1 in the model iff $\widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_1) < \widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0)$ is equivalent to a likelihood ratio test of $H : \beta_1 = 0$ vs $K : \beta_1 \neq 0$. Show that the rule and the test can be written as

$$\text{"Keep } \beta_1 \text{ in the model iff } t^2 > 2\text{"}$$

where $t = \widehat{\beta}_1/(s/\sqrt{n}\widehat{\sigma}_z)$ with $\widehat{\sigma}_z^2 = n^{-1}\sum_{i=1}^n z_i^2$ is the t-statistic for regression.
Hint. See Example 6.1.2, pages 381–382.

(d) Show that the limit as $n \to \infty$ of the significance level of the test in (c) equals $P(|Z| > \sqrt{2}) = 0.16$ where $Z \sim \mathcal{N}(0,1)$. That is, we decide to keep β_1 if the p-value is less than 0.16. Show this result without the normality assumption. Instead assume that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. as ε with $E(\varepsilon) = 0,\ E(\varepsilon^2) = \sigma^2,\ E(\varepsilon^4) < \infty$. Also assume $\max_i\{z_i^2/\Sigma z_i^2\} \to 0$.
Hint. Use Lindeberg-Feller and Slutsky's theorems. See Example 5.3.3.

(e) Using (a), show that $R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_1) < R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0)$ iff $\tau^2 > 1$ where $\tau = \beta_1/(\sigma/\sqrt{n}\widehat{\sigma}_z)$ is the noncentrality parameter of the distribution of the t-statistic.
Hint. See page 260, Section 4.9.2.

(f) Show that $\widehat{\tau}^2 \equiv ct^2 - 1$ with $c = (n-4)/(n-2)$ is an unbiased estimate of τ^2. Using $\widehat{\tau}^2$, deciding in favor of keeping β_1 is now equivalent to $t^2 > 2(n-2)/(n-4)$ for $n \geq 5$.
Hint. $t \stackrel{\mathcal{L}}{=} (Z+\tau)/\sqrt{V/(n-2)}$ where $Z \sim \mathcal{N}(0,1),\ V \sim \mathcal{X}^2_{n-2}$, and Z and V are independent. Use Problem B.2.4.

(g) Show that $\widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_1) < \widehat{R}(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0)$ is also equivalent to $r^2 > 2/n$, where r^2 is the sample correlation coefficient and we assume $n \geq 3$. Also show $t^2 \geq 2(n-2)/(n-4)$ iff $r^2 \geq 2/(n-2)$.
Hint. $t^2 = n\widehat{\beta}_1^2\widehat{\sigma}_z^2/RSS_1$ and $r^2 = \widehat{\beta}_1^2\widehat{\sigma}_z^2/\widehat{\sigma}_Y^2$ where $\widehat{\sigma}_Y^2 = n^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$. Now use the ANOVA decomposition $n\widehat{\sigma}_Y^2 = RSS_1 + \widehat{\beta}^2 n\widehat{\sigma}_z^2$ (see Section 6.1.3).

4. *Selecting the model to minimize MSE. The F-test revisited.* Consider the linear model (see 6.1.26)

$$\boldsymbol{Y} = \boldsymbol{Z_1}\boldsymbol{\beta_1} + \boldsymbol{Z_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon} \tag{I.8.3}$$

where $\boldsymbol{Z}_1$ is $n \times q$ and $\boldsymbol{Z}_2$ is $n \times (p-q)$, $\boldsymbol{\beta}_1$ is $q \times 1$, $\boldsymbol{\beta}_2$ is $(p-q) \times 1$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ with $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. as $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The entries of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are constants. Assume that $\boldsymbol{Z} \equiv (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ has rank p. We are interested in estimating the effect of the covariates on the response mean $E(Y)$. Thus our parameters are

$$\mu_i \equiv \mu(\boldsymbol{z}_i) \equiv E(Y_i) = \boldsymbol{z}_i^{(1)}\boldsymbol{\beta}_1 + \boldsymbol{z}_i^{(2)}\boldsymbol{\beta}_2,\ i = 1, \ldots, n$$

where $\boldsymbol{z}_i^{(j)}$ is the ith row of $\boldsymbol{Z}_j,\ j = 1, 2$. It may be that the coefficients in $\boldsymbol{\beta}_2$ are so small that in the bias-variance tradeoff we get more efficient estimates of the μ_i's if we use the

model with $\boldsymbol{\beta}_2 = \mathbf{0}$. Thus we want to compare the risks of the two estimates $\widehat{\boldsymbol{\mu}}_0 = H_1\boldsymbol{Y}$ and $\widehat{\boldsymbol{\mu}} = H\boldsymbol{Y}$ where H_1 and H are the hat matrices (e.g., $H_1 = \boldsymbol{Z}_1(\boldsymbol{Z}_1^T\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^T$) for the models with $\boldsymbol{\beta}_2 = 0$ and general $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, respectively. The risk for estimating $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is

$$R(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}) = n^{-1}E|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}|^2 = n^{-1}\sum_{i=1}^{n} E[\widehat{\mu}_i - \mu_i]^2$$

where the expectation is computed for the full model.

(a) Set $\boldsymbol{\mu}_0 = H_1\boldsymbol{\mu}$. Show that

$$R_q \equiv R_q(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}_0) = \frac{q\sigma^2}{n} + \frac{|\boldsymbol{\mu} - \boldsymbol{\mu}_0|^2}{n},\ 1 \le q \le p\ .$$

Note that when $q = p$, then $\widehat{\boldsymbol{\mu}}_0 = \widehat{\boldsymbol{\mu}}$, and $R_p(\boldsymbol{\mu}, \widehat{\boldsymbol{\mu}}) = p\sigma^2/n$.
Hint. See (6.1.15).

(b) (i) Let $s^2 = [n - (p+1)]^{-1}\sum[Y_i - \widehat{Y}_i]^2$ where $\widehat{Y}_i$ is the predicted value of Y_i in the full model (I.8.3). Show that an unbiased estimate of $R_p - R_q$ for model (I.8.3) is

$$\widehat{R}_p - \widehat{R}_q = \frac{2(p-q)s^2}{n} - \frac{|\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0|^2}{n}\ .$$

(ii) Show that deciding to keep $Z_2\boldsymbol{\beta}_2$ in the model when $\widehat{R}_p < \widehat{R}_q$ is equivalent to deciding $\boldsymbol{\beta}_2 = 0$ when $F > 2(p-q)$ where "$F > 2(p-q)$" is a likelihood ratio test of $H : \boldsymbol{\beta}_2 = \mathbf{0}$ vs $K : \boldsymbol{\beta}_2 \neq \mathbf{0}$.
Hint. See Example 6.1.2.

(c) (i) Use (a) to show that $R_p < R_q$ is equivalent to $\theta^2 > p - q$, where $\theta^2 = \sigma^{-2}|\boldsymbol{\mu} - \boldsymbol{\mu}_0|^2$ is the noncentrality parameter of the distribution of the F-statistic of Example 6.1.2.

(ii) Show that $\widehat{\theta^2} = (n-p)^{-1}(p-q)(n-p-2)F - (p-q)$ is an unbiased estimate of θ^2. Thus, using $\widehat{\theta^2} > (p-q)$, we select the full model iff

$$F > \frac{(p-q+1)(n-p)}{(p-q)(n-p-2)},\ n > p+2\ .$$

Hint. By Problem 8.3.13, $F = [(p-q)^{-1}(Z_1+\theta)^2 + \sum_{i=1}^{p-q} Z_i^2(n-p)^{-1}V^{-1}]$ where $Z_i \sim \mathcal{N}(0,1),\ V \sim \mathcal{X}^2_{n-p}$, and $Z_1, \ldots, Z_{p-q},\ V$ are independent.

5. *Prediction and estimation are equivalent for squared error.* Assume that $E(Y_i) = \mu_i$ and $\mathrm{Var}(Y_i) = \sigma_i^2$ where μ_i depends on a vector $\boldsymbol{z}_i$ of covariates while σ_i^2 does not, $i = 1, \ldots, n$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T, \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, and let $\widehat{\boldsymbol{\mu}}$ be any estimate of $\boldsymbol{\mu}$ based on $\boldsymbol{Y}$ with $0 < \mathrm{Var}\widehat{\mu}_i < \infty$ for $i = 1, \ldots, n$. Let $Y_1^0, \ldots, Y_n^0$ be variables to be predicted using $Y_1, \ldots, Y_n$. We assume that $\boldsymbol{Y}^0 = (Y_1^0, \ldots, Y_n^0)^T$ is independent of $\boldsymbol{Y}$,

and that Y_i^0 has the same mean and variance as $Y_i,\ i = 1, \ldots, n$. We will use $\widehat{\boldsymbol{\mu}}$ to predict $\boldsymbol{Y}^0$ and define the mean squared prediction error as

$$\text{MSPE} = n^{-1} E|\widehat{\boldsymbol{\mu}} - \boldsymbol{Y}^0|^2 \ .$$

Show that selecting the model (covariates, as in Problem 4 preceding) that minimizes MSPE is equivalent to selecting the model (covariates) that minimizes the mean squared error MSE $= n^{-1} E|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}|^2$.
Hint. $\widehat{\mu}_i - Y_i^0 = [\mu_i - Y_i^0] + [\widehat{\mu}_i - \mu_i]$. Complete the square keeping the square brackets intact. Because Y_i^0 is independent of $\widehat{\mu}_i$, we have

$$E[\widehat{\mu}_i - Y_i^0]^2 = \sigma_i^2 + E[\widehat{\mu}_i - \mu_i]^2 \ . \tag{I.8.4}$$

Note that the equivalence fails if $\sum_{i=1}^n \sigma_i^2$ depends on the covariates.

I.9 Notes

Notes for Section I.1.
(1) Important exceptions are Sections 1.4 and 6.6.